

# QTL mapping and GWAS

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

3/16/2021

# Outline

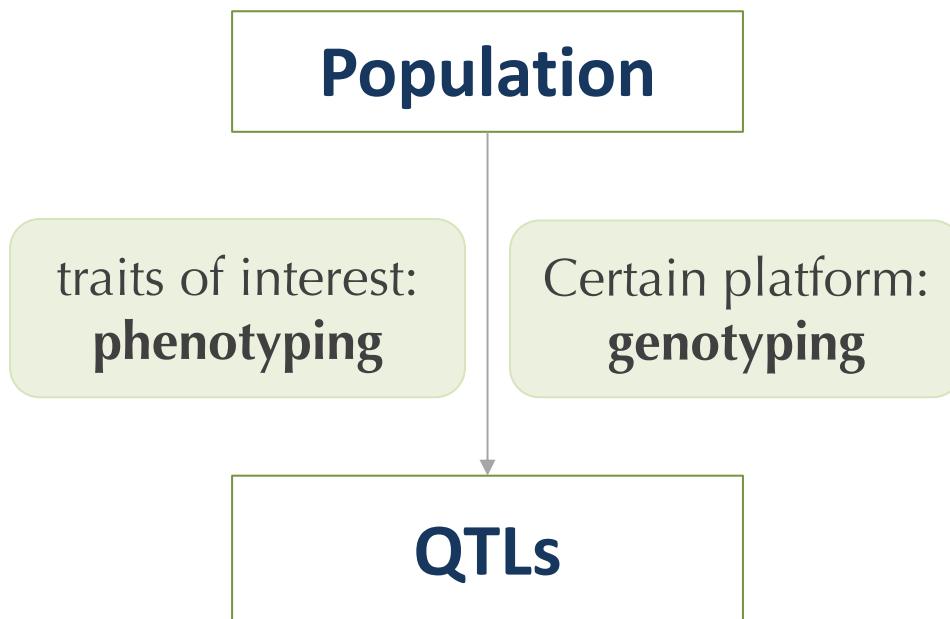
- QTL mapping
- Genome-wide association study (GWAS)

*Acknowledgements: some slides were prepared by Dr. Lei Li.*

What is the goal to perform  
QTL or GWAS?

# QTL mapping

A Quantitative Trait Locus (QTL) is a genomic locus that genetically influence variation in a phenotype of a quantitative trait.



Genetic linkage map or a physical map would be helpful to identify QTLs and locate the QTL on a map

# Sequencing technology is an excellent tool to genotype many loci in parallel



ATCGCTGCCGATCTGC~~G~~TCATA~~C~~GGAA~~T~~CGTC~~G~~G~~T~~T~~C~~A~~G~~  
ATCGCTGCCGATCTGC~~G~~TGATA~~C~~GGAA~~T~~CGTC~~G~~G~~T~~T~~C~~A~~G~~  
ATCGCTGCCGATCTGC~~G~~TCATA~~C~~GGAA~~T~~CGTC~~G~~G~~T~~T~~C~~A~~G~~  
ATCGCTGCCGATCTGC~~G~~TGATA~~C~~GGAA~~T~~CGTC~~G~~G~~T~~T~~C~~A~~G~~  
ATCGCTGCCGATCTGC~~G~~TGATA~~C~~GGAA~~T~~CGTC~~G~~G~~T~~T~~C~~A~~G~~  
ATCGCTGCCGATCTGC~~G~~TGATA~~C~~GGAA~~T~~CGTC~~G~~G~~T~~T~~C~~A~~G~~  
ATCGCTGCCGATCTGC~~G~~TCATA~~C~~GGAA~~T~~CGTC~~G~~G~~T~~T~~C~~A~~G~~  
ATCGCTGCCGATCTGC~~G~~TGATA~~C~~GGAA~~T~~CGTC~~G~~G~~T~~T~~C~~A~~G~~  
ATCGCTGCCGATCTGC~~G~~TCATA~~C~~GGAA~~T~~CGTC~~G~~G~~T~~T~~C~~A~~G~~  
ATCGCTGCCGATCTGC~~G~~TGATA~~C~~GGAA~~T~~CGTC~~G~~G~~T~~T~~C~~A~~G~~

