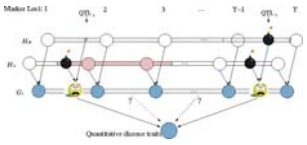
Computational Genomics

10-810/02-710, Spring 2009

Quantitative Trait Locus (QTL) Mapping

Eric Xing

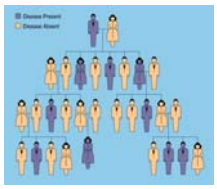
Lecture 23, April 13, 2009



Reading: DTW book, Chap 13

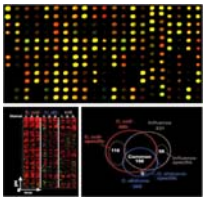
Phenotypical Traits

- Body measures:

- 

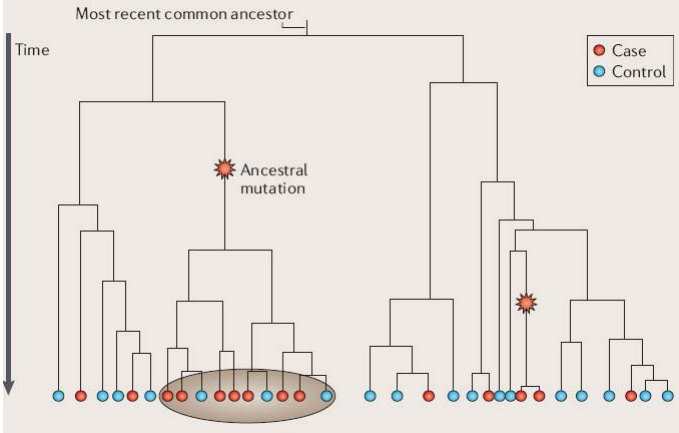
Disease susceptibility and drug response

- Gene expression (microarray)



Population-Based Association Study

- Case/control data are collected from unrelated individuals
  - All individuals are related if we go back far enough in the ancestry



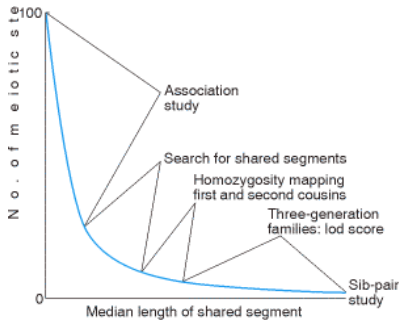
Linkage Analysis vs. Association Analysis

- Linkage analysis: Use the linkage disequilibrium between marker locus and disease locus to localize disease locus
- Association : co-occurrence of alleles and phenotypes in population. Association is observed when
  - A) There is a direct causation from the allele to disease
  - B) The marker and the disease locus are in linkage disequilibrium
  - C) There are confounding factors such as population stratification or admixture
    - It is important to try to exclude C) from A) and B) in association analysis!
- The marker locus found in linkage analysis of family data may not show association to the disease in the population of unrelated individuals.

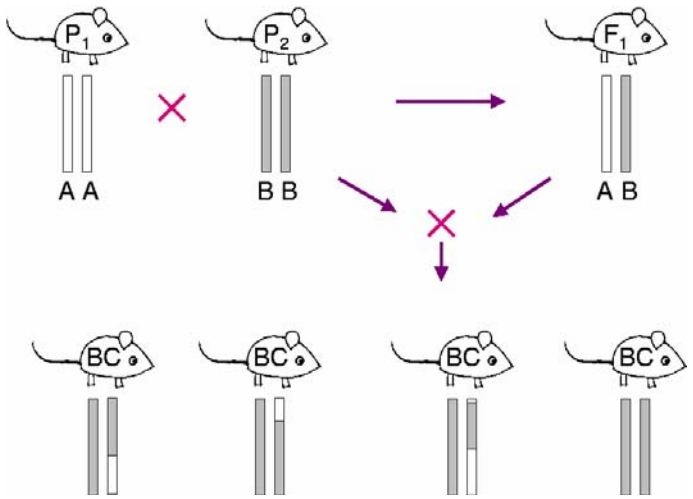
Linkage Analysis vs. Association Analysis

