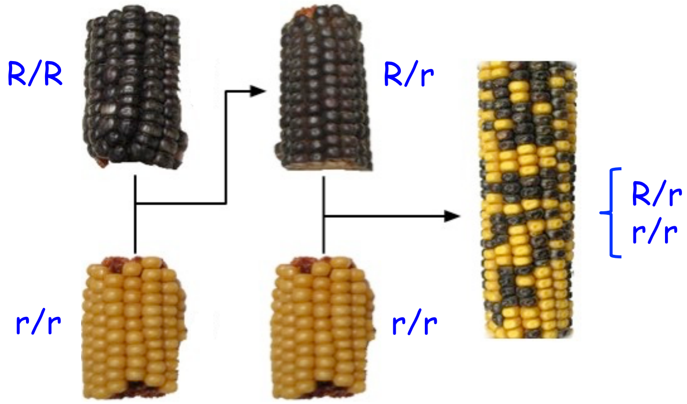
Lecture 9
QTL and Association Mapping

Guilherme J. M. Rosa

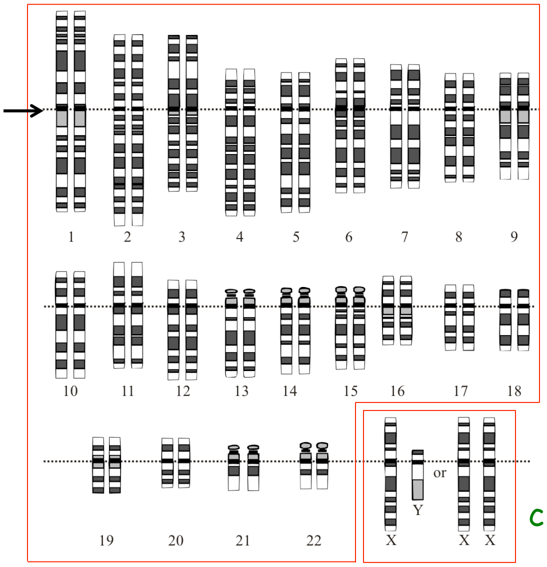
University of Wisconsin-Madison

Introduction to Quantitative Genetics
SISG, Seattle
16 – 18 July 2018

Linkage Analysis and QTL Mapping



Human Genome, Chromosomes



Centromeres →

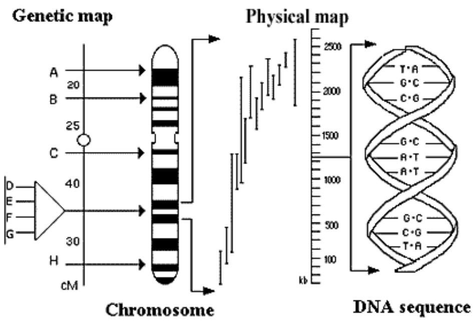
Autosomes

Mitochondrial DNA not shown

Sex chromosomes

Graphical representation of the idealized human diploid karyotype

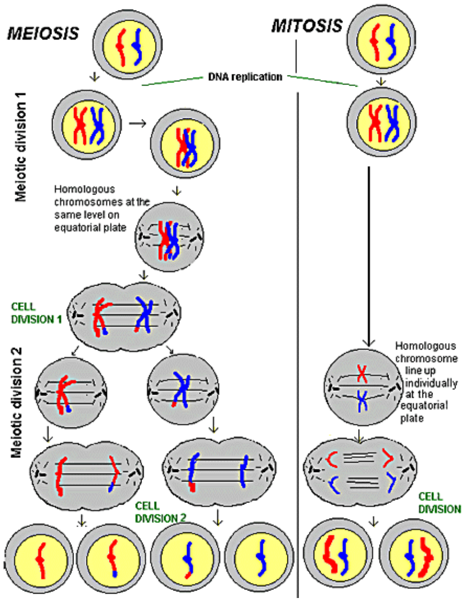
Sequences of Base Pairs Mapping



Genetic maps: relative positions of loci in chromosomes or linkage groups. Distances in genetic maps are measured in centimorgans (cM, about 1 million base pairs)

Physical maps: overlapping collections of DNA fragments (measured in kilobases, kb) which are assembled together to build the base-by-base sequence of DNA

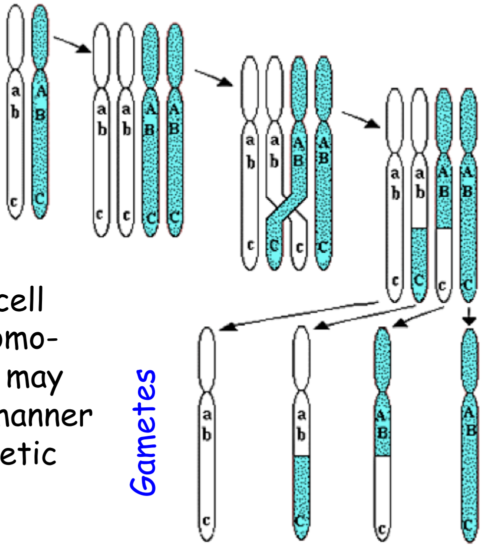
Comparison of Meiosis and Mitosis


MEIOSIS
MITOSIS
DNA replication
Meiotic division 1
Homologous chromosomes at the same level on equatorial plate
CELL DIVISION 1
Meiotic division 2
CELL DIVISION 2
Homologous chromosome line up individually at the equatorial plate
CELL DIVISION

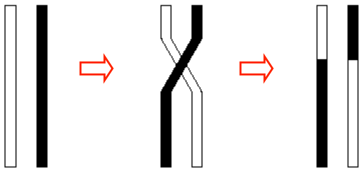
Crossing-Over and Recombination During Meiosis

In meiosis, the precursor cells of the sperm or ova must multiply and at the same time reduce the number of chromosomes to one full set.

During the early stages of cell division in meiosis, two chromosomes of a homologous pair may exchange segments in the manner shown above, producing genetic variations in germ cells.



Crossing Over and Recombination



An odd number of crossovers between two loci results in a recombination between them

Because crossing over takes place at random, the probability of recombination (r) is higher for loci that are farther apart than for loci that are closer to each other

$$0 \leq r \leq 0.5$$

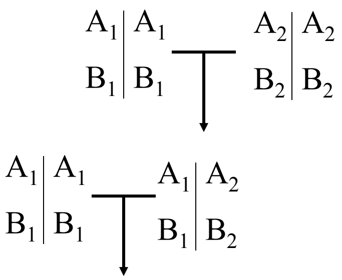
completely linked loci

unlinked loci

Two Point Linkage Analysis

- ⇨ Backcross experiment
- ⇨ Genotypic information for two loci (A and B)
- ⇨ Estimate the recombination rate r_{AB}
- ⇨ Are these two loci linked?

Individual	A	B
1	0	0
2	0	1
⋮	⋮	⋮
n	1	1



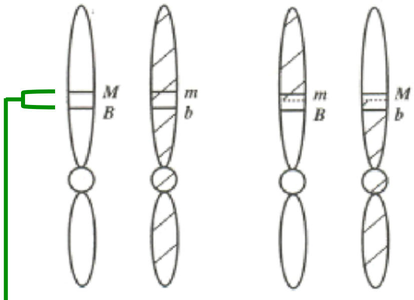
Four possible genotypes

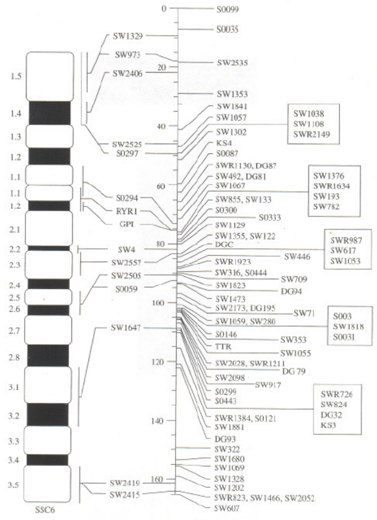
Two Point Linkage Analysis

- ⇨ Suppose n = 80 and y = 16 (recombinants)
- ⇨ Point estimate of r_{AB} : $\hat{r}_{AB} = \frac{y}{n} = 0.20$
- ⇨ Confidence interval (95%) of r_{AB} :

$$\mathrm{CI}(r_{AB};\ 95\%) = [0.1189;\ \ 0.3044]$$

Recombination Rate and Linkage Map

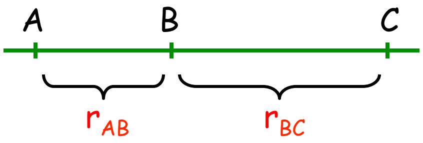




Estimates of recombination rates between pairs of markers are used to order markers and to infer their genetic distances (centimorgans; cM)