Genotyping score

C / ~~G~~

1 / ~~0~~

Marker

# Phenotyping



wikimedia.org

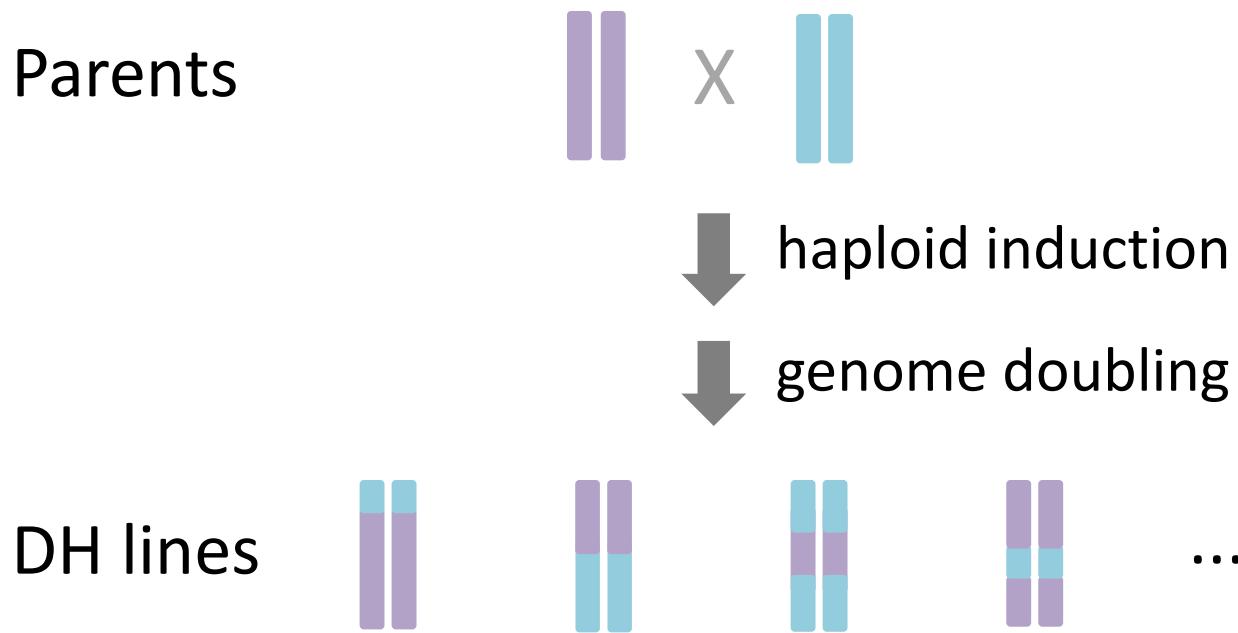


agphd.com

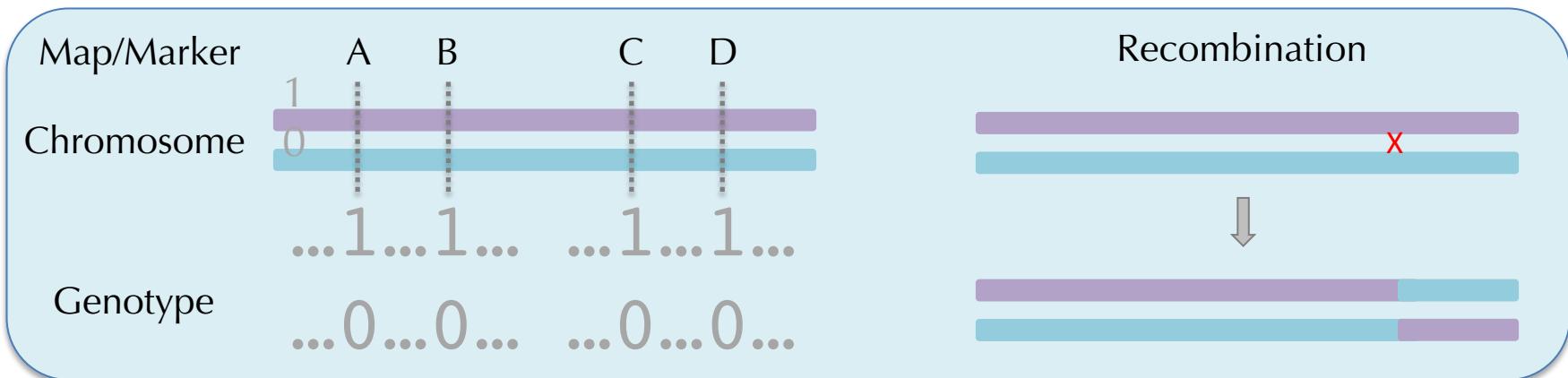
High-throughput phenotyping

# Mapping populations

1. F1, F2
2. Recombinant Inbred Lines
3. Double haploid (DH) lines



# Mapping a causal genetic controlling component (X)



**Mapping population**

A	B	C	D	X	X'
1	1	●	1	1	35
0	0	0	0	0	3
1	1	●	1	1	24
0	0	0	1	0	12
1	1	●	0	1	45
1	0	0	0	0	18
0	1	●	1	1	20

**Mapping result**

A	B	X	C	D	Phenotype
1	1	●	1	1	Phenotype 8

**Genotype**      **Phenotype**      **Phenotype**

# Approach 1: t-test or ANOVA

1. Based on the genotype data, individuals are divided into groups

2. Perform t-test or ANOVA

3. Repeat for all markers

(use t-test if only two groups exist)

genotype	phenotype
1	35
1	24
1	20
1	45

## Pros:

- Simple
- No genetic map required

0	3
0	18
0	12

## Cons:

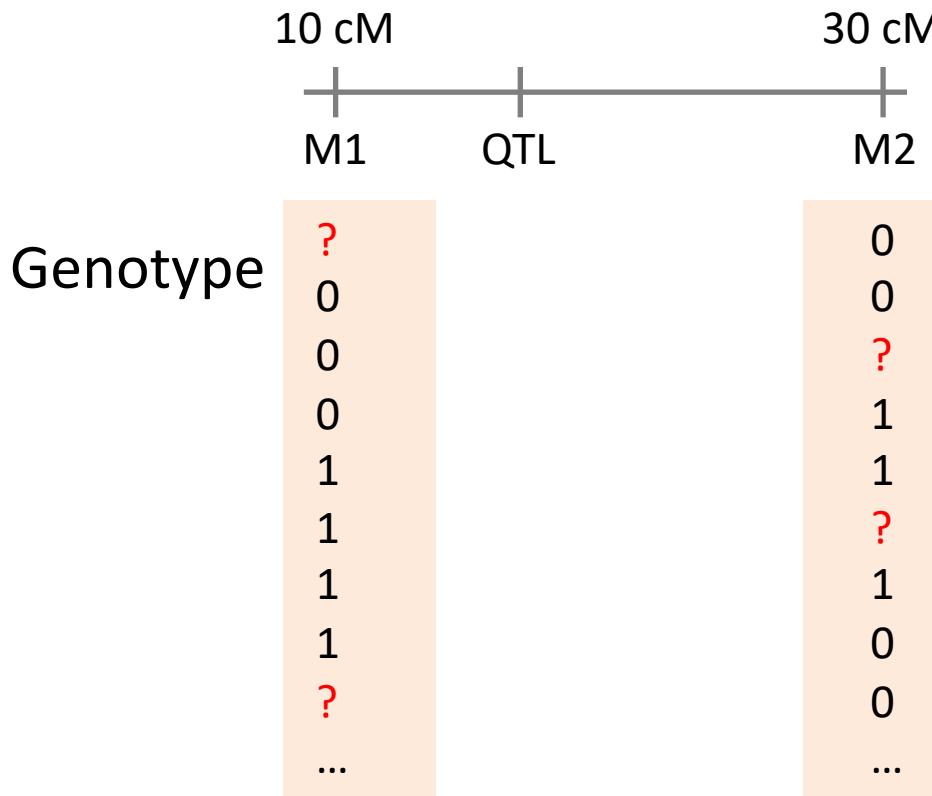
- Individuals with missing data are excluded
- Suffers in low density markers

## Approach 2: Interval mapping (IM)



- Assume a single QTL model (QTL at a certain genetic position)
- Determine the ***confidence*** of each QTL model
- Scan the whole genome (interval by interval)