- Linkage Analysis
  - Based on pedigree data
  - Only a very small number of meiosis (and recombinations) separates two individuals in a given family.
    - When the disease locus is mapped to be linked to a marker locus, the mapped segment of chromosome is usually too large. A follow-up study is necessary to further narrow down the candidate region
    - A relatively small number of markers need to be genotyped
- Association Analysis
  - Based on controls/cases for a given disease, unrelated individuals in population
  - A large number of meiosis separates two individuals.
    - The chromosome segment in linkage disequilibrium is small.
    - A relatively large number of markers are required.
    - SNPs are commonly used genetic markers because of the availability of high-throughput genotyping technology

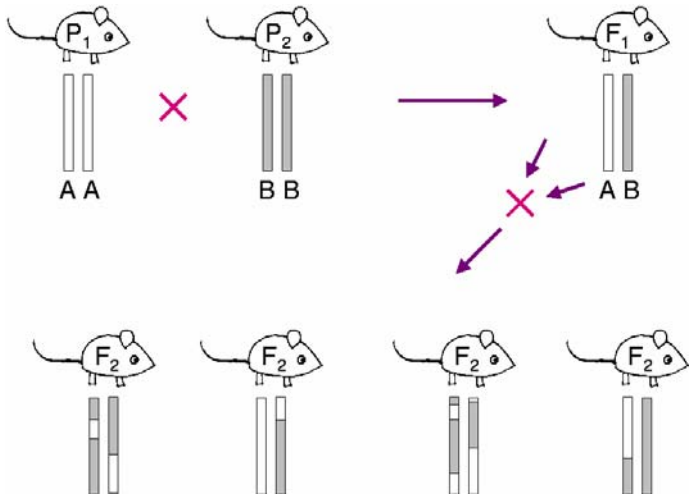
Linkage Analysis vs. Association Analysis