Interference

⇨ Lack of independence in recombinations at different intervals on a chromosome



- If r_{AB} and r_{BC} are independent, the probability of double recombination is $Pr(DR) = r_{AB} \times r_{BC}$
- If r_{AB} and r_{BC} are not independent, the above probability is given by $Pr(DR) = c \times r_{AB} \times r_{BC}$ where c is called "coefficient of coincidence"
- Interference: $I = 1 - c$

Map Distance

The map distance x between two loci, in Morgan units, is defined as the expected number of crossovers between them

Unlike recombination rates, map distances are additive

The relationship between map distances and recombination rates is discussed next

Map Functions

Map functions provide a transformation from map distance to recombination rate. Two approaches have been used to derive map functions:

In the first case, a probability model is assumed for the number of crossovers in an interval of length *x*. Then, recombination rate is calculated as the probability of an odd number of crossovers in the interval

In the second approach, recombination events in two adjacent intervals are modeled, allowing for interference

Examples of map functions: Haldane, Binomial, Kosambi

Haldane Map Function

Haldane (1919) suggested that the number of crossovers in any chromosomal interval follows a Poisson distribution, with no interference

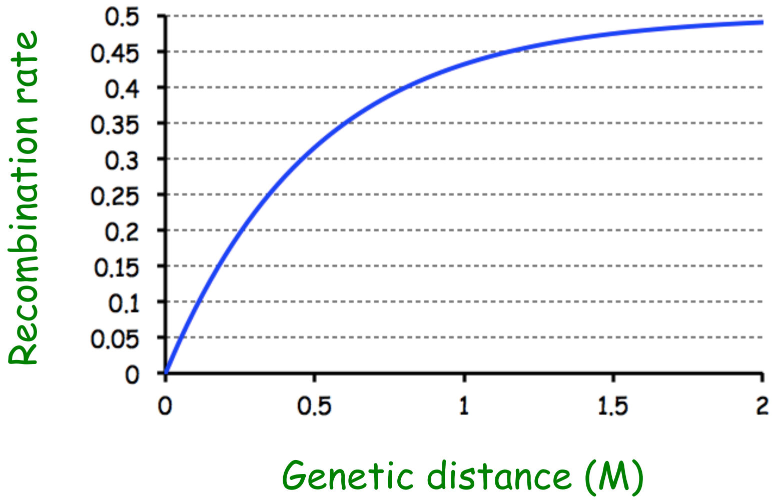
If P_k is the probability of k crossovers, then the probability of recombination (r) is $r = P_1 + P_3 + P_5 + \ldots$

This leads to the Haldane's map function:

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

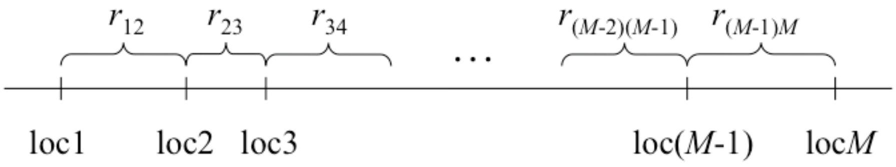
The inverse of which is: $x = \begin{cases} -\frac{1}{2}\ln(1-2r) & , \text{ if } 0 \le r < 0.5 \\ \infty & , \text{ if } r = 0.5 \end{cases}$

Haldane Map Function



Multipoint Point Linkage Analysis

- ⇨ Instead of two loci, suppose there are M loci
- ⇨ If order is unknown: M!/2 alternatives



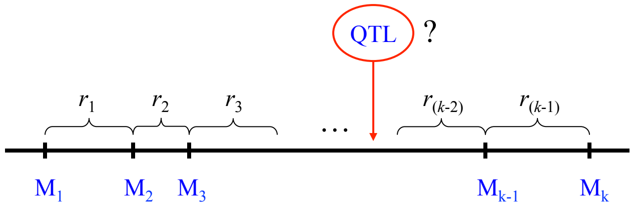
Goal: Determine the order of the loci and estimate recombination fractions between neighboring loci, i.e. "Map Construction"

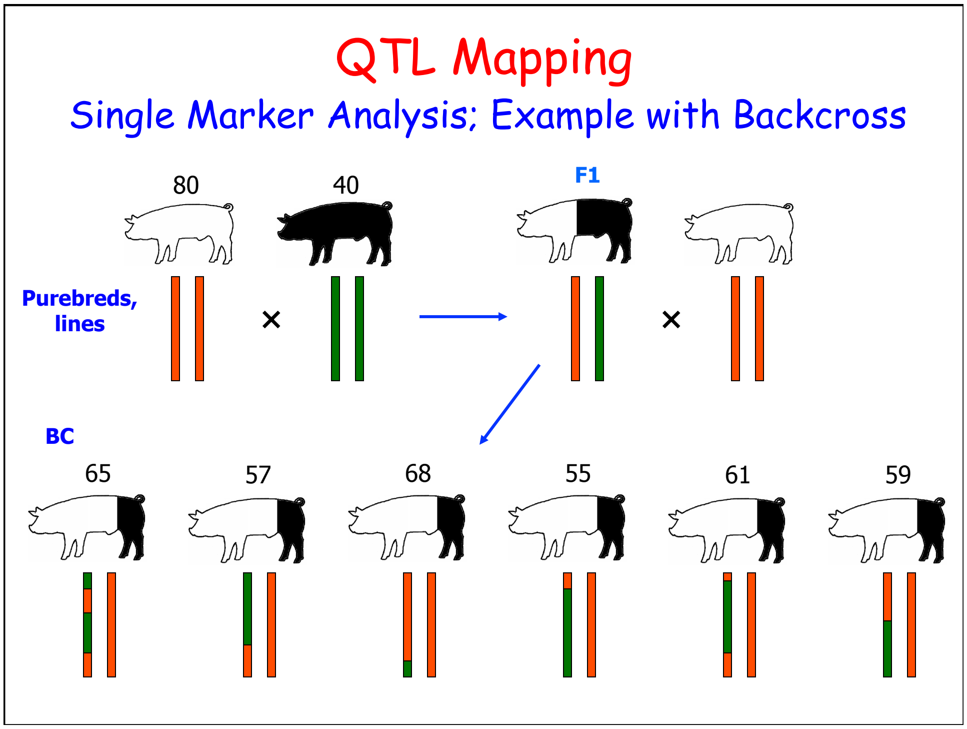
Methods for Mapping QTL

- ⇨ Single Marker Analysis
- ⇨ Interval Mapping
- ⇨ Composite Interval Mapping
- ⇨ Bayesian Methods

QTL Mapping

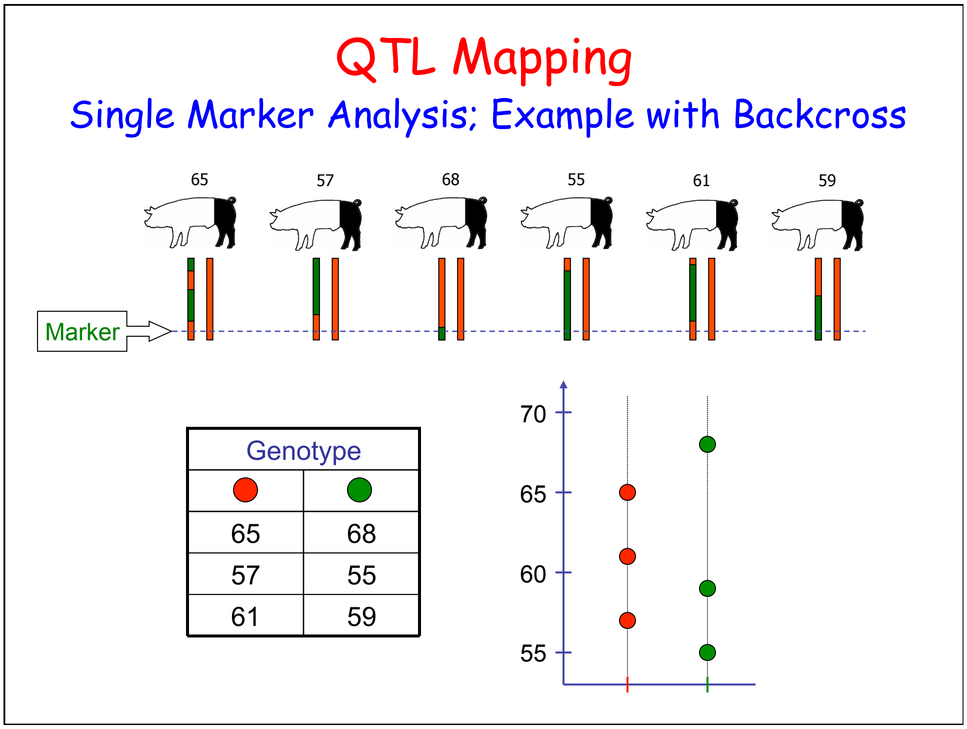
- ⇨ Methods based on linkage disequilibrium between markers and QTL (line crossing or segregating population)
- ⇨ Requirements:
  - ① Linkage (marker) maps
  - ② Variation for the quantitative trait





