

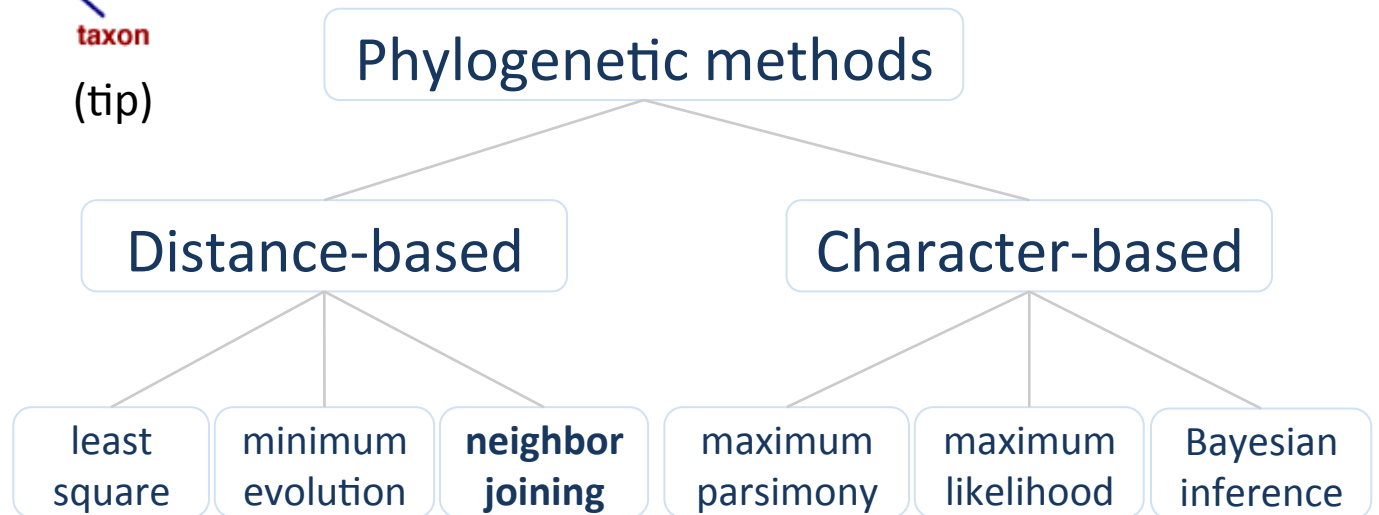
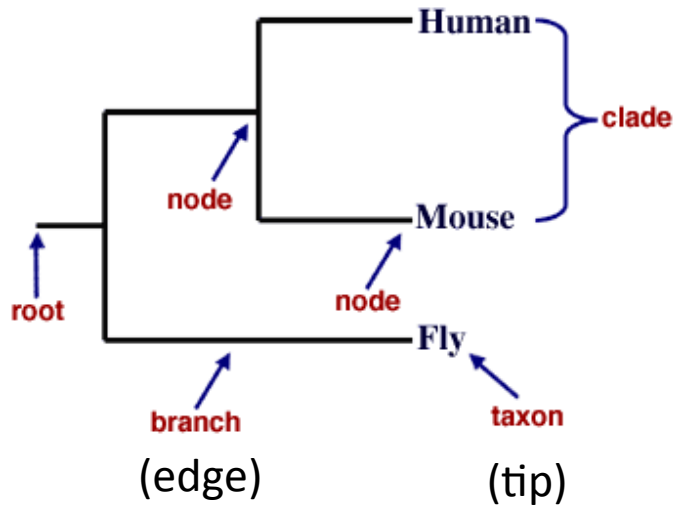
# QTL mapping and GWAS