# Interval mapping – estimate genotypes



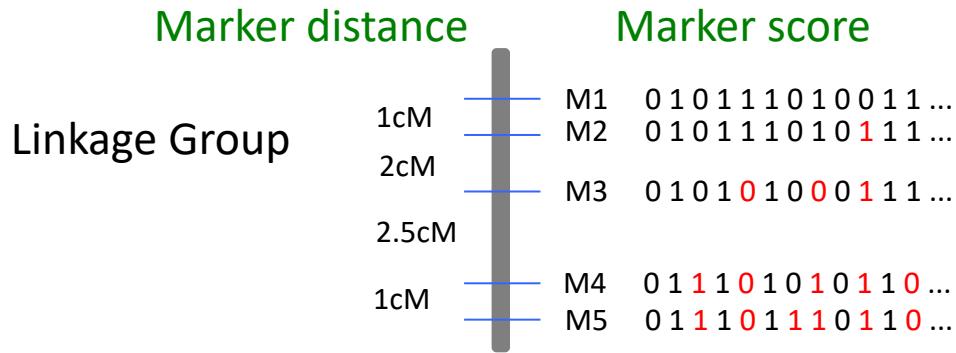
**Estimate genotypes:**  
each estimated genotype  
is associated with a certain  
probability

- Genetic linkage map

Assume a single QTL model (QTL at a certain genetic position)

# Genetic linkage map

- Describe the linear order of markers within a linkage group



- Recombination frequency:** the percentage of recombinant gametes produced in a cross

$$\text{Recombination frequency } (r) = \# \text{recombinants} / \text{total} \times 100\%$$

- 1 centimorgan (cM) apart on a genetic map indicates approximately 1% of recombination events.

# Mapping function

- Conversion between recombination frequencies and genetic distances
- Different formula (Haldane and Kosambi)
- Haldane's mapping function

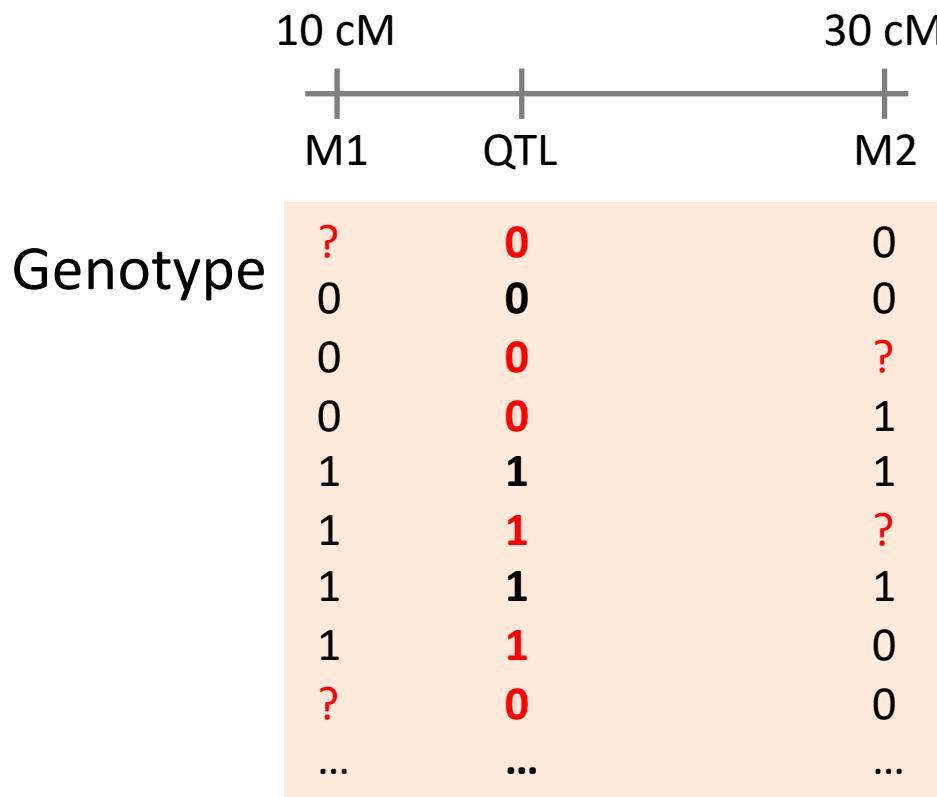
$$r = \frac{1}{2} \left( 1 - e^{-2d} \right)$$

$$d = -\frac{1}{2} \ln(1 - 2r)$$

$r$  = recombination rate (0-0.5)

$d$  = distance in Morgans

# Interval mapping – estimate genotypes



each estimated genotype is associated with a certain probability

# Estimate likelihood of a QTL model

- ***Maximum likelihood estimates (MLE)***

$\text{Prob}(\text{pheno data} \mid \text{geno data}; \text{a QTL at a given position})$

e.g., EM algorithm, Haley-Knott regression (HK)

- ***No QTL Likelihood***

$\text{Prob}(\text{pheno data} \mid \text{geno data}; \text{no QTL})$

# LOD (logarithm of the odds)

$$LOD = \log_{10} \frac{Prob(\text{pheno data} \mid \text{geno data; a QTL at a given position})}{Prob(\text{pheno data} \mid \text{geno data; no QTL})}$$

LOD =  **$\log_{10}$  likelihood ratio**, comparing a single-QTL model to the “no QTL anywhere”.

The **LOD score** is a measure of the strength of evidence for the presence of a QTL at a particular location.

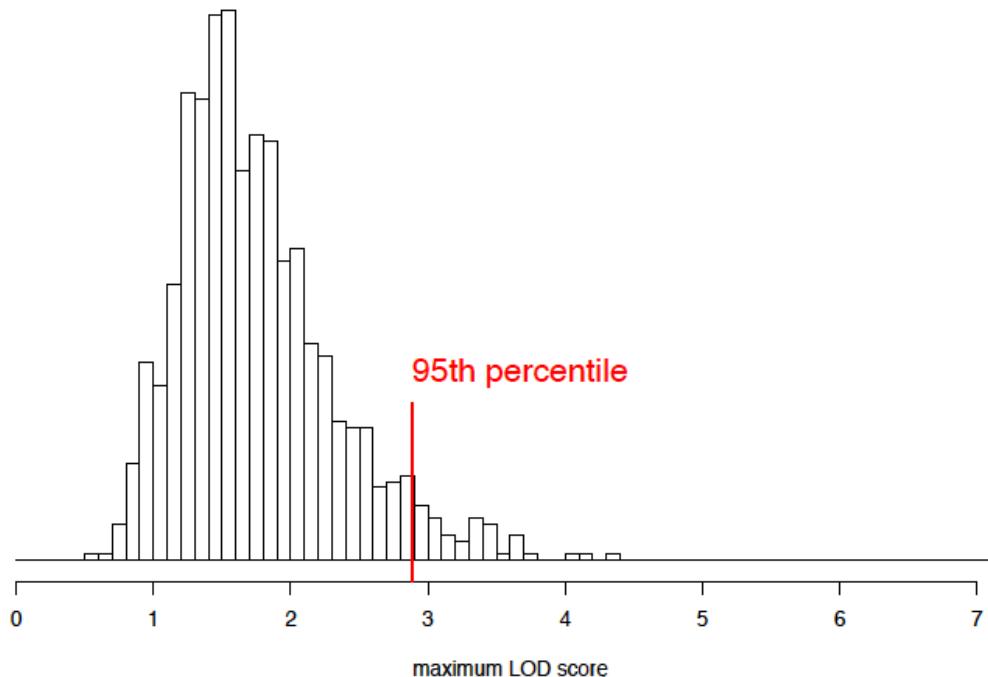
LOD scores must be closer to 3 before they will generally be deemed interesting. - Broman, Lab Animal, 30(7):44–52, 2001