QTL Mapping

Single Marker Analysis; Example with Backcross



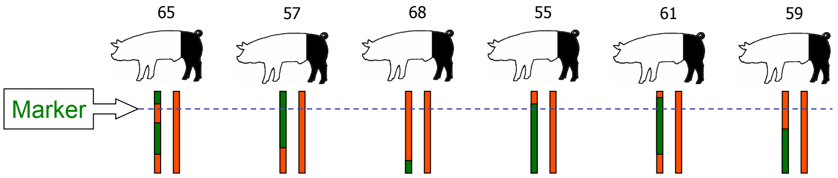
QTL Mapping

Single Marker Analysis; Example with Backcross

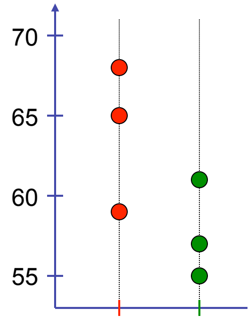
Genotype	
🔴	🟢
65	68
57	55
61	59

QTL Mapping

Single Marker Analysis; Example with Backcross

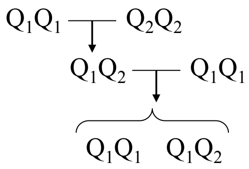


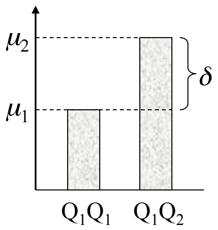
Genotype	
🔴	🟢
65	57
68	55
59	61



Single Marker Analysis

☞ Simple example with candidate gene and BC population





Genotype	Obs.	Mean	STD
Q_1Q_1	n_1	m_1	s_1
Q_1Q_2	n_2	m_2	s_2

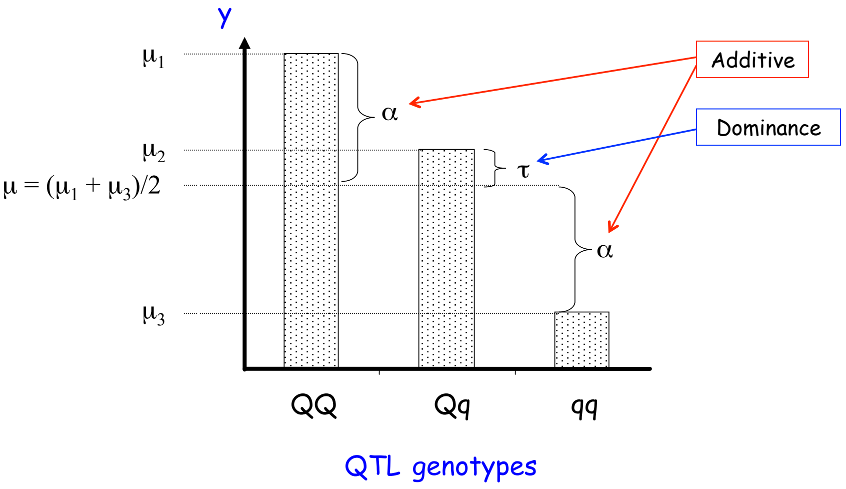
⇨ H_0: $\delta = 0$ vs H_1: $\delta \neq 0$

$$t = \frac{m_1 - m_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)}$$

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$CI[\delta; (1-\alpha)]:\ (m_2 - m_1) \pm t_{(n_1+n_2-2;\alpha/2)}\sqrt{\frac{s^2}{n_1 + n_2 - 2}}$$

Example with F2 Population

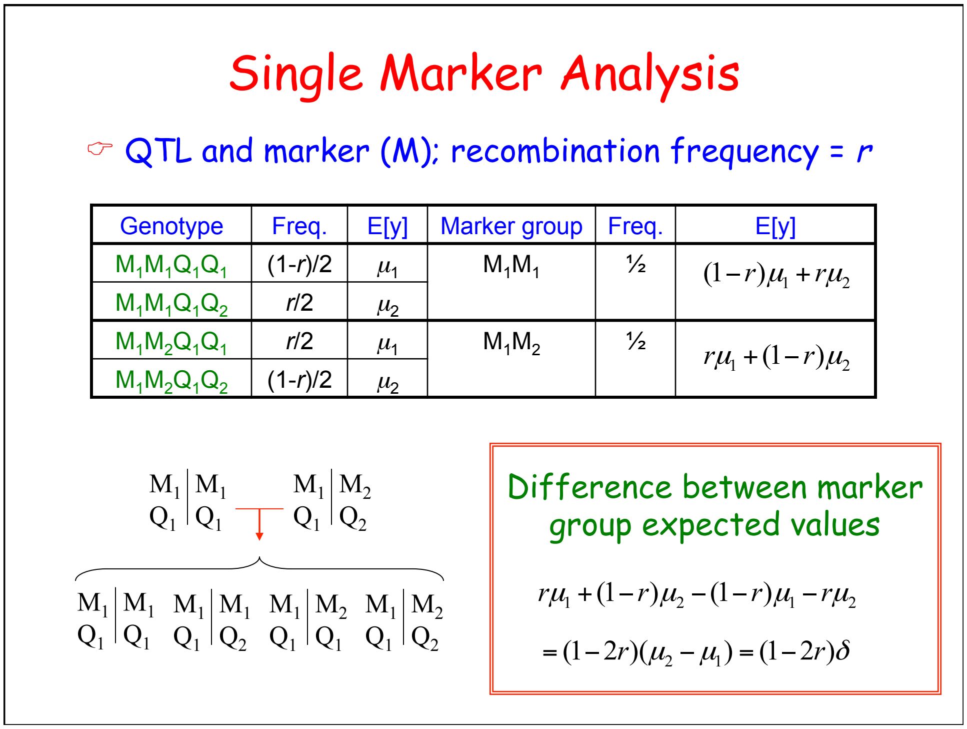


Example with F2 Population

Candidate gene

Information on phenotypes and genotypes for a specific marker

Marker Genotype	Phenotype (8 individuals per group)
MM	95.9, 108.0, 96.5, 92.9 101.0, 94.5, 93.7, 89.8
Mm	105.2, 107.9, 89.9, 113.4 109.7, 102.4, 97.1, 107.1
mm	117.1, 95.2, 106.4, 104.7 92.5, 123.9, 97.8, 100.5



Single Marker Analysis

☞ QTL and marker (M); recombination frequency = *r*

Genotype	Freq.	E[y]	Marker group	Freq.	E[y]
$M_1M_1Q_1Q_1$	(1-*r*)/2	μ_1	M_1M_1	½	$(1-r)\mu_1 + r\mu_2$
$M_1M_1Q_1Q_2$	*r*/2	μ_2			
$M_1M_2Q_1Q_1$	*r*/2	μ_1	M_1M_2	½	$r\mu_1 + (1-r)\mu_2$
$M_1M_2Q_1Q_2$	(1-*r*)/2	μ_2			