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

3/28/2016

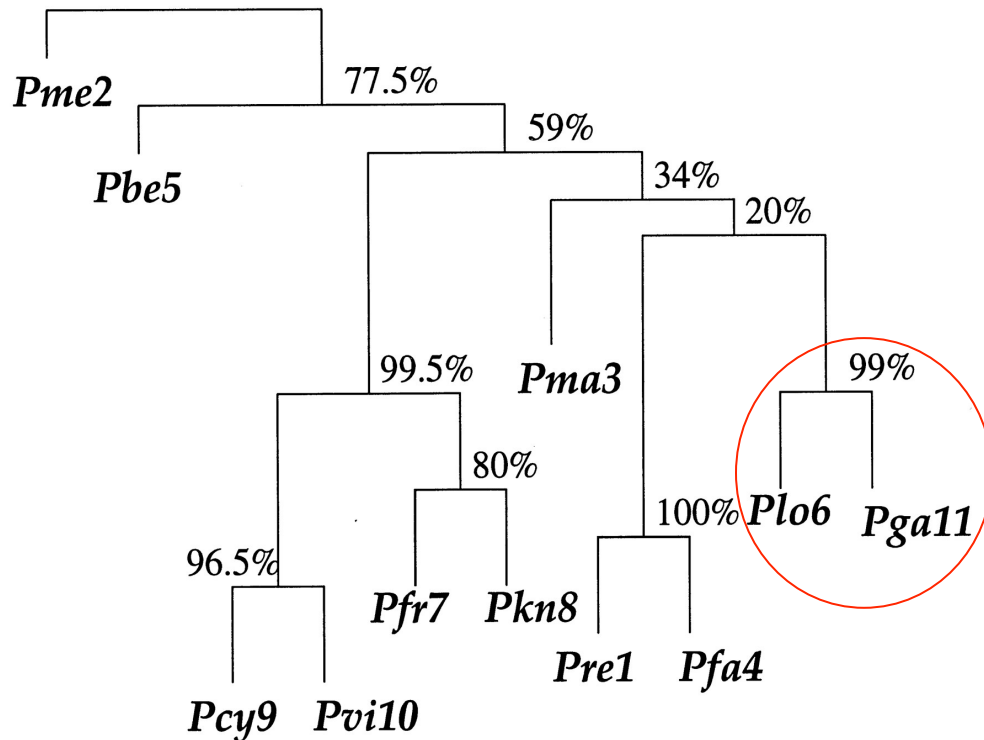
# Phylogenetic trees



# Outgroup rooting

- Many methods (NJ) construct unrooted tree. An outgroup can be introduced to identify the “root”. Although the inferred tree for all species is still unrooted, the root is believed to be located along the branch that leads to the outgroup so that the tree for the ingroup species is rooted. This strategy is called outgroup rooting.
- A good outgroup needs to satisfy:
  1. not a member of the ingroup
  2. close related to the ingroup

# Tree evaluation: Bootstrap analysis



Bootstrapping measures how consistently the data support given taxon bipartitions (Hedges, 1992).

*Plo6* and *Pga11* are grouped together in 99% bootstrap replicates.

\* B = 200 bootstrap replications.

# An R package - ape

Analysis of Phylogenetics and Evolution ("ape") is an R software package for use in molecular evolution and phylogenetics.

**Table 1.** Special functions available in APE 1.1

Application	Available commands
Input/output	read.dna, write.dna, read.nexus, write.nexus, read.tree, write.tree, read.GenBank
Graphics	add.scale.bar, plot.mst, plot.phylo, plot.skyline, lines.skyline, ltt.plot
Tree manipulation	bind.tree, drop.tip, is.binary.tree, is.ultrametric
Comparative method	compar.gee, compar.lynch, pic, vcv.phylo
Diversification	birthdeath, cherry, diversi.gof, diversi.time, gamma.stat
Population genetics	branching.times, coalescent.intervals, collapsed.intervals, find.skyline.epsilon, heterozygosity, skylineplot, skyline, theta.h, theta.k, theta.s
Molecular dating	chronogram, ratogram, NPRS.criterion
Miscellaneous	all.equal.phylo, balance, base.freq, dist.dna, dist.gene, dist.phylo, GC.content, klastorin, mantel.test, mst, summary.phylo
Data sets	bird.families, bird.orders, hivtree, landplants, opsin, woodmouse, xenarthra

# Outline

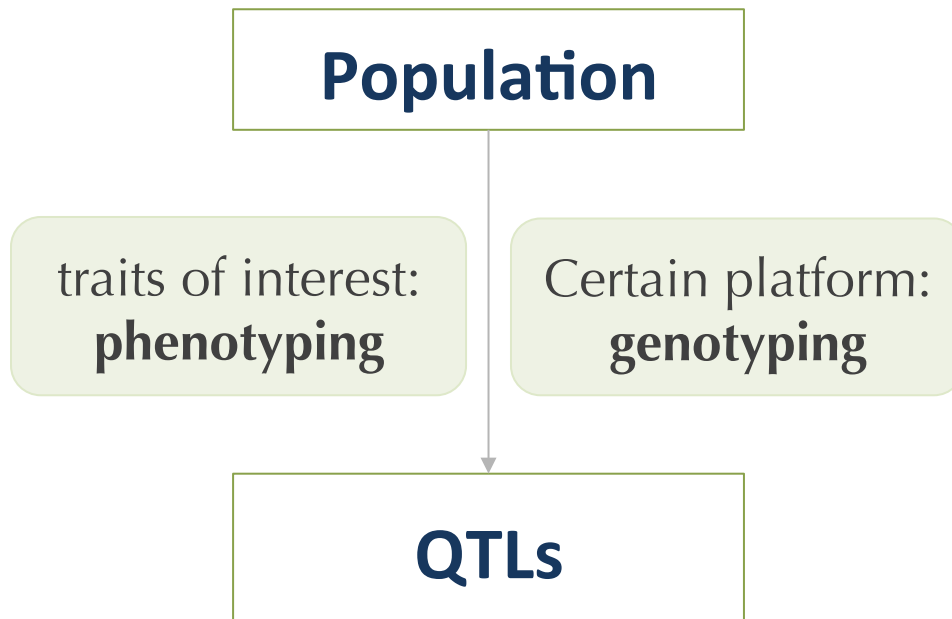
- QTL mapping
- Genome-wide association study (GWAS)

What is the goal to perform QTL or GWAS?

