Computational Genomics

Population Genetics:
Quantitative Trait Locus (QTL) Mapping

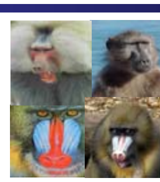
Eric Xing

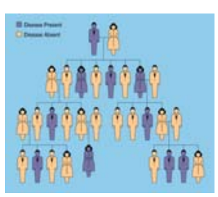
Lecture 4, January 25