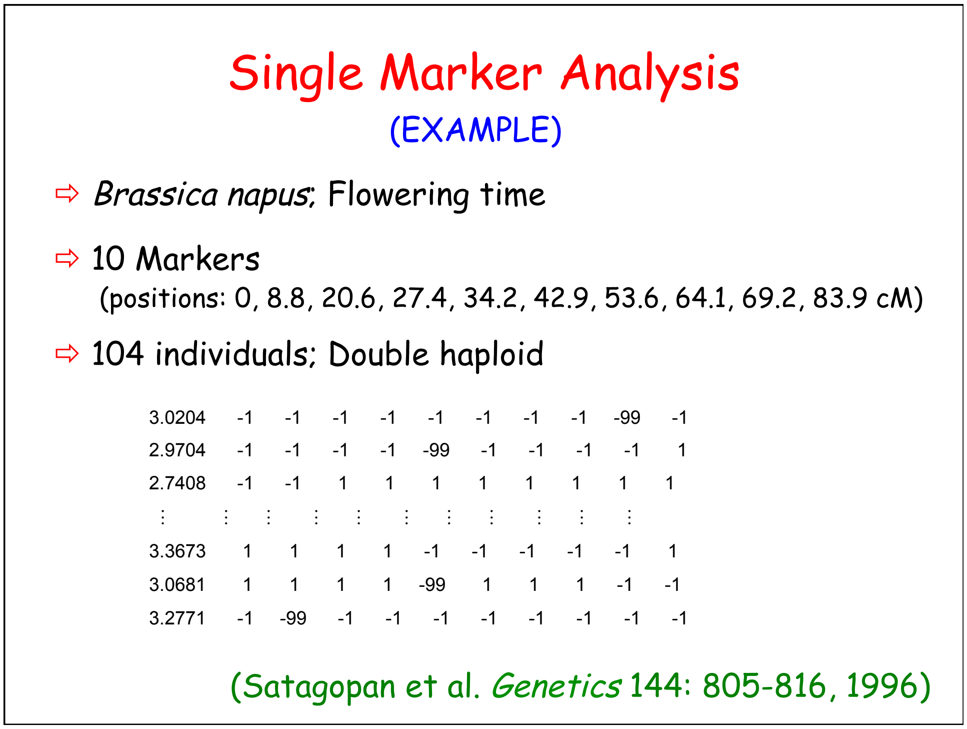
$$\frac{M_1}{Q_1}\Big|\frac{M_1}{Q_1} \times \frac{M_1}{Q_1}\Big|\frac{M_2}{Q_2}$$

$$\frac{M_1}{Q_1}\Big|\frac{M_1}{Q_1} \quad \frac{M_1}{Q_1}\Big|\frac{M_1}{Q_2} \quad \frac{M_1}{Q_1}\Big|\frac{M_2}{Q_1} \quad \frac{M_1}{Q_1}\Big|\frac{M_2}{Q_2}$$

Difference between marker group expected values

$$r\mu_1 + (1-r)\mu_2 - (1-r)\mu_1 - r\mu_2$$

$$= (1-2r)(\mu_2 - \mu_1) = (1-2r)\delta$$



Single Marker Analysis

(EXAMPLE)

⇨ *Brassica napus*; Flowering time

⇨ 10 Markers

(positions: 0, 8.8, 20.6, 27.4, 34.2, 42.9, 53.6, 64.1, 69.2, 83.9 cM)

⇨ 104 individuals; Double haploid

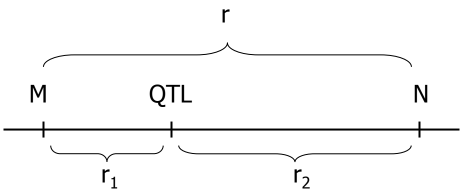
3.0204	-1	-1	-1	-1	-1	-1	-1	-1	-99	-1
2.9704	-1	-1	-1	-1	-99	-1	-1	-1	-1	1
2.7408	-1	-1	1	1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.3673	1	1	1	1	-1	-1	-1	-1	-1	1
3.0681	1	1	1	1	-99	1	1	1	-1	-1
3.2771	-1	-99	-1	-1	-1	-1	-1	-1	-1	-1

(Satagopan et al. *Genetics* 144: 805-816, 1996)

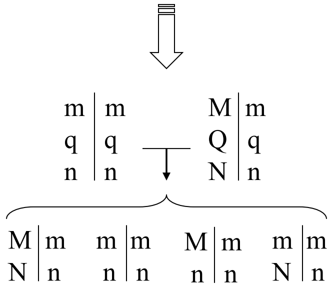
Chrom.	Marker	μ	τ	LRT	F	p-value
1	1	3.184	-0.202	9.379	9.624	0.002 **
1	2	3.204	-0.230	11.378	11.789	0.001 ***
1	3	3.232	-0.266	14.706	15.485	0.000 ***
1	4	3.229	-0.259	13.885	14.562	0.000 ***
1	5	3.240	-0.276	15.554	16.446	0.000 ****
1	6	3.259	-0.307	19.518	21.041	0.000 ****
1	7	3.252	-0.302	19.747	21.312	0.000 ****
1	8	3.257	-0.318	23.450	25.775	0.000 ****
1	9	3.258	-0.330	25.156	27.884	0.000 ****
1	10	3.252	-0.362	31.518	36.059	0.000 ****

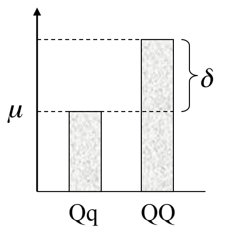
Interval Mapping

(Lander & Botstein, 1989)



Backcross





$$y_i = \mu + q_i\delta + \varepsilon_i$$

phenotype

QTL genotype

residual

$$q_i = \begin{cases} 0 \text{ , if qq} \\ 1 \text{ , if Qq} \end{cases}$$

Interval Mapping

If $\varepsilon_i \sim N(0,\sigma^2) \longrightarrow y_i \,|\, q_i \sim N(\mu + q_i\delta, \sigma^2)$

$$\boxed{p(y_i \,|\, q_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu - q_i\delta)^2\right\}}$$

$$L(\mu,\delta,\sigma^2,\lambda,\boldsymbol{q} \,|\, \boldsymbol{y}) \propto \prod_{i=1}^{N}\left[f(y_i \,|\, q_i = 0)\Pr(q_i = 0) + f(y_i \,|\, q_i = 1)\Pr(q_i = 1)\right]$$

$$L(\mu,\delta,\sigma^2,\lambda,\boldsymbol{q} \,|\, \boldsymbol{y}) \propto \prod_{i=1}^{N}\left[\frac{1}{\sqrt{\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\}\Pr(q_i = 0 \,|\, \lambda) + \frac{1}{\sqrt{\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu - \delta)^2\right\}\Pr(q_i = 1 \,|\, \lambda)\right]$$

QTL position

Interval Mapping

Pr(q_i|λ) is modeled in terms of recombinations between flanking markers and QTL:

Marker Genotypes	Pr(q_i = QQ)	Pr(q_i = Qq)
M,N	$(1 - r_1)(1 - r_2)/(1 - r)$	$r_1 r_2/(1 – r)$
M,n	$(1 - r_1) r_2 / r$	$r_1 (1 - r_2)/ r$
m,N	$r_1 (1 - r_2)/ r$	$(1 - r_1) r_2 / r$
m,n	$r_1 r_2 /(1 - r)$	$(1 - r_1)(1 - r_2)/(1 - r)$

Approximation:

(no double recombination)

Markers	Pr(q_i = QQ)	Pr(q_i = Qq)
M,N	1	0
M,n	(1 - p)	p
m,N	p	(1 - p)
m,n	0	1

$p = \frac{r_1}{r}$

Interval Mapping

➪ Likelihood estimation: EM algorithm to estimate parameters, including λ (position of QTL)

➪ Alternatively: Fix λ (grid search) and evaluate LOD

$$\mathrm{LOD}_\lambda = \log_{10}\left[\frac{\mathrm{L}(\hat{\mu},\hat{\delta},\hat{\sigma}^2,\hat{\boldsymbol{q}}\mid \boldsymbol{y})}{\mathrm{L}(\hat{\mu},\hat{\sigma}^2,\hat{\boldsymbol{q}}\mid \boldsymbol{y},\delta=0)}\right]$$

☞ A QTL is detected whenever the LOD score gets larger than a threshold; estimated position of the QTL maximizes LOD

Interval Mapping

REGRESSION APPROACH

(Haley & Knott, 1992)