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

# 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



ATCGCTGCCGATCTGCGT**C**ATACGGAATCGTCGGCTTTCAG  
ATCGCTGCCGATCTGCGT**G**ATACGGAATCGTCGGCTTTCAG  
ATCGCTGCCGATCTGCGT**C**ATACGGAATCGTCGGCTTTCAG  
ATCGCTGCCGATCTGCGT**G**ATACGGAATCGTCGGCTTTCAG  
ATCGCTGCCGATCTGCGT**G**ATACGGAATCGTCGGCTTTCAG  
ATCGCTGCCGATCTGCGT**G**ATACGGAATCGTCGGCTTTCAG  
ATCGCTGCCGATCTGCGT**C**ATACGGAATCGTCGGCTTTCAG  
ATCGCTGCCGATCTGCGT**C**ATACGGAATCGTCGGCTTTCAG  
ATCGCTGCCGATCTGCGT**G**ATACGGAATCGTCGGCTTTCAG  
ATCGCTGCCGATCTGCGT**C**ATACGGAATCGTCGGCTTTCAG

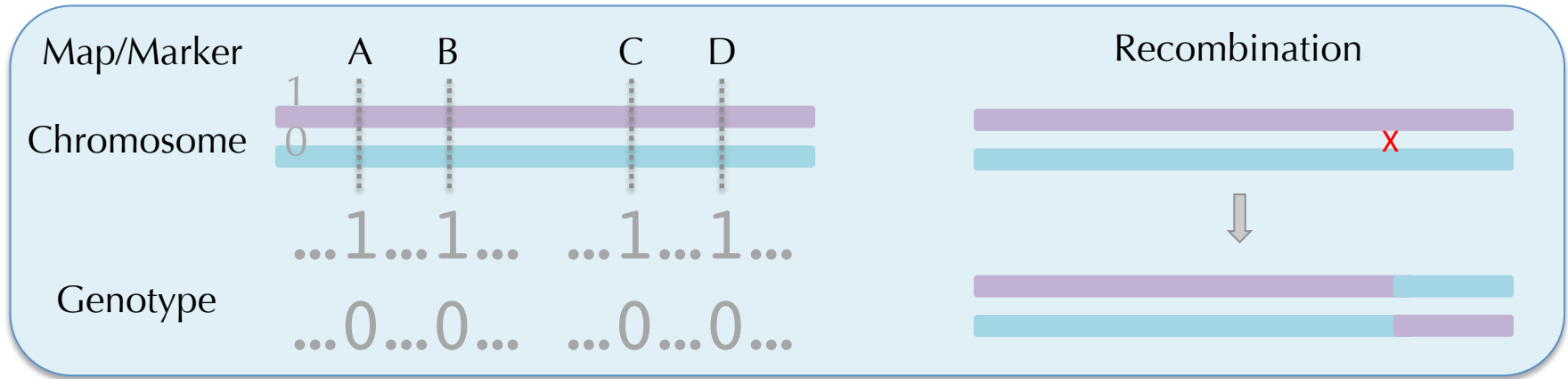
Genotyping score

-----**C**/**G**-----  
-----**1**/**0**-----

Marker



# Mapping a causal genetic controlling component (X)



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

Mapping result

A B **X** C D

# 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:**

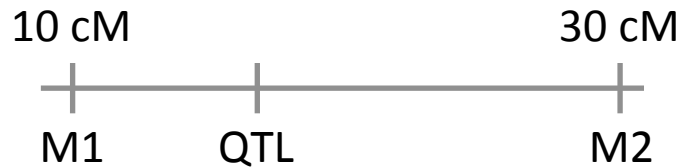
- Individuals with missing data are excluded
- Suffers in low density markers

## 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



Genotype

?	?	0
0	0 w/high prob	0
0	?	?
0	?	1
1	1 w/high prob	1
1	?	?
1	1 w/high prob	1
1	?	0
?	?	0
...	...	...