# LOD = 3?

$$LOD = \log_{10} \frac{Prob(\text{pheno data} \mid \text{geno data; a QTL at a given position})}{Prob(\text{pheno data} \mid \text{geno data; no QTL})}$$

# Permutation tests to infer a *LOD* threshold

- Permute/shuffle the phenotypes; keep the genotype data intact.
- QTL analysis and get the max(LOD) ( $\text{maxLOD}_1$ )
- Repeat 1000 times to have ( $\text{maxLOD}_1$ ,  $\text{maxLOD}_2$ , ...  $\text{maxLOD}_{1000}$ )
- The 95<sup>th</sup> percentile of MaxLOD is a genome-wide LOD threshold.

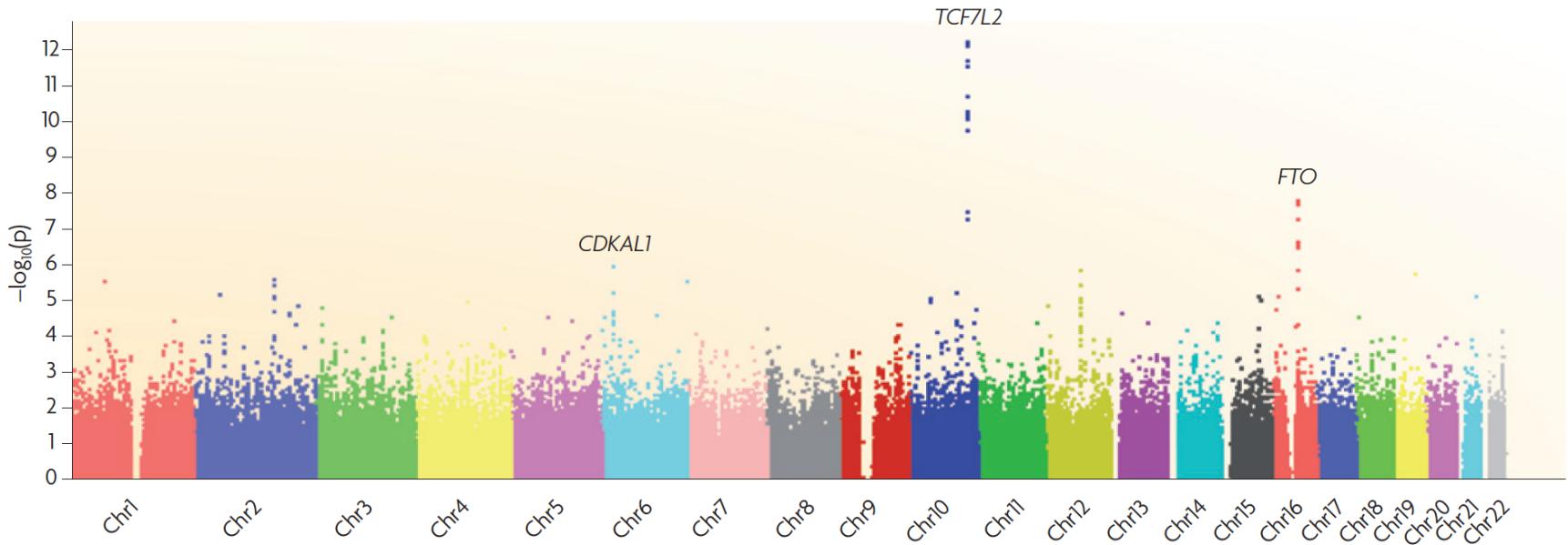


# Question

Can we perform a QTL study on a human population?

# Genome-wide association study (GWAS)

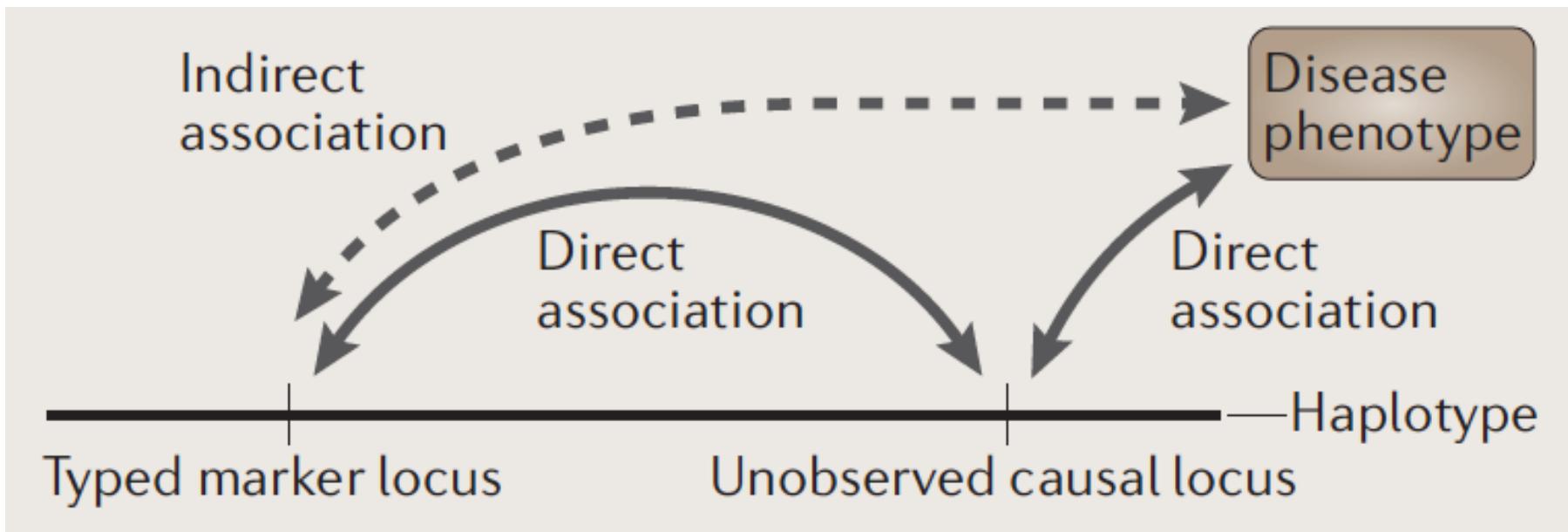
GWAS is the study to correlate a great number of **genomic variants** with a large number of individuals to identify variants that are significantly associated with **the phenotype of interest**.