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \\ \vdots & \vdots \\ p_{N1} & p_{N2} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

alternatively

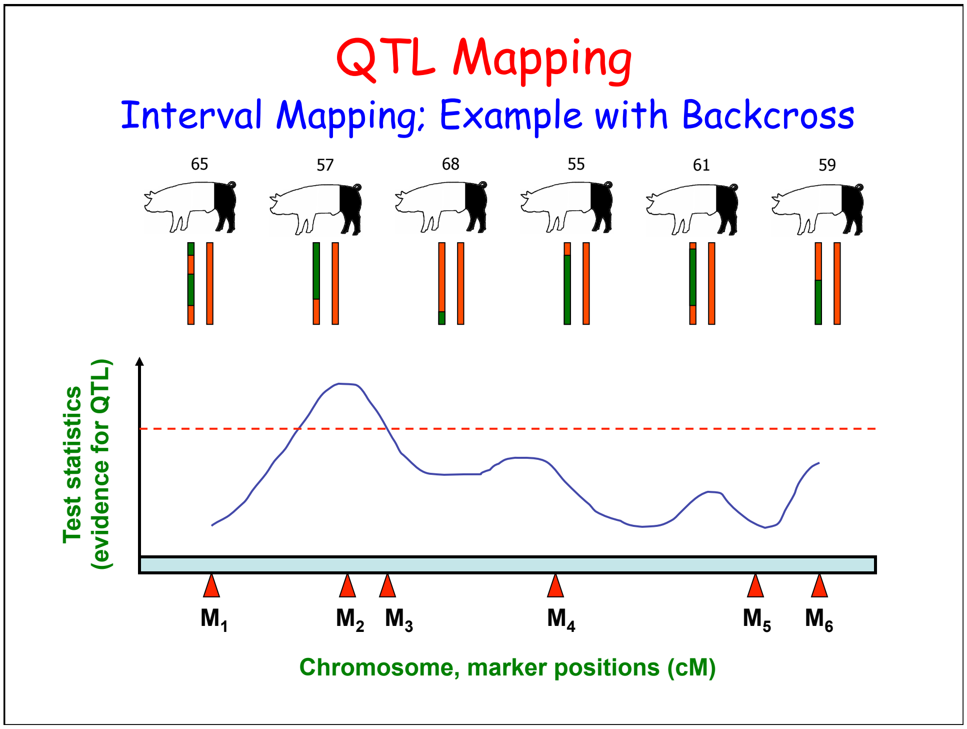
$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & p_{12} \\ 1 & p_{22} \\ \vdots & \vdots \\ 1 & p_{N2} \end{bmatrix} \begin{bmatrix} \mu \\ \delta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Residual Sum of Squares:

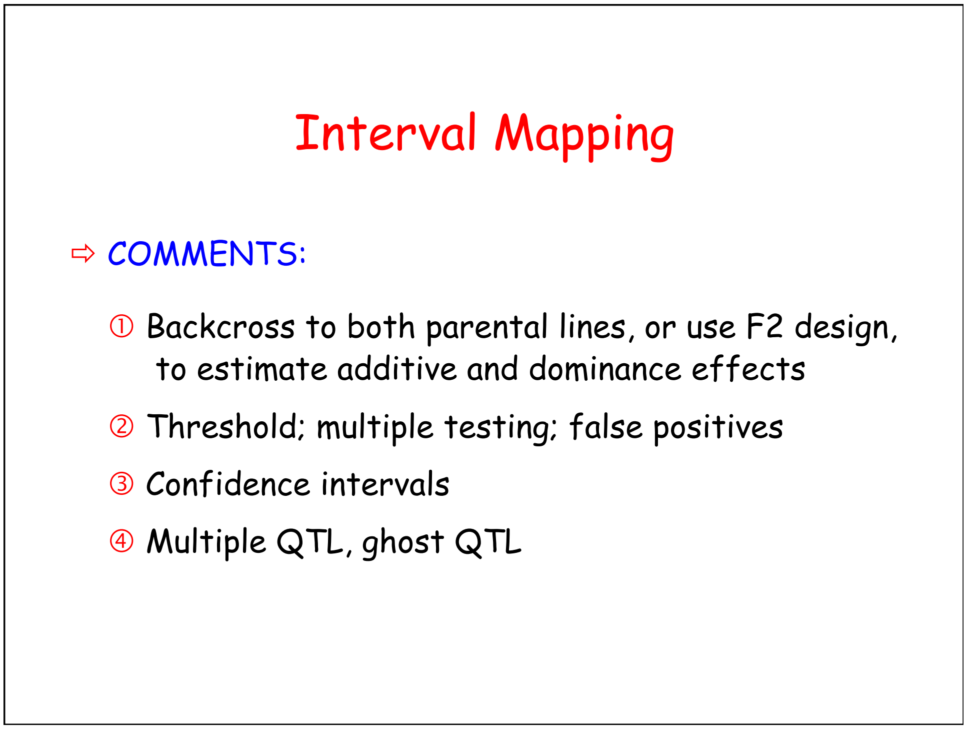
$$\mathrm{RSS} = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$$

Estimated position of the QTL minimizes RSS.



QTL Mapping

Interval Mapping; Example with Backcross



Interval Mapping

⇨ COMMENTS:

① Backcross to both parental lines, or use F2 design, to estimate additive and dominance effects

② Threshold; multiple testing; false positives

③ Confidence intervals

④ Multiple QTL, ghost QTL

Interval Mapping Example

R/QTL package in R: Simulated backcross data (Broman and Saunak, 2009) with 400 individuals (200 males and 200 females; sex == 1 and 0, respectively) with a single quantitative phenotype.

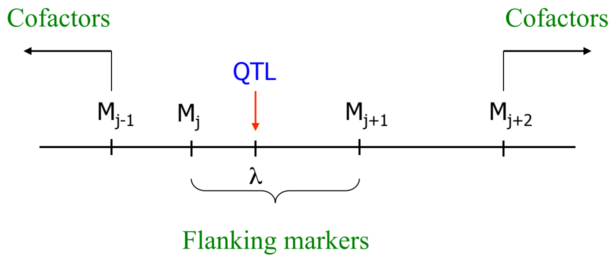
Interval mapping with sex as an additive covariate and sex as an interactive covariate, and also with males and females separately. Detection of regions of the genome affecting the phenotype, and also QTL × sex interactions?

QTL scan

Composite Interval Mapping

(Zeng, 1993, 1994)

⇨ Interval analysis adding marker cofactors (to account for the effects of unlinked QTLs); combination of single interval mapping and multiple linear regression



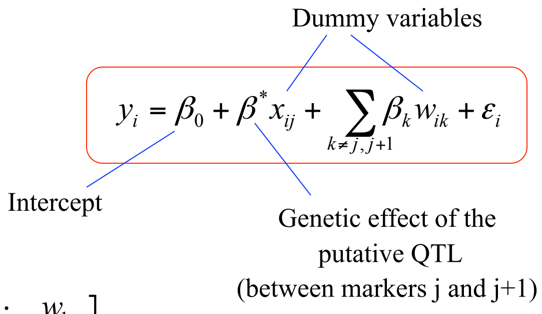
Composite Interval Mapping

(Zeng, 1993, 1994)

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\Downarrow$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$



Dummy variables

$$y_i = \beta_0 + \beta^* x_{ij} + \sum_{k \neq j, j+1} \beta_k w_{ik} + \varepsilon_i$$

Intercept

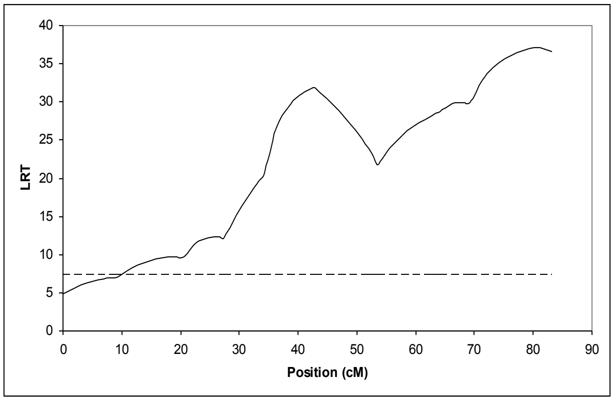
Genetic effect of the putative QTL (between markers j and j+1)

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1j} & w_{11} & \cdots & w_{1p} \\ 1 & x_{2j} & w_{21} & \cdots & w_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{Nj} & w_{N1} & \cdots & w_{Np} \end{bmatrix}$$

Interval Mapping

(Example)

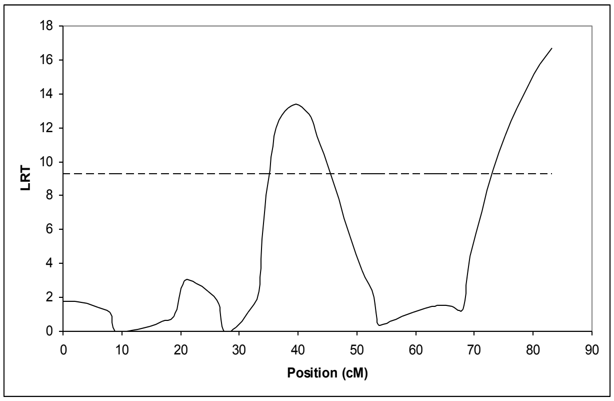
⇨ *Brassica napus*; Flowering time (Satagopan et al., 1996)