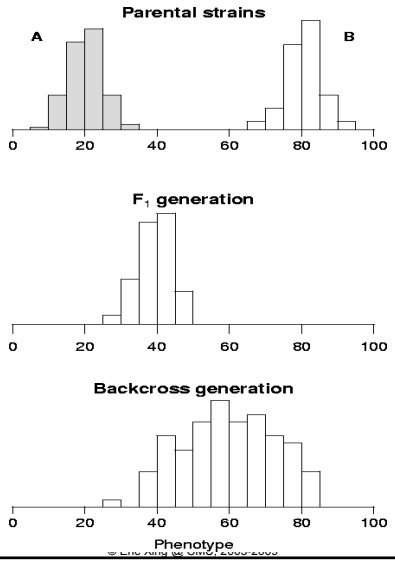
Backcross experiment



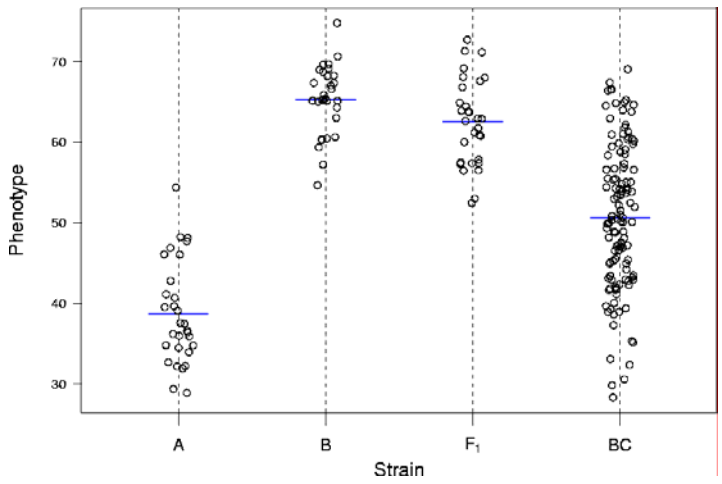
F_2 intercross experiment



Trait distributions: a classical view

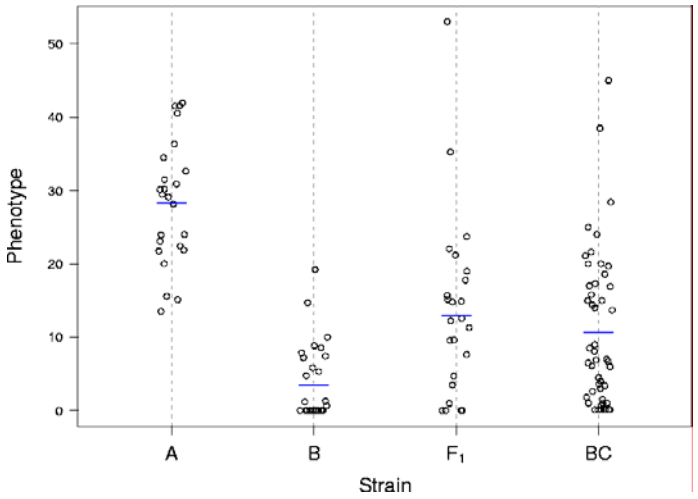


Another representation of a trait distribution



Note the equivalent of dominance in our trait distributions.

A second example

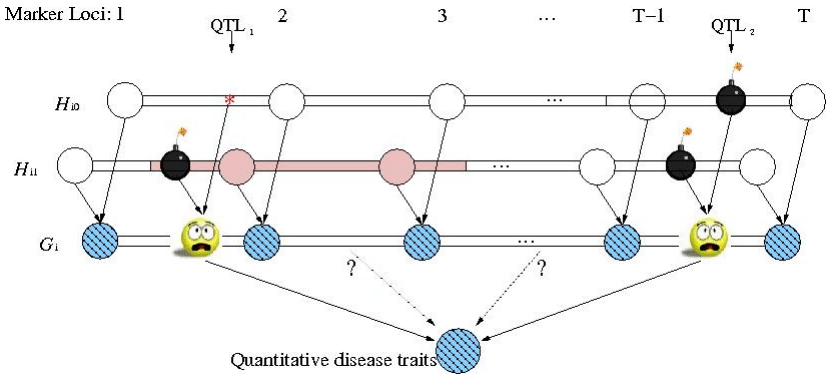


Note the approximate additivity in our trait distributions here.

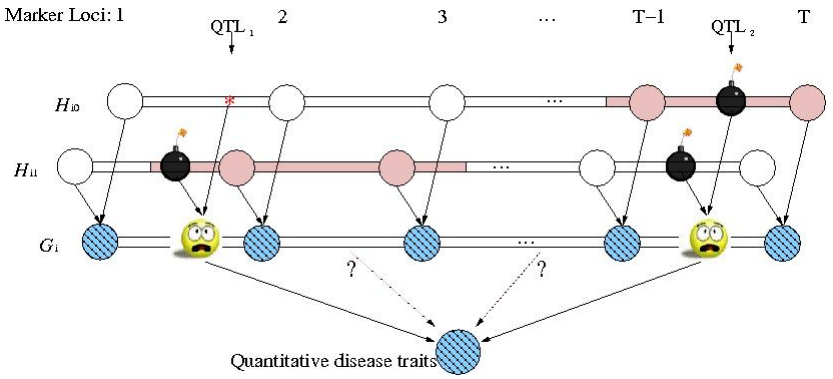
QTL mapping

- Data
  - Phenotypes: y_i = trait value for mouse *i*
  - Genotype: x_{ij} = 1/0 (i.e., A/H) of mouse *i* at marker *j*(backcross); need three states for intercross
  - Genetic map: Locations of markers
- Goals
  - Identify the (or at least one) genomic region, called quantitative trait locus = QTL, that contributes to variation in the trait
  - Form confidence intervals for the QTL location
  - Estimate QTL effects

QTL mapping (BC)



QTL mapping (F2)



Discrete Traits:
Case/Control Association Analysis

- Case/control data are collected from unrelated individuals

	Case(1)/Control(0)	Genotype
Individual 1	0	C C / T A / C C / T T
Individual 2	1	G C / A T / G C / A T
⋮		
Individual N	1	G G / T T / C G / T T

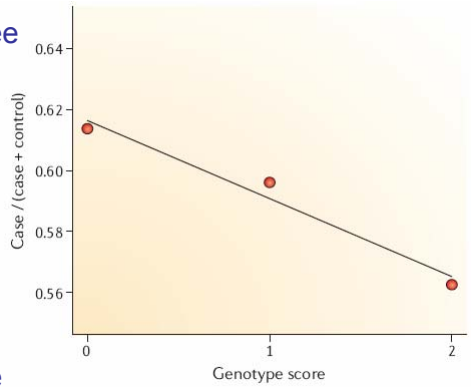
Markers with no association

Marker associated with the disease

Discrete Case/Control Association Analysis

- For each marker locus, find the 3x2 contingency table containing the counts of three genotypes