Goal: to identify causal variants

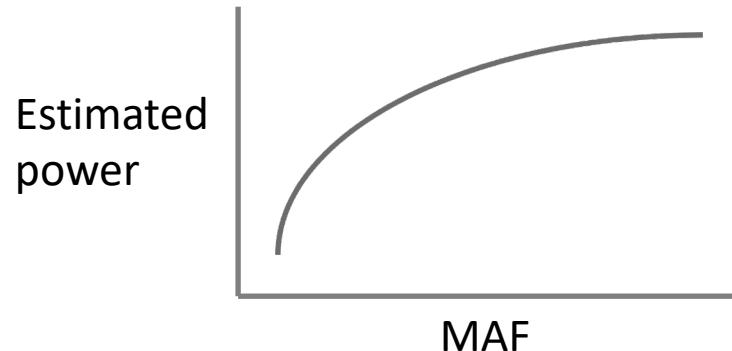
# Linkage disequilibrium (LD)

**Linkage disequilibrium (LD):** a non-random association of alleles at different loci; genotyping data at two loci have some level of correlations



# Genotyping data and filtering

- Typically only bi-allelic markers are used.
- Of two alleles, the allele with a smaller frequency is the minor allele. Its frequency is **minor allele frequency (MAF)**. A MAF cutoff is needed to filter SNPs (e.g., 1%).
- Filter out markers with high missing data (e.g., 30%).
- Imputation can reduce missing data.



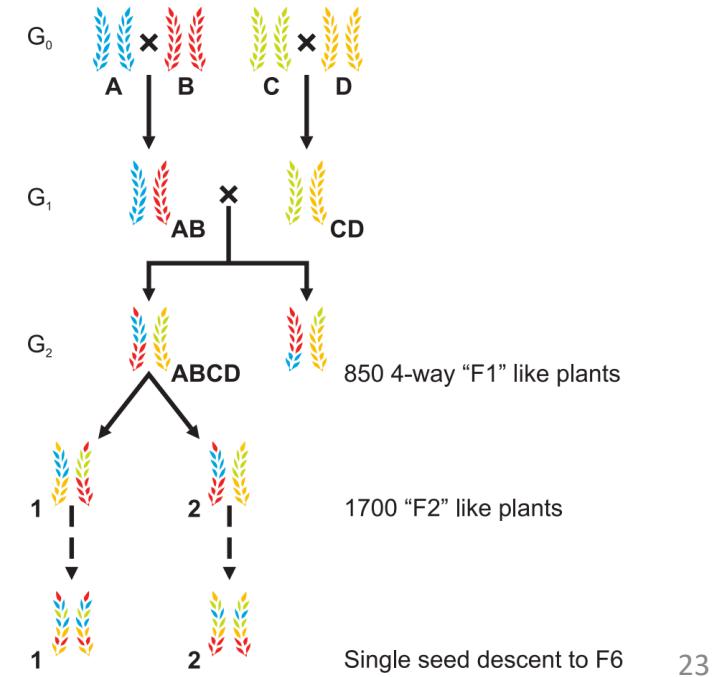
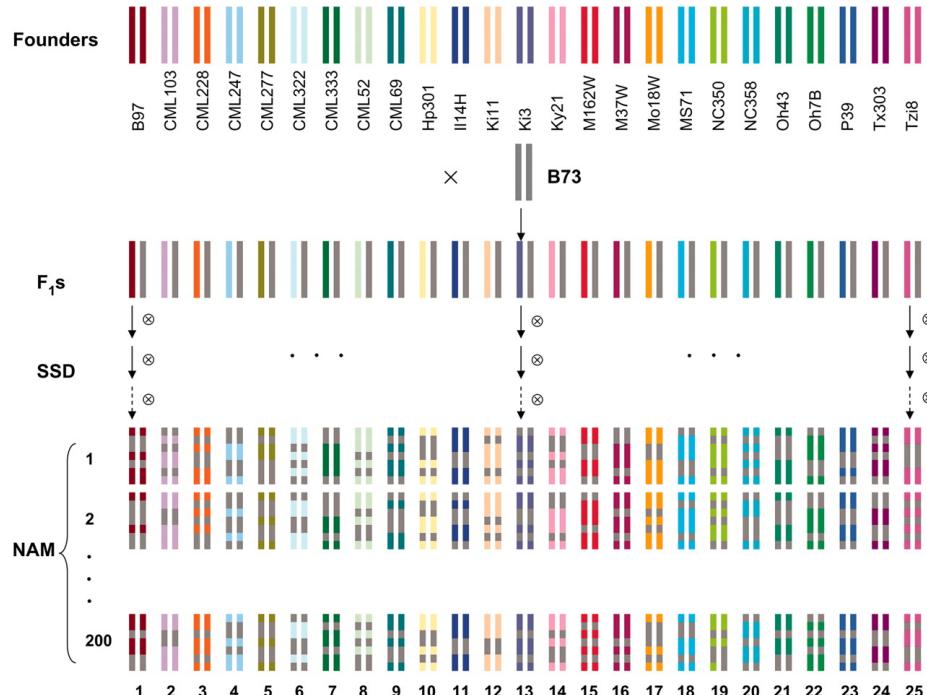
# Mapping populations

- **Natural population**

Diverse individual plant lines/animals/human beings.