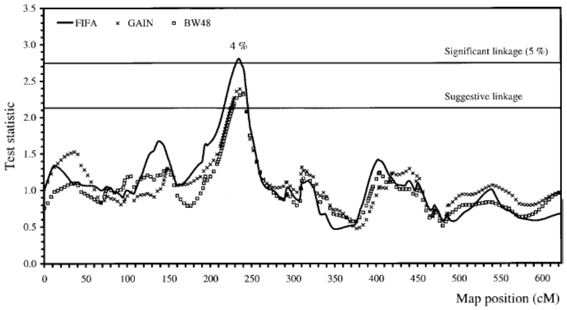
Composite Interval Mapping

(Example)

- ⇨ *Brassica napus;* Flowering time (Satagopan et al., 1996)



QTL Database (Livestock)



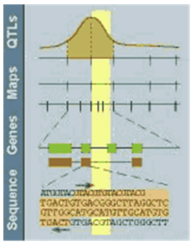
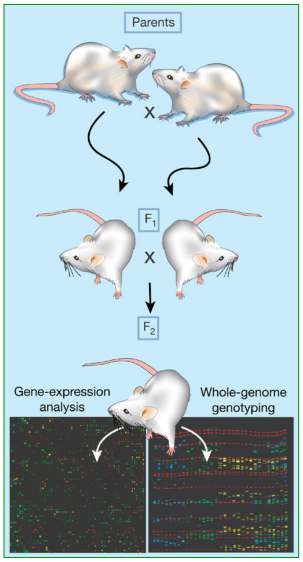
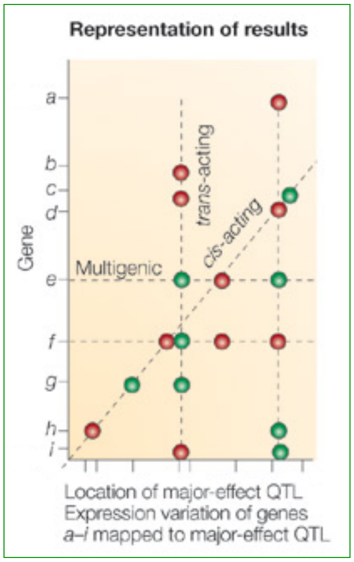


FIGURE 1. Test statistic values from the analysis of body weight at 48 d (BW48), growth between 23 and 48 d (GAIN), and feed intake between 23 and 48 d (FIFA) for quantitative trait loci on Chromosome 1. Significant and suggestive linkage thresholds of FIFA are included. The thresholds for BW48 and GAIN were slightly higher. Map positions are given using the Haldane scale.

EXPRESSION QTL (eQTL)



Darvasi (2003)



Jansen and Nap (2001)

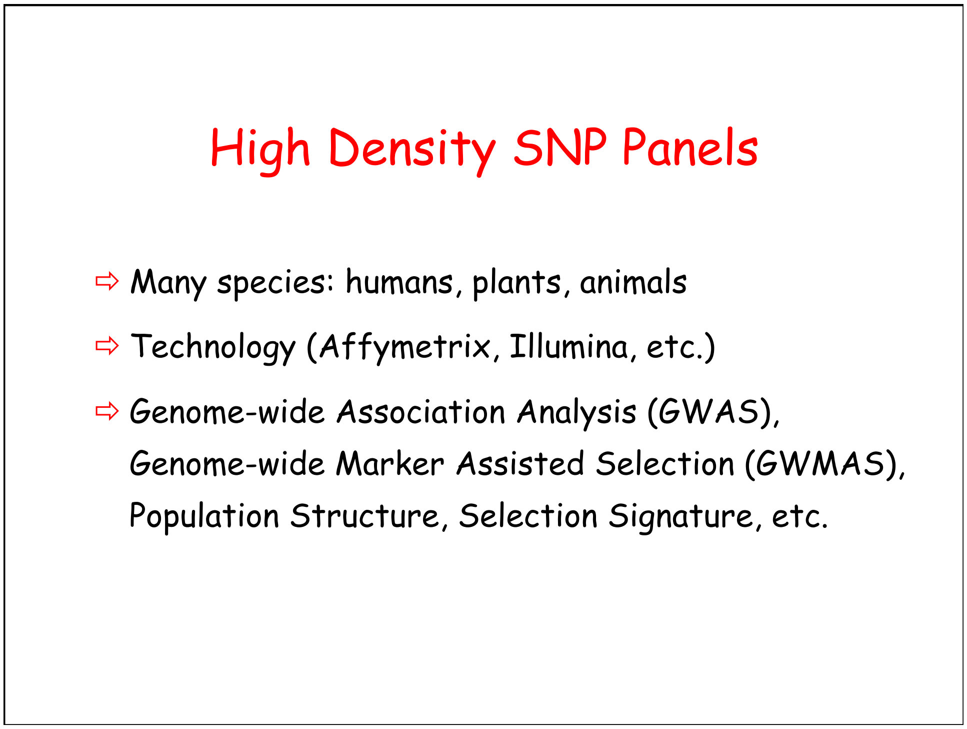
Genome-Wide Association Analysis (GWAS)

Guilherme J. M. Rosa

University of Wisconsin-Madison

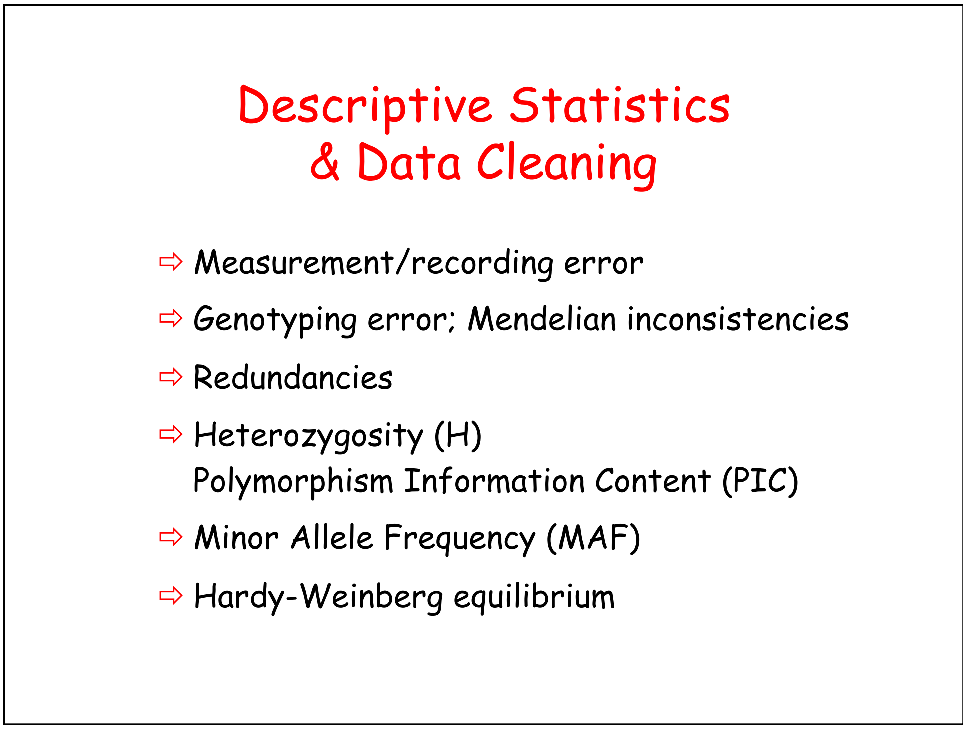
Gene Mapping

- ⇨ Linkage Analysis (QTL Analysis)
- ⇨ Fine Mapping Strategies (LDLA approach, Selective Genotyping, etc.)
- ⇨ Association Analysis, Candidate Gene Approach
- ⇨ Genome-wide Association Analysis (GWAS)



High Density SNP Panels

- ⇨ Many species: humans, plants, animals
- ⇨ Technology (Affymetrix, Illumina, etc.)
- ⇨ Genome-wide Association Analysis (GWAS),
  Genome-wide Marker Assisted Selection (GWMAS),
  Population Structure, Selection Signature, etc.



Descriptive Statistics & Data Cleaning

- ⇨ Measurement/recording error
- ⇨ Genotyping error; Mendelian inconsistencies
- ⇨ Redundancies
- ⇨ Heterozygosity (H)
  Polymorphism Information Content (PIC)
- ⇨ Minor Allele Frequency (MAF)
- ⇨ Hardy-Weinberg equilibrium

Single Marker Regression

⇨ Series of models, one for each marker j (j = 1, 2,…, k):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{m}\mathbf{g}_j + \mathbf{e}$$

where:

y: vector of phenotypic observations (n individuals)

β: environmental covariates, such as gender, age, etc.