Genotype	Case	Control
AA	$N_{case,AA}$	$N_{cont,AA}$
Aa	$N_{case,Aa}$	$N_{cont,Aa}$
aa	$N_{case,aa}$	$N_{cont,aa}$
Total	N_{case}	N_{cont}



- Pearson test with 2 df, or Fisher's exact test under the null hypothesis of no association

Discrete Case/Control Association Analysis

- Alternatively, assume an additive model, where the heterozygote risk is approximately between the two homozygotes
- Form a 2x2 contingency table. Each individual contributes twice from each of the two chromosomes.

Genotype	Case	Control
A	$G_{case,A}$	$G_{cont,A}$
a	$G_{case,a}$	$G_{cont,a}$
Total	$2xN_{case}$	$2xN_{cont}$

- Pearson test with 1df

Models: Recombination

- We assume no chromatid or crossover interference.

⇒ points of exchange (crossovers) along chromosomes are distributed as a Poisson process, rate 1 in genetic distance

⇒ the marker genotypes $\{x_{ij}\}$ form a Markov chain along the chromosome for a backcross; what do they form in an F_2 intercross?

Models: Genotype → Phenotype

- Let y = phenotype,
  g = whole genome genotype
- Imagine a small number of QTL with genotypes $g_1, \ldots, g_p$ (2^p or 3^p distinct genotypes for BC, IC resp, why?).

We assume

$$E(y|g) = \mu(g_1, \ldots g_p), \quad var(y|g) = \sigma^2(g_1, \ldots g_p)$$

Models: Genotype → Phenotype

- **Homoscedacity** (constant variance)

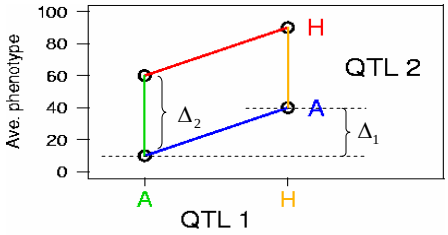
  $\sigma^2(g_1, \ldots g_p) = \sigma^2$ *(constant)*
- **Normality** of residual variation

  $y|g \sim N(\mu_g, \sigma^2)$
- **Additivity**:

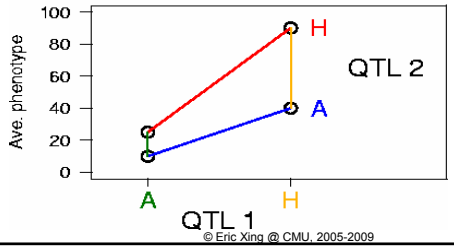
  $\mu(g_1, \ldots g_p) = \mu + \sum \Delta_j g_j$ (g_j = 0/1 for BC)
- **Epistasis**: Any deviations from additivity.

  $\mu(g_1, \ldots g_p) = \mu + \sum \Delta_j g_j + \sum \omega_{ij} g_i g_j$

Additivity, or non-additivity (BC)



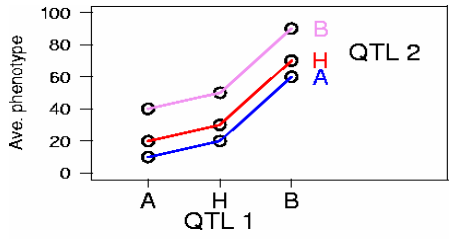
The effect of QTL 1 is the same, irrespective of the genotype of QTL 2, and vice versa.