## Estimate genotypes:

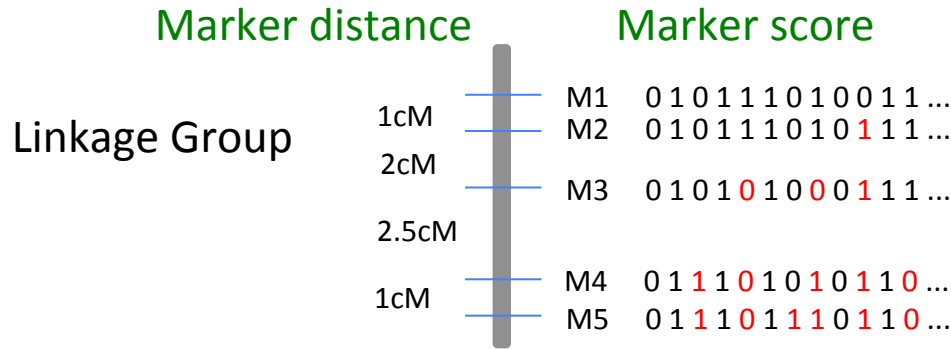
each estimated genotype is associated with a certain probability

1. Genetic map
2. Mapping function

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

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

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

# Mapping function

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

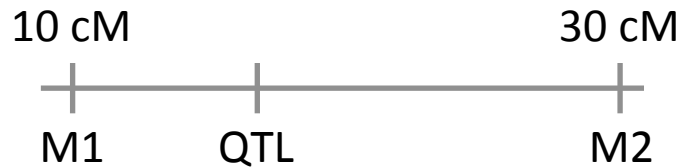
$$r = \frac{1}{2}(1 - e^{-2d})$$

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

$r$  = recombination rate (0-0.5)

$d$  = distance in Morgans

# Interval mapping – estimate genotypes



Genotype

?	?	0
0	<b>0 w/high prob</b>	0
0	?	?
0	?	1
1	<b>1 w/high prob</b>	1
1	?	?
1	<b>1 w/high prob</b>	1
1	?	0
?	?	0
...	...	...

## Estimate genotypes:

each estimated genotype is associated with a certain probability

1. Genetic map
2. Mapping function

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

# Estimate likelihood of a QTL model

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

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

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

- ***No QTL Likelihood***

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



# LOD (logarithm of the odds)

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

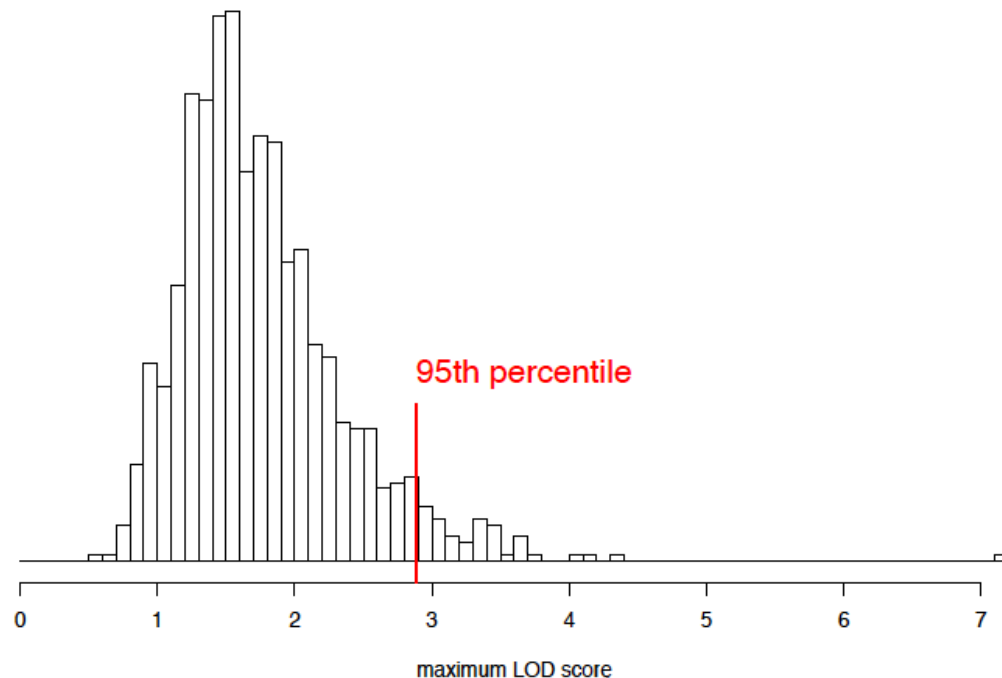
LOD =  **$\log_{10}$  likelihood ratio**, comparing a single-QTL model to the “no QTL anywhere”.