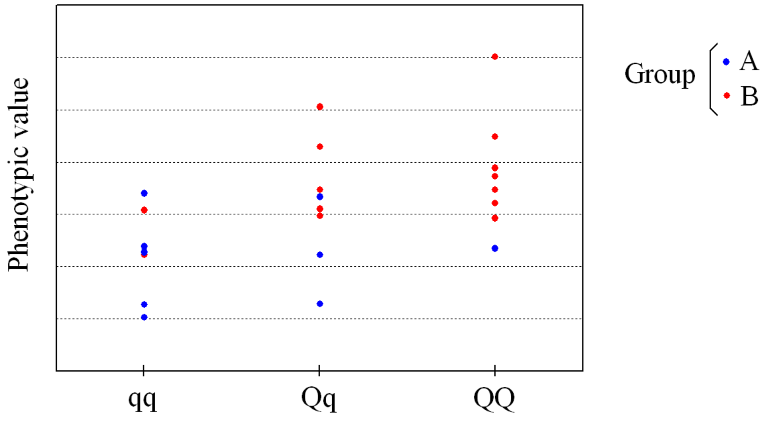
X: incidence matrix relating β to **y**

g_j: 'effect' of marker j (j = 1, 2,…, k)

m = $[m_{1j}, m_{2j},…, m_{nj},]^T$: vector of genotypes for marker j, with m_{ij} = -1, 0 or 1

e: residual vector

Confounding



⇨ True model: $y_{ij} = \mu + \text{Group}_i + e_{ij}$

Accounting for Population Stratification

⇨ Series of models, one for each marker j (j = 1, 2,…, k):

$$\mathbf{y} = \mathbf{X}\beta + \psi + \mathbf{m}\mathbf{g}_j + \mathbf{e}$$

where: **Ψ** is a population structure term (e.g. PC built from genotypes)

Mixed Model Approach

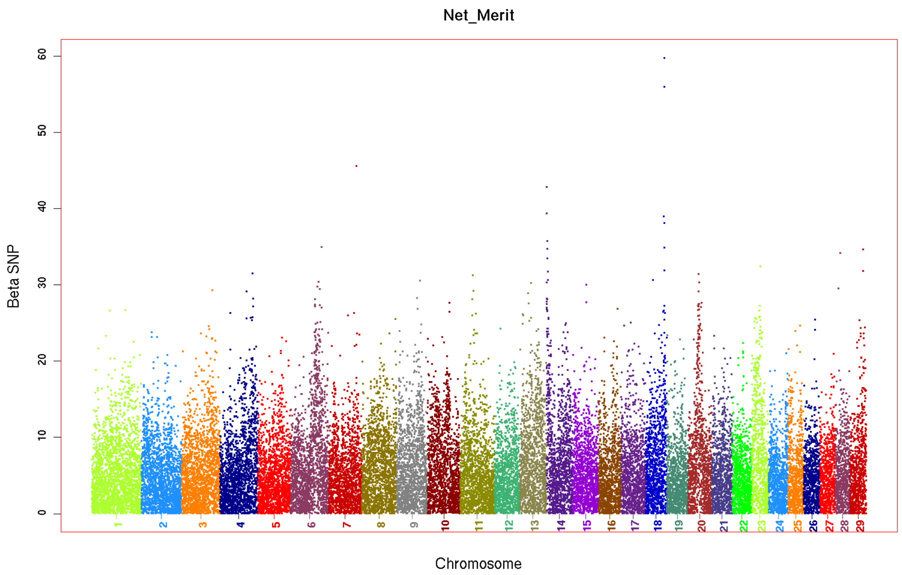
⇨ The model now is expressed as:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} + \mathbf{m}\mathbf{g}_j + \mathbf{e}$$

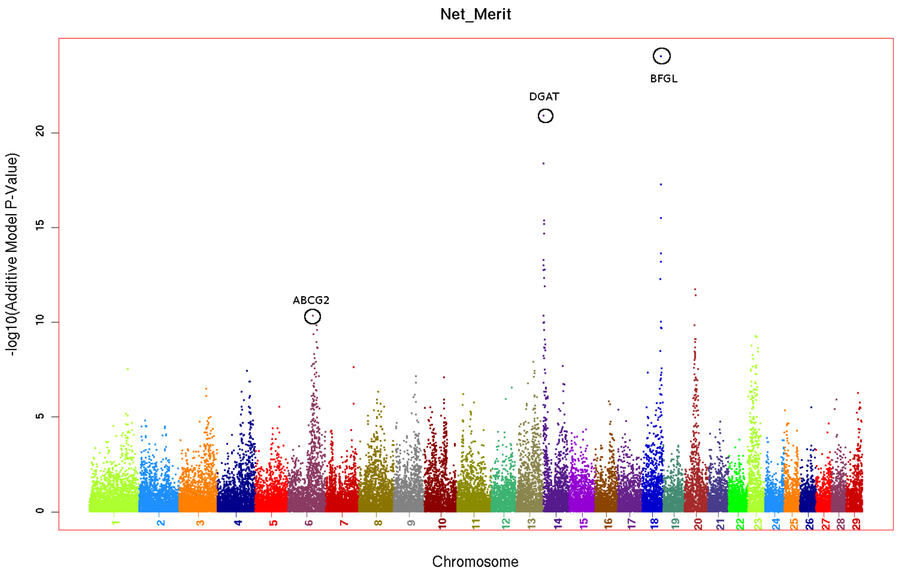
where all terms are as before, except that a polygenic (infinitesimal) term **u** is included to account for population sub-structure, with **u** ~ N(0, $\mathbf{K}\sigma_u^2$); **K** is a kinship matrix built from pedigree information (e.g. ***A***) or genotypic information (e.g. ***G***)

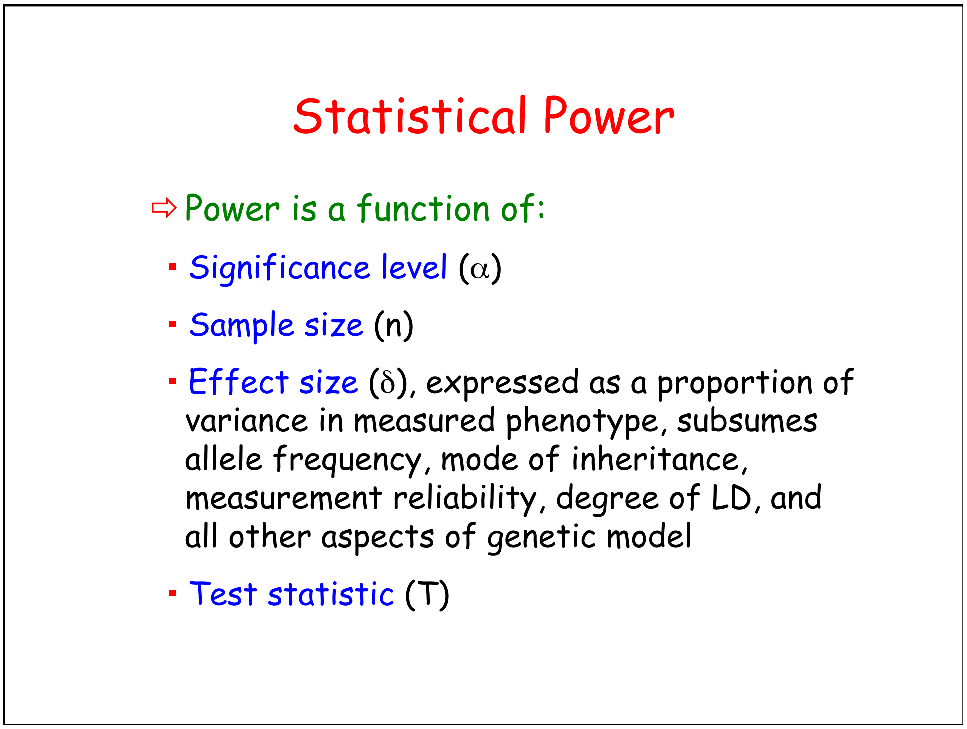
Note: Efficient computation, e.g. EMMA and GEMMA