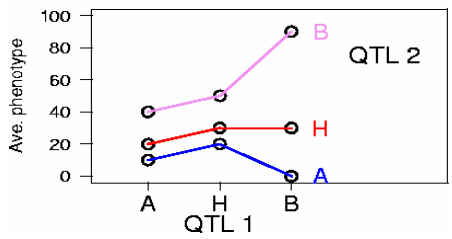


Epistatic QTLs

$\Delta_i \sim p(\ \ | g_j)$

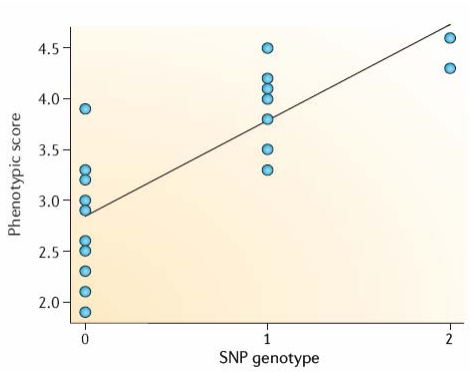
Additivity or non-additivity: F2





Association Analysis with Continuous-valued Traits

- Continuous-valued traits
  - Cholesterol level, blood pressure etc.
- For each locus, fit a linear regression using the number of minor alleles at the given locus of the individual as covariate
- Alternatively, for each locus perform ANOVA

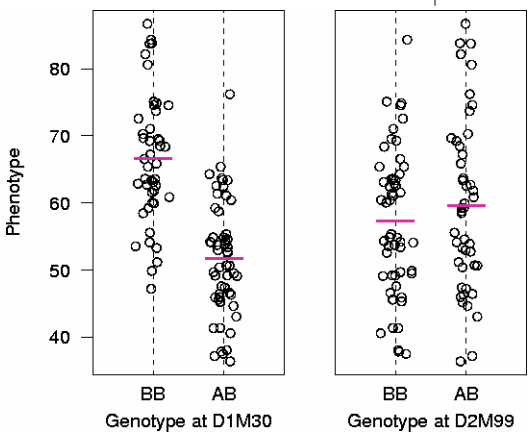


The simplest method: ANOVA

One marker at a time:

- Split subjects into groups according to genotype at a single marker
- Do a t-test/ANOVA
- Repeat for each marker

t-test/ANOVA will tell whether there is sufficient evidence to say that measurements from one condition (i.e., genotype) differ significantly from another



- LOD score = $\log_{10}$ likelihood ratio, comparing single-QTL model to the "no QTL anywhere" model.

ANOVA at marker loci

Advantages

- Simple
- Easily incorporate covariates (sex, env, treatment ...)
- Easily extended to more complex models

Disadvantages

- Must exclude individuals with missing genotype data
- Imperfect information about QTL location
- Suffers in low density scans
- Only considers one QTL at a time

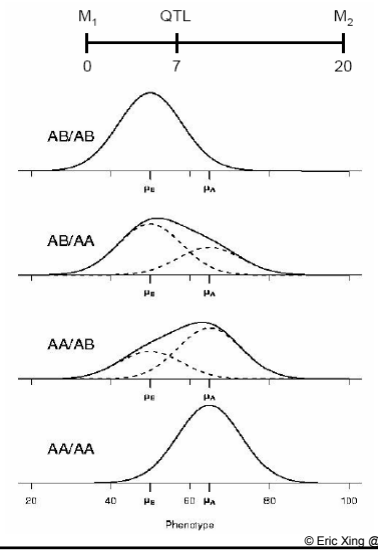
Interval mapping (IM)

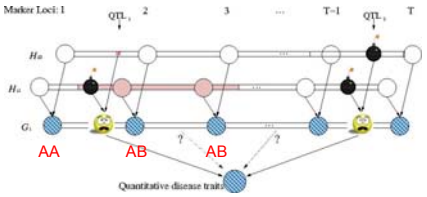
- Consider any one position in the genome as the location for a putative QTL
- For a particular mouse, let z = 1/0 if (unobserved) genotype at QTL is AB/AA
- Calculate Pr(z = 1 | marker data of an interval bracketing the QTL)
  - Assume no meiotic interference
  - Need only consider flanking typed markers
  - May allow for the presence of genotyping errors
- Given genotype at the QTL, phenotype is distributed as