- **Multi-parent crosses**

1. Nested association mapping lines (NAM)
2. Multiparent Advanced Generation Inter-Cross (MAGIC)



## Statistical test for each SNP

$$y \sim X\beta + S\alpha + e$$

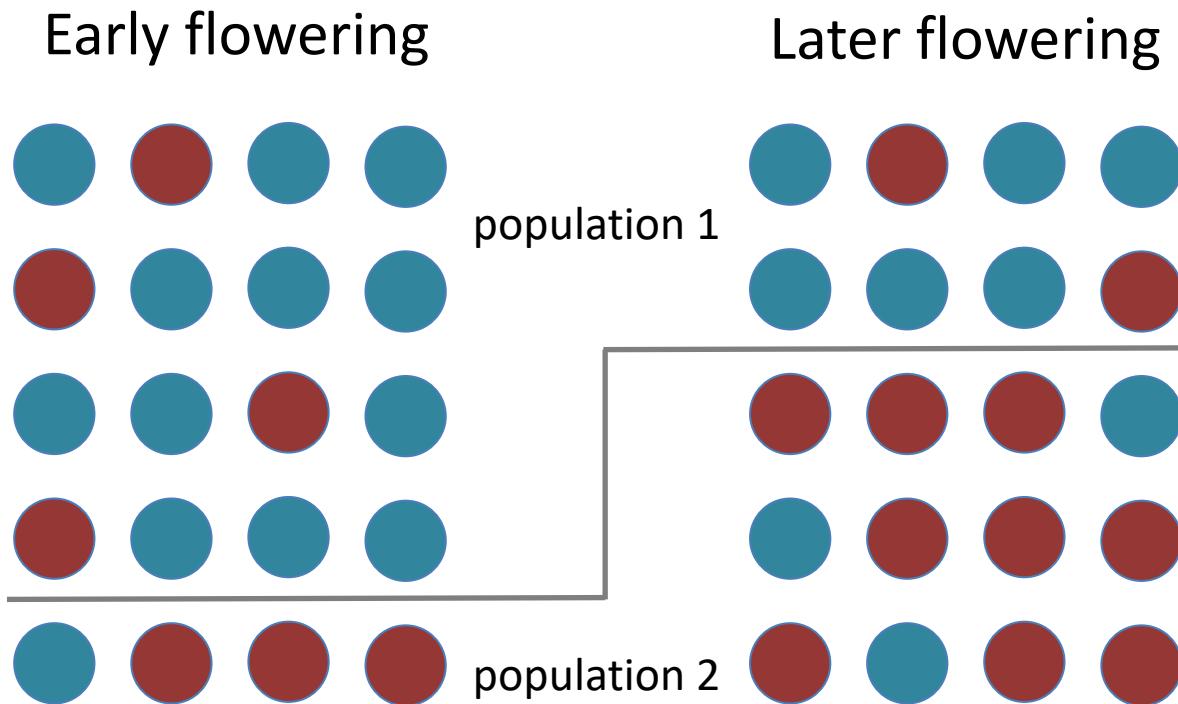
$y$ : trait data

$X\beta$  : all non-variant fixed effect

$S\alpha$  : variant effects

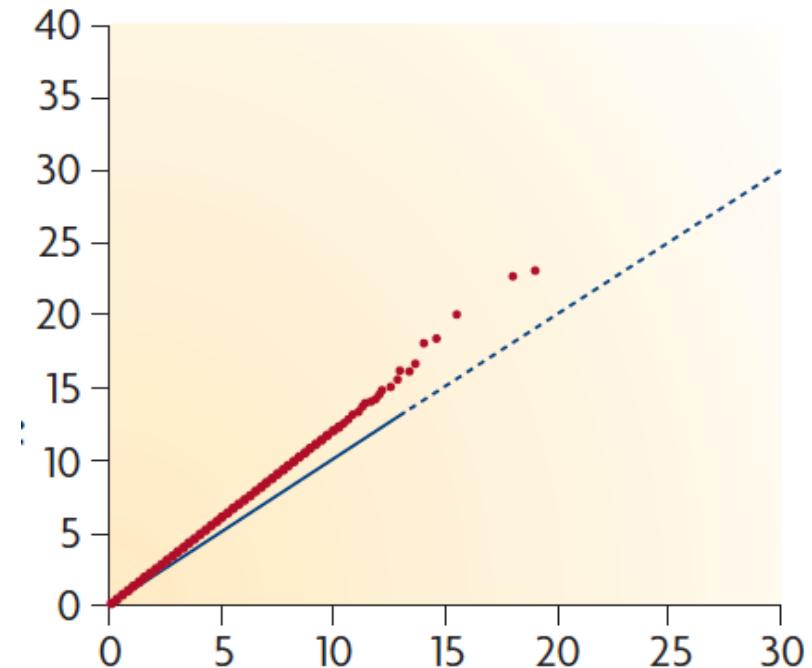
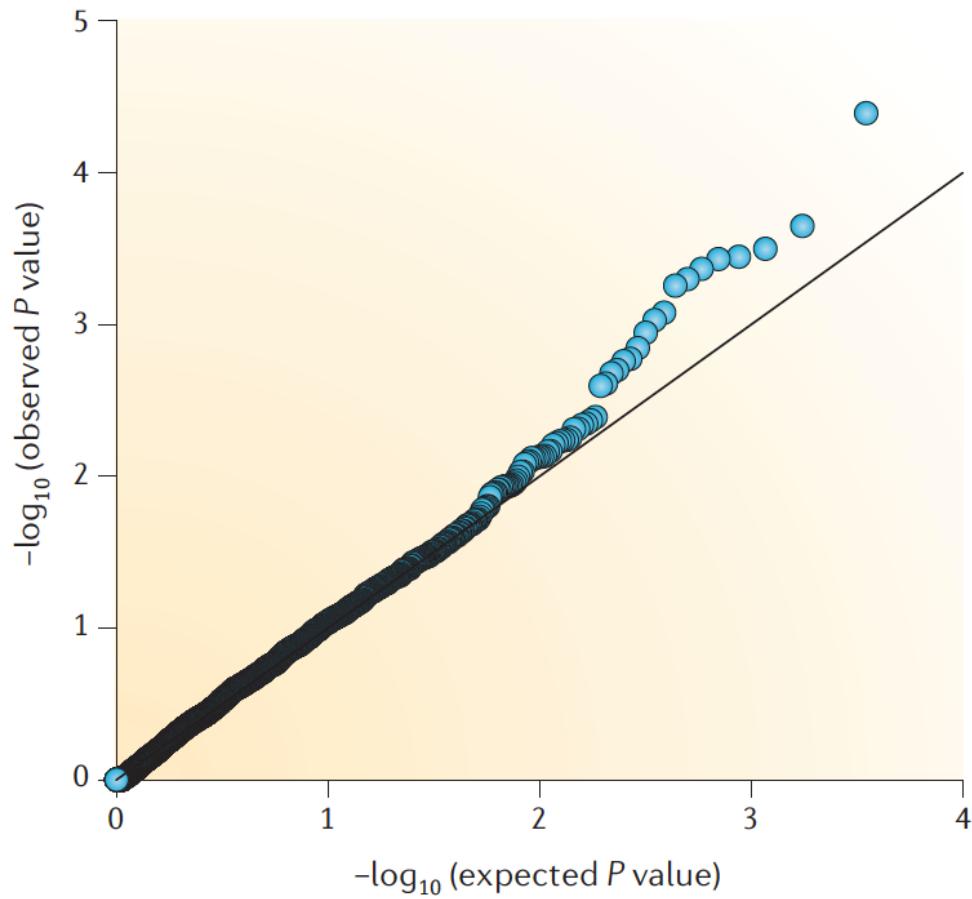
This model is not sufficient to explain phenotypic data.