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

# 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 ( $\text{maxLOD}_1, \text{maxLOD}_2, \dots, \text{maxLOD}_{1000}$ )
- The 95<sup>th</sup> percentile of MaxLOD is a genome-wide LOD threshold.



# Question

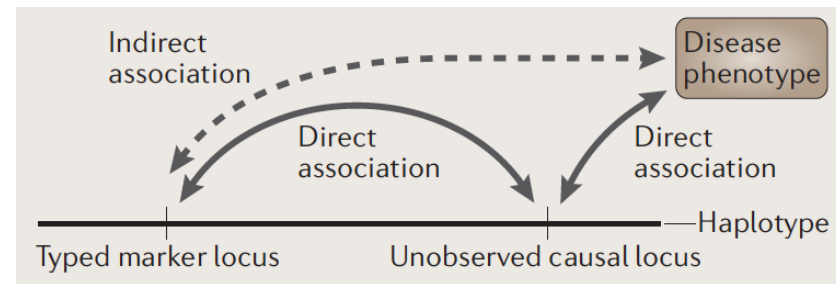
Can we perform a QTL study on a human population?

# 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

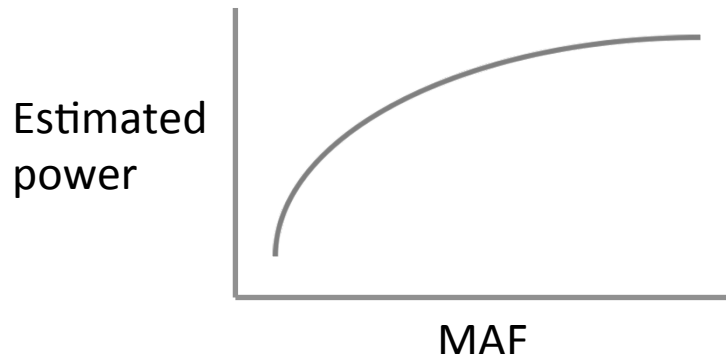
**Linkage disequilibrium (LD):** a non-random association of alleles at different loci; genotyping data at two loci have some level of correlations