$$y_i \mid z_i \sim Normal(\mu_{zi}, \sigma^2)$$

- Given marker data, phenotype follows a *mixture* of normal distributions

IM: the mixture model





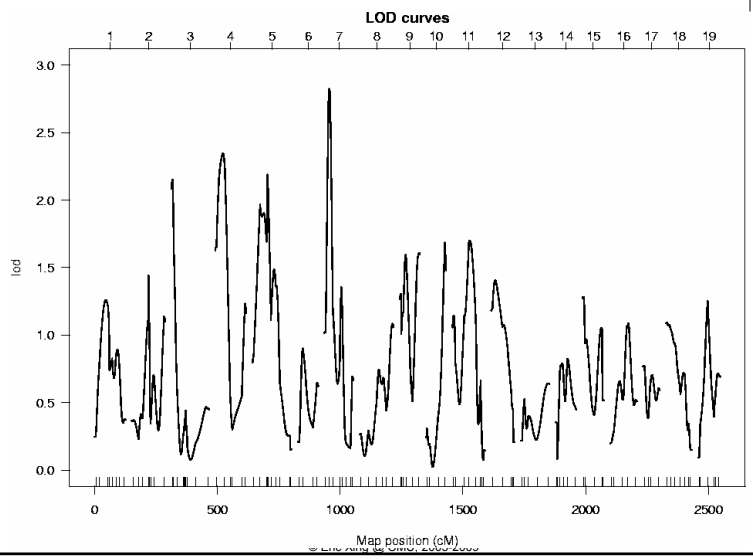
IM: estimation and LOD scores

- Use a version of the EM algorithm to obtain estimates of μ_{AA}, μ_{AB}, and σ (an *iterative* algorithm)
- Calculate the LOD score

$$\text{LOD} = \log_{10}\left\{\frac{P(\text{data}|\hat{\mu}_{AA},\hat{\mu}_{AB})}{P(\text{data}|\text{no QTL})}\right\}$$

- Repeat for all other genomic positions (in practice, at 0.5 cM steps along genome)

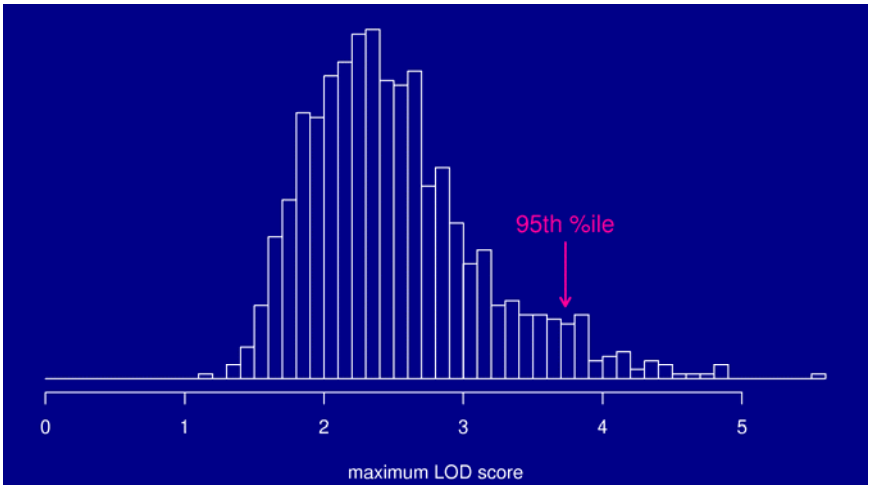
LOD score curves



LOD thresholds

- To account for the genome-wide search, compare the observed LOD scores to the distribution of the maximum LOD score, genome-wide, that would be obtained if there were no QTL anywhere.
- **LOD threshold** = 95th %ile of the distribution of genome-wide maxLOD, when there are no QTL anywhere
- **Derivations:**
  - Analytical calculations (Lander & Botstein, 1989)
  - Simulations
  - Permutation tests (Churchill & Doerge, 1994).

Permutation distribution for trait4



Interval mapping

Advantages

- Make proper account of missing data
- Can allow for the presence of genotyping errors
- Pretty pictures
- Higher power in low-density scans
- Improved estimate of QTL location

Disadvantages

- Greater computational effort
- Requires specialized software
- More difficult to include covariates?
- Only considers one QTL at a time

Multiple QTL methods

Why consider multiple QTL at once?

- To separate linked QTL. If two QTL are close together on the same chromosome, our one-at-a-time strategy may have problems finding either (e.g. if they work in opposite directions, or interact). Our LOD scores won't make sense either.
- To permit the investigation of interactions. It may be that interactions greatly strengthen our ability to find QTL, though this is not clear.
- To reduce residual variation. If QTL exist at loci other than the one we are currently considering, they should be in our model. For if they are not, they will be in the error, and hence reduce our ability to detect the current one. See below.

The problem

- *n* backcross subjects; *M* markers in all, with at most a handful expected to be near QTL

x_{ij} = genotype (0/1) of mouse *i* at marker *j*
y_i = phenotype (trait value) of mouse *i*

$$Y_i = \mu + \sum_{j=1}^{M} \Delta_j x_{ij} + \varepsilon_j \qquad \text{Which } \Delta_j \neq 0 \text{ ?}$$

$\Rightarrow$ Variable selection in linear models (regression)

Finding QTL as model selection

Select class of models

- Additive models
- Additive plus pairwise interactions
- Regression trees

Compare models (γ)

- $BIC_\delta(\gamma) = logRSS(\gamma) + \gamma(\delta \log n/n)$
- Sequential permutation tests

Search model space

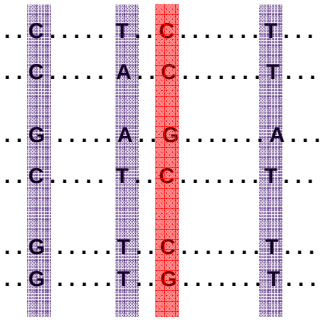
- Forward selection (FS)
- Backward elimination (BE)
- FS followed by BE
- MCMC

Assess performance

- Maximize no QTL found;
- control false positive rate

Logistic Regression for Multiple SNPs in Case/control Association

	Case(1)/Control(0)	Genotype
Individual 1	0	C . . . . . T . . C . . . . T . . . C . . . . . A . . C . . . . T . . .
Individual 2	1	G . . . . . A . . G . . . . A . . . C . . . . . T . . C . . . . T . . .
⋮		
Individual N	1	G . . . . . T . . C . . . . T . . . G . . . . . T . . G . . . . T . . .



$$p(y_i=\text{case}) = f\left(\sum_{k=1}^{K} x_{ik}\beta_k\right)$$

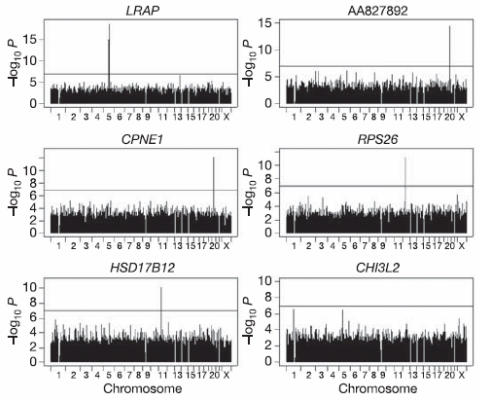
- f : logistic function
- β_k : weight for the kth SNP
- x_{ik} : genotype of the kth SNP for the ith individual (0,1, or 2 depending on the number of minor alleles)

Logistic Regression for Multiple SNPs in Case/control Association

- Variable selection methods
  - Stepwise selection procedure
  - Shrinkage methods such as Lasso
- Similarly, for continuous-valued traits, multivariate regression with variable selection methods

Expression QTL

- Microarray gene expression level data as phenotype
  - cis eQTL : eQTL for the given gene is located near the gene
  - trans eQTL : eQTL for the given gene is located far from the gene or on a different chromosome



Multi-marker Approach

- Form a new allele by combining multiple SNPs

SNP A	SNP B
0	0
0	1
1	0
1	1

Auxiliary Markers for Haplotypes			
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

- Pros : multi-marker approach can capture dependencies across multiple markers
  - SNPs in LD form a haplotype that can be tested as a single allele
- Cons: Haplotype of *K* SNPs result in 2^K haplotypes
  - The number of samples corresponding to each haplotype decreases quickly as we increase *K*

Acknowledgements

Melanie Bahlo, WEHI
Hongyu Zhao, Yale
Karl Broman, Johns Hopkins
Nusrat Rabbee, UCB

References

www.netspace.org/MendelWeb

HLK Whitehouse: **Towards an Understanding of the Mechanism of Heredity,** 3rd ed. Arnold 1973

Kenneth Lange: **Mathematical and statistical methods for genetic analysis,** Springer 1997

Elizabeth A Thompson: **Statistical inference from genetic data on pedigrees,** CBMS, IMS, 2000**.**

Jurg Ott : **Analysis of human genetic linkage**, 3rd edn Johns Hopkins University Press 1999

JD Terwilliger & J Ott : **Handbook of human genetic linkage**, Johns Hopkins University Press 1994