# Spurious associations



Different proportions of sub-populations in two groups lead to spurious associations.

# quantile-quantile (Q-Q) p-value plot



# Population structure (Q)

## Population structure (Q)

Confounding structure leads to false positive.

- Define a set of markers
- Population structure:
  1. Principal Component Analysis (PCA) ([EIGENSOFT](#))
  2. Distance-based cluster ([R/stats](#))
  3. Model-based clustering ([STRUCTURE](#) )

$$y \sim \underbrace{X\beta + S\alpha + Q\nu}_{\text{Fixed effect}} + e$$

# $Q + K$ model explains more phenotypic variants

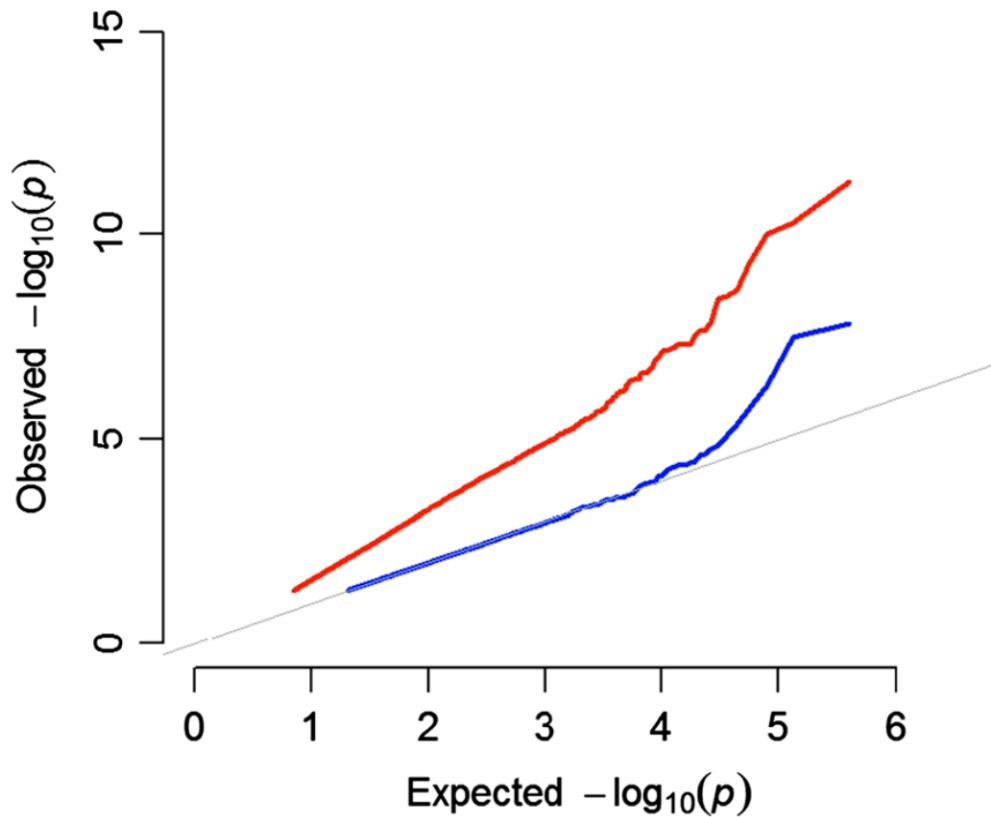
- **Population structure (Q)**
- **Kinship (K) - cryptic relatedness:** The probability that two homologous genes are identical by descent, estimated by using all genotyped markers.

## Mixed linear model (MLM)

$$y \sim X\beta + S\alpha + Qv + \boxed{Zu} + e$$

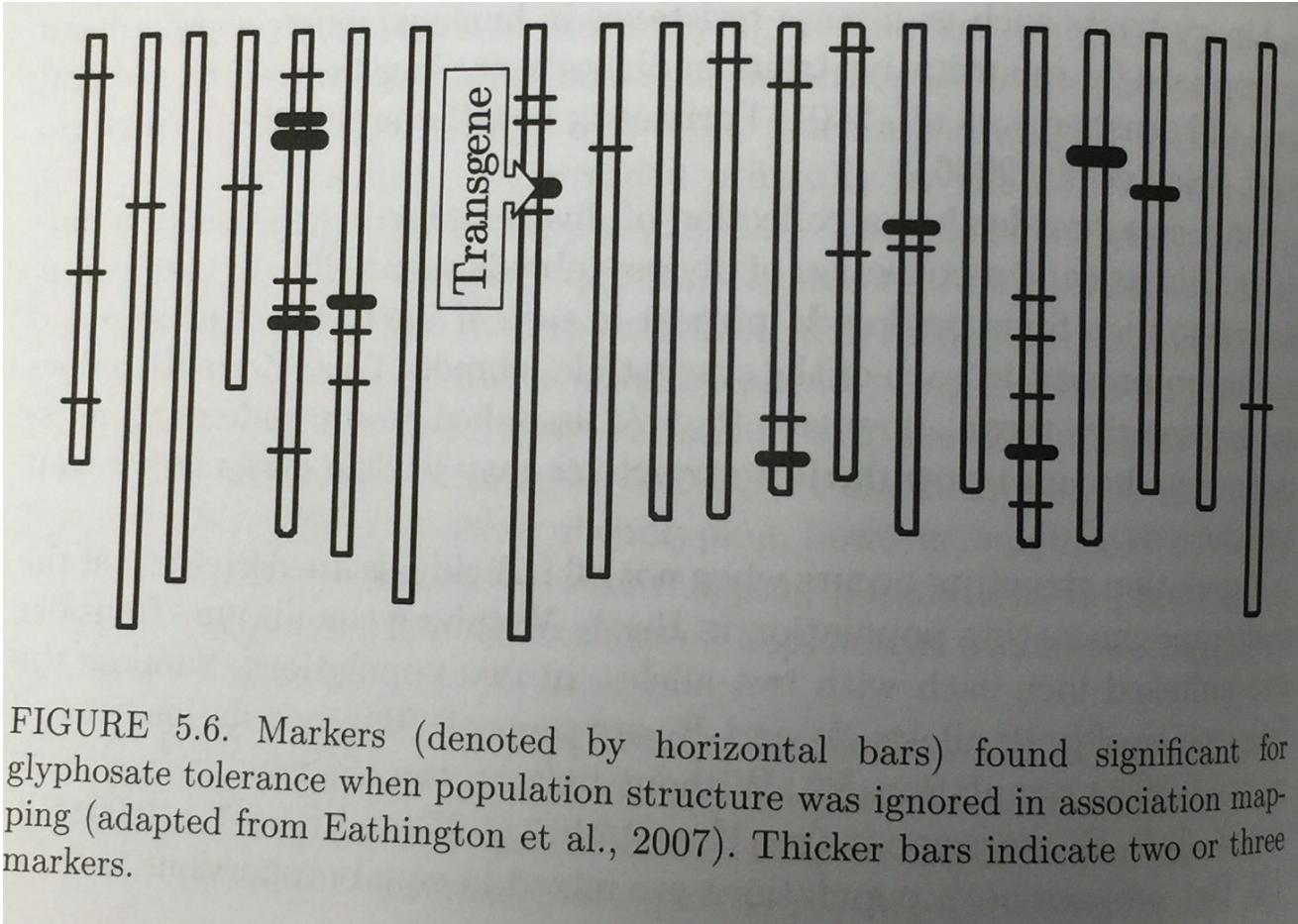
Random effect

# Mixed linear model (Q+K MLM)



The mixed model (blue) dramatically reduces inflation of p-values

# GWAS w/o accounting for population structure

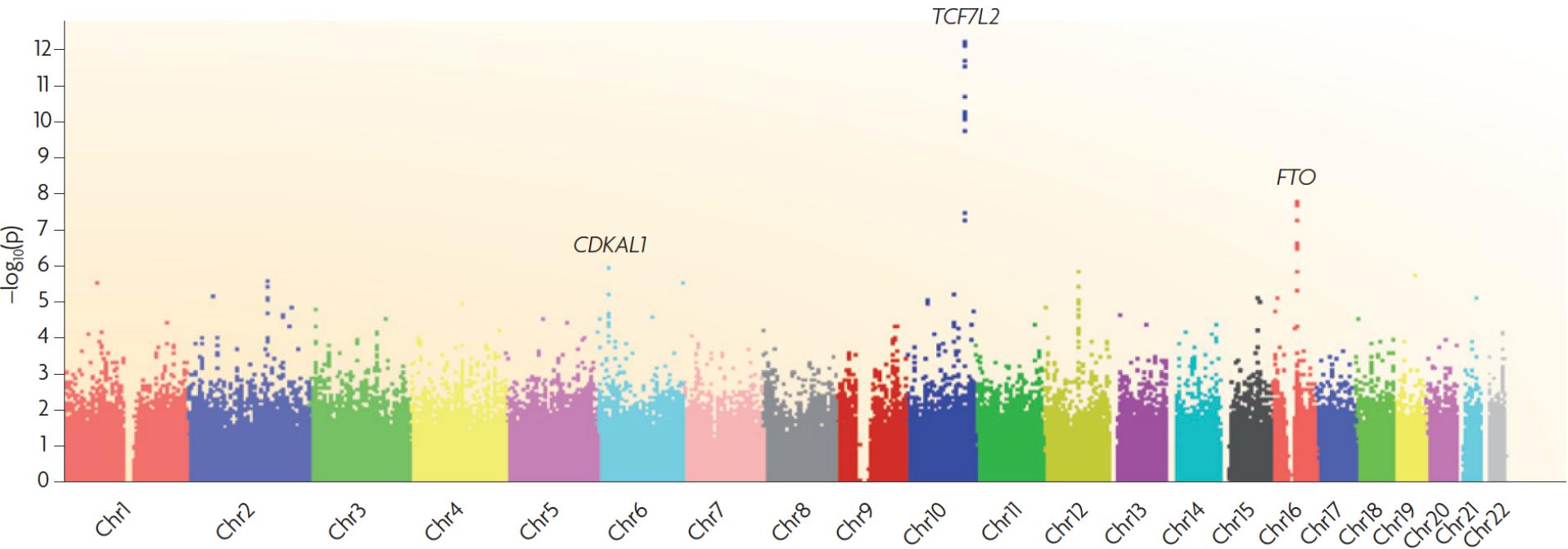


750 soybean  
inbred lines

49 markers on 15  
chromosomes

FIGURE 5.6. Markers (denoted by horizontal bars) found significant for glyphosate tolerance when population structure was ignored in association mapping (adapted from Eathington et al., 2007). Thicker bars indicate two or three markers.

# Manhattan plot



McCarthy et al., Nature Review Genetics, 2008: 9:356-369

association does not imply causation

# GWAS p-value threshold

- $5 \times 10^{-8}$  has become a standard (Human GWAS)
- naive Bonferroni correction (conservative due to the assumption that every genetic variant tested is independent of the rest)
- false discovery rate procedures
- permutation based-approaches
- Bayesian approaches

What is the difference between QTL and GWAS?

# Comparison between QTL and GWAS

<b>Attribute</b>	<b>QTL mapping</b>	<b>Association genetics</b>
Populations	Typically from biparental lines; Limited recombination	from diverse lines, taking advantage of historic recombination
Markers for genome coverage	No high-density markers required	high-density markers required
Resolution	Limited	High