Balding et al., Nature Review Genetics, 2006, 7:781

# 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



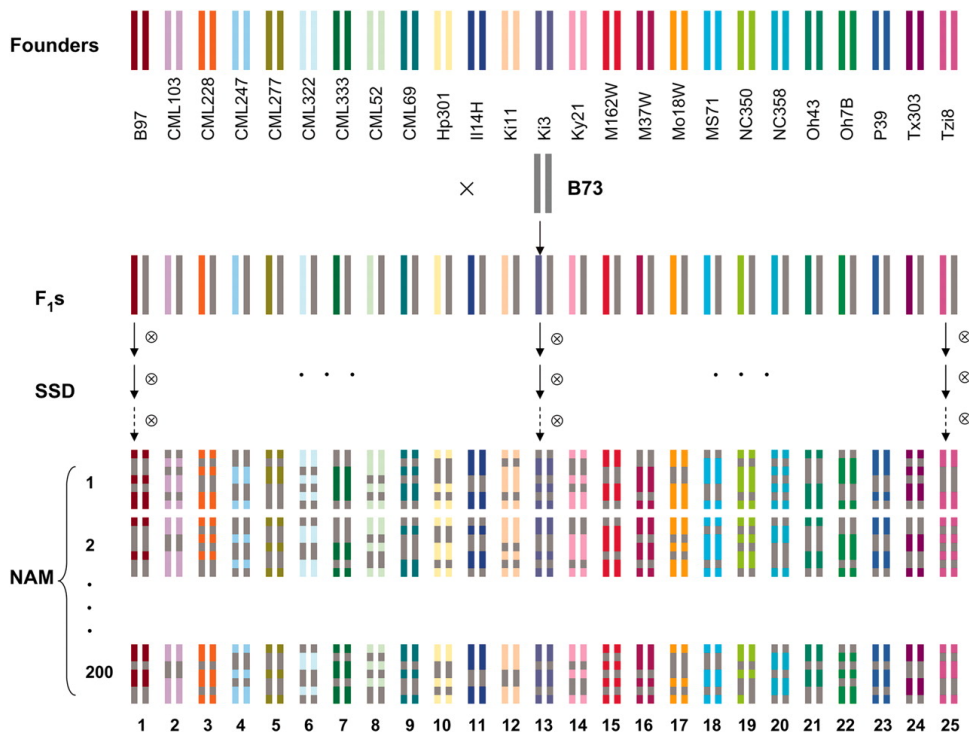
# Mapping populations

- **Natural population**

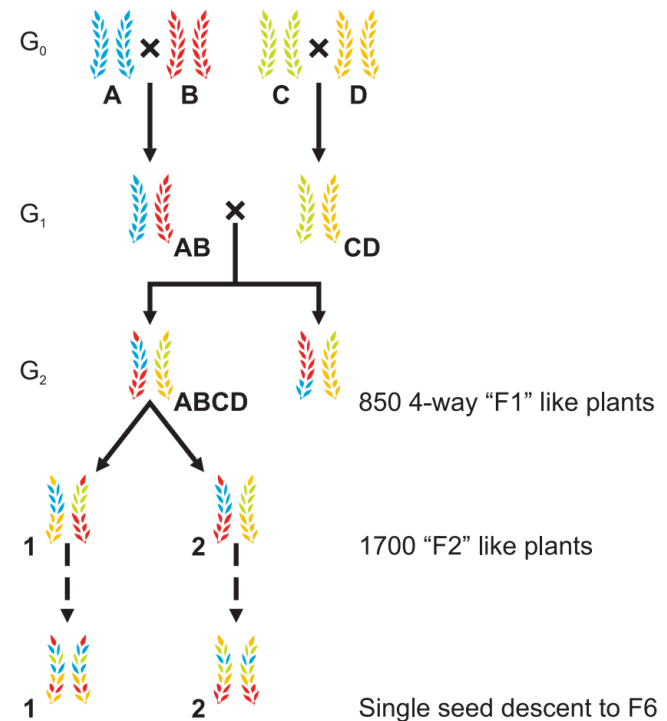
Diverse individual plant lines/animals/human beings.

- **Multi-parent crosses**

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



Yu et al., Genetics 2008;178:539-551



Huang et al., Plant Biotechnology Journal 2012; 10:826-839

## Statistical test for each SNP

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

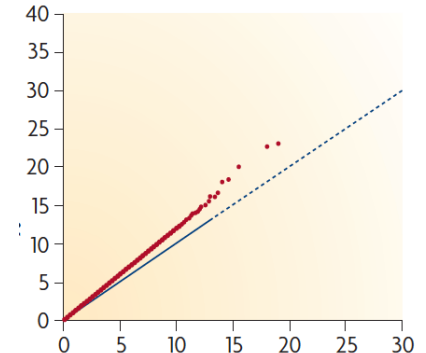
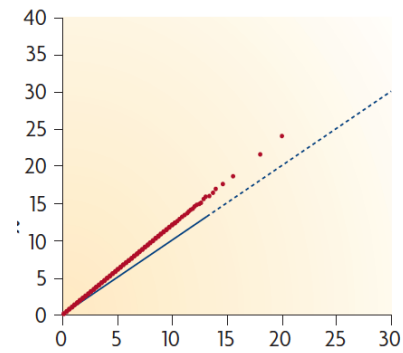
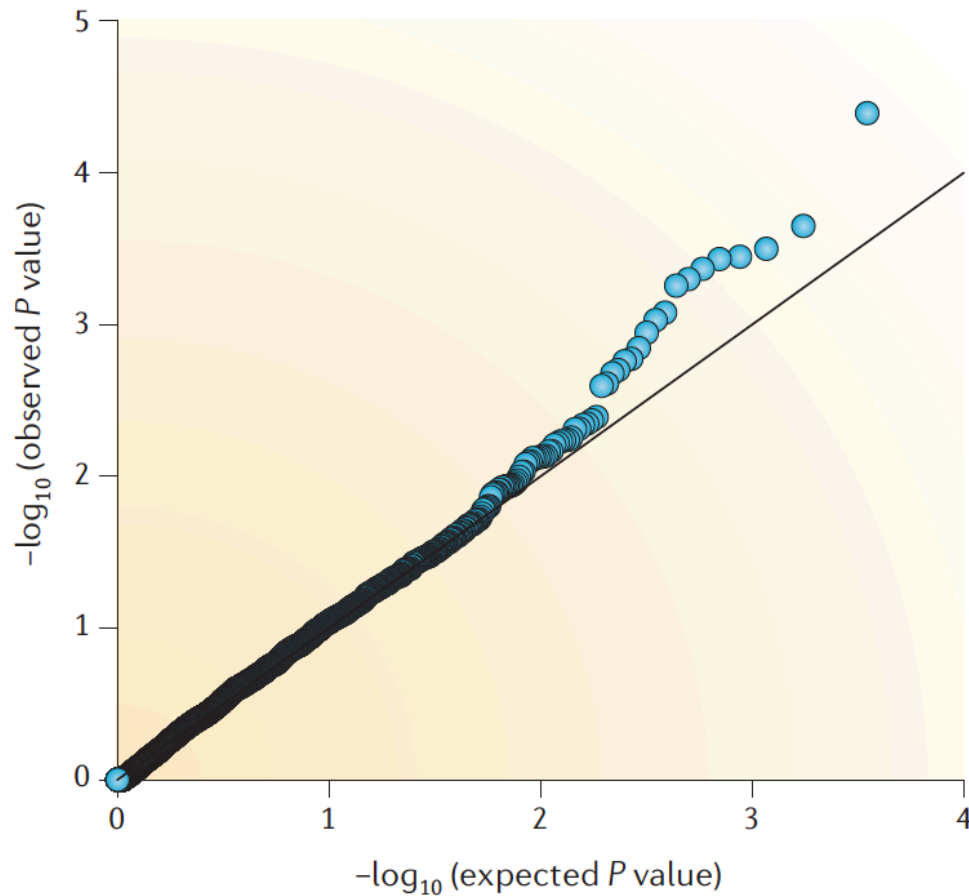
$y$ : trait data

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

$S\alpha$  : variant effects

This simple model is not sufficient to explain phenotypic data.

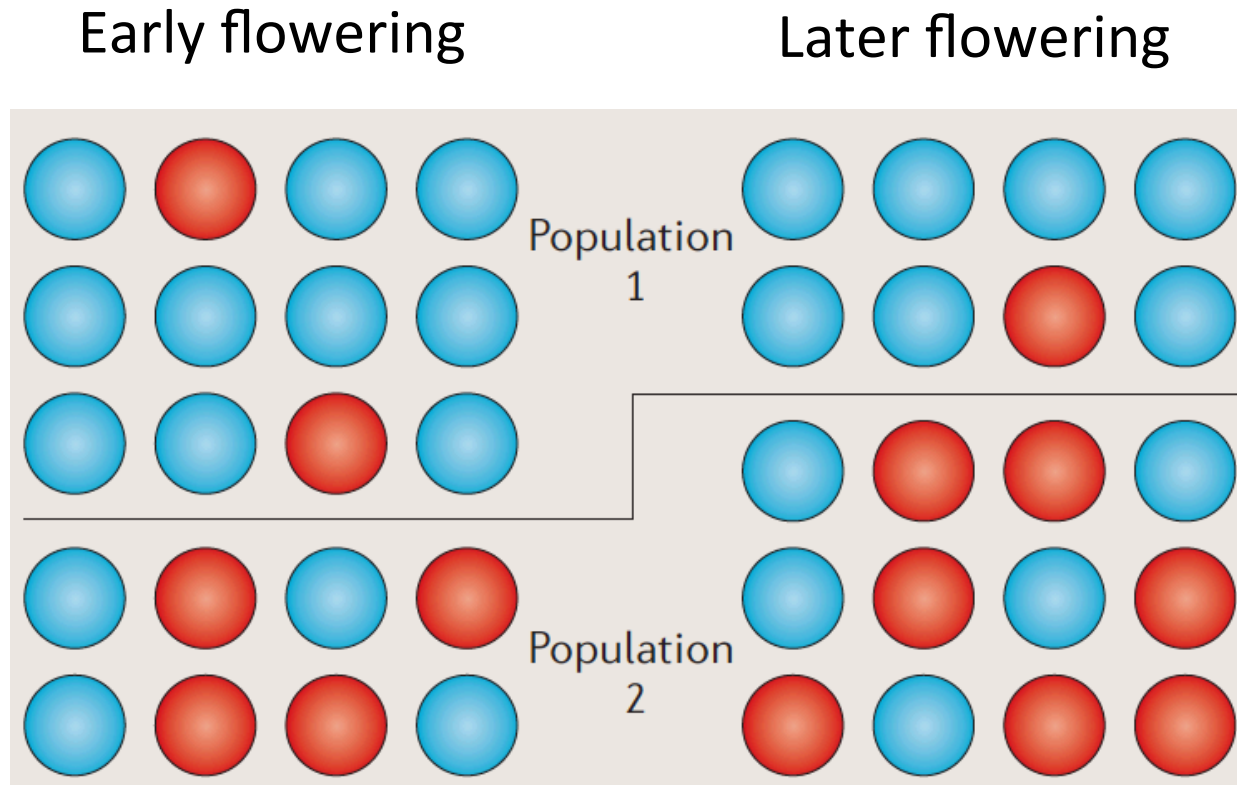
# quantile-quantile (Q-Q) p-value plot



Balding et al., Nature Review Genetics, 2006, 7:781



# Spurious associations



Flowering time is confounded with Populations

Modified from Balding et al., Nature Review Genetics, 2006, 7:781


# Population structure (Q)

## Population structure (Q)

Confounding structure leads to false positive.

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

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

  
Fixed effect

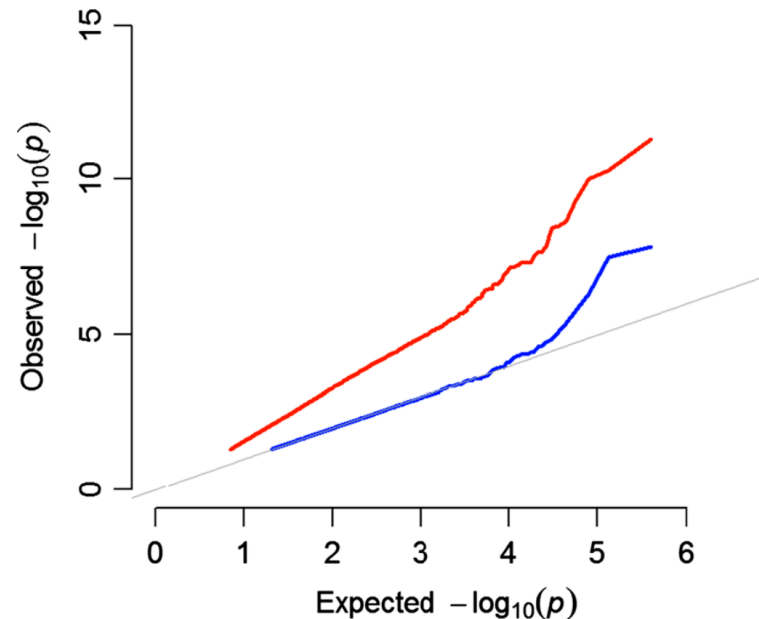
# Q + K model explains more phenotypic variants

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

**Mixed** linear model (MLM)

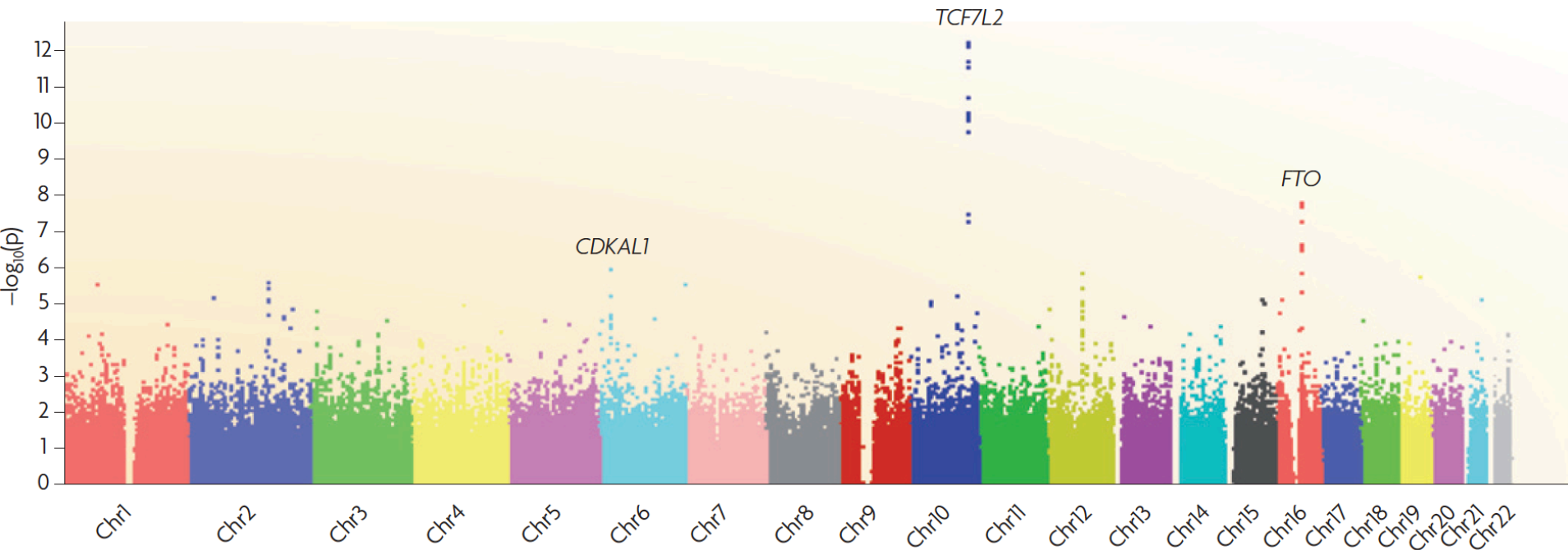
$$y \sim X\beta + S\alpha + Qv + \boxed{Zu} + e$$

Random effect



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

# Manhattan plot



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

association does not imply causation

What is the difference between QTL and GWAS?

# Comparison between QTL and GWAS

Attribute	QTL mapping	Association genetics
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