Manhattan Plot with Marker Effects



Manhattan Plot with Significance Tests

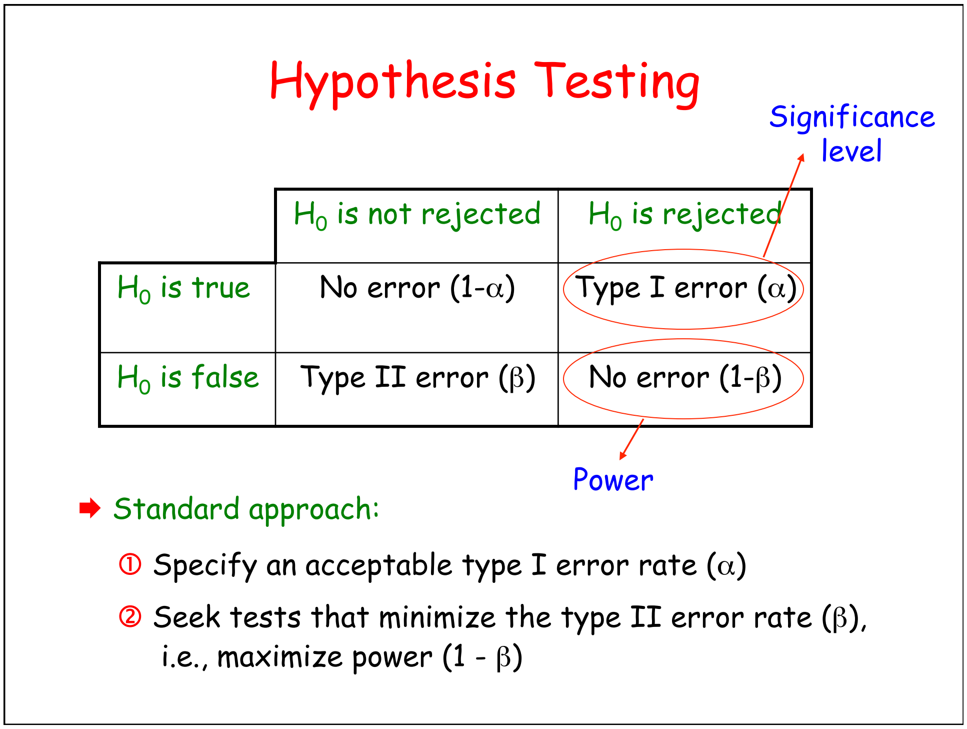




Statistical Power

⇨ Power is a function of:

- Significance level (α)
- Sample size (n)
- Effect size (δ), expressed as a proportion of variance in measured phenotype, subsumes allele frequency, mode of inheritance, measurement reliability, degree of LD, and all other aspects of genetic model
- Test statistic (T)



Hypothesis Testing

Significance level

	H_0 is not rejected	H_0 is rejected
H_0 is true	No error ($1-\alpha$)	Type I error (α)
H_0 is false	Type II error (β)	No error ($1-\beta$)

Power

➡ Standard approach:

① Specify an acceptable type I error rate (α)

② Seek tests that minimize the type II error rate (β), i.e., maximize power ($1 - \beta$)

The Multiple Testing Issue

Suppose you carry out 10 hypothesis tests at the 5% level (assume independent tests)

The probability of declaring a particular test significant under its null hypothesis is 0.05

But the probability of declaring at least 1 of the 10 tests significant is $\boxed{0.401} \rightarrow 1 - 0.95^{10}$

If you perform 20 hypothesis tests, this probability increases to 0.642…

- Typically thousands of markers tested simultaneously
- Example: Suppose trait with $H^2 = 0$ and association analysis considering 100 markers and $\alpha = 5\%$ (for each test)
  - Expected 100 x 0.05 = 5 false associations…

The Multiple Testing Issue

	# H_0 not rejected	# H_0 rejected	
# true H_0	A	B	m_0
# false H_0	C	D	m_1
	m - R	R	m

Observable quantity (nº rejected H_0)

known quantity (number of tests)

The Multiple Testing Issue

- Family-wise error rate (FWER):

$$\mathrm{FWER} = \Pr(B \geq 1) = 1 - \Pr(B = 0)$$

- False discovery rate (FDR):

$$\mathrm{FDR} = \underbrace{E[B/R \,|\, R > 0]}_{\text{Positive FDR (pFDR); Storey (2002)}} \Pr(R > 0)$$

Controlling the FWER at level α:

Pr[V ≥ 1]

- Bonferroni: Rejects any hypothesis H_j with p-value less than or equal to α/m, i.e.:

$$\boxed{\tilde{p}_j = \min[mp_j, 1]}$$

adjusted p-value

unadjusted p-value

- Sidák: Rejects any hypothesis H_j with p-value less than or equal to $1-(1-\alpha)^{1/g}$, i.e.:

$$\boxed{\tilde{p}_j = \min[1-(1-p_j)^g, 1]}$$

  - Very similar to Bonferroni adjustment.
  - Both are too conservative...

Controlling the FDR:

Definition: FDR = $E[V/R \mid R>0]\Pr[R>0]$; expected proportion of false positive findings among all rejected hypotheses times the probability of making at least one rejection.

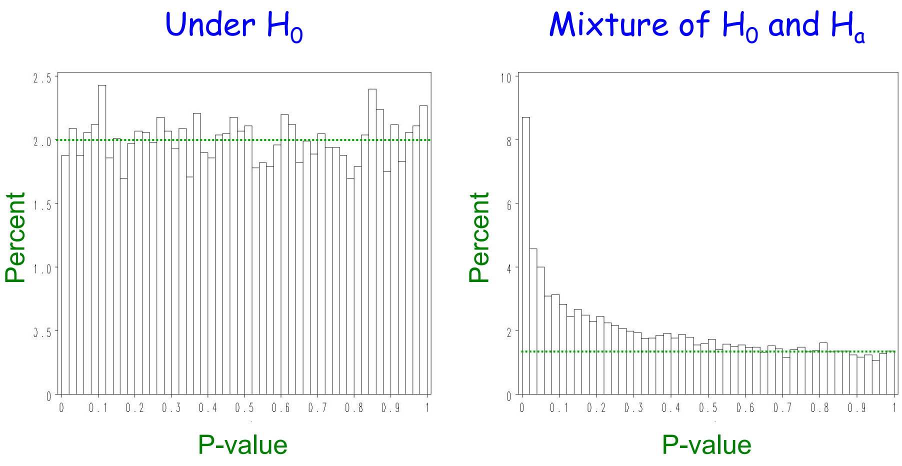
Positive FDR (pFDR); Storey (2002)

- Benjamini and Hochberg (1995) algorithm:

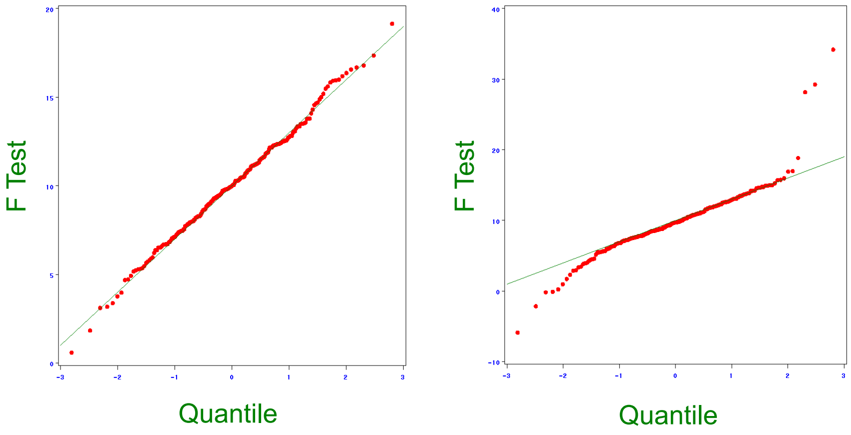
- Fix a value $\alpha^* \in (0,1)$
- Let $p_{(1)}, p_{(2)},\ldots,p_{(m)}$ be the ordered observed p-values
- Let $\hat{k} = \max\{k\colon p_{(k)} \le \alpha^*(k/m)\}$

  (If $p_{(k)} > \alpha^*(k/m)$ for all $k = 1,\ldots,m$, let $\hat{k} = 0$)
- If $\hat{k} \ge 1$, reject the hypotheses corresponding to $p_{(1)}, p_{(2)},\ldots,p_{(\hat{k})}$
- If $\hat{k} = 1$, do not reject any hypothesis

Distribution of P-values (Histogram)



Distribution of P-values (Q-Q Plot)


Under H_0

Mixture of H_0 and H_a

Replication

- ➪ Confounding factors, population structure and stratification, Type I error, etc.
- ➪ Biased estimates of gene effects due to significance threshold
- ➪ Multiple genes, with modest individual effects
- ➪ Gene × gene and gene × environment interactions
- ➪ Inter population heterogeneity
- ➪ Low statistical power
- ➪ Validation of association findings
- ➪ But what constitutes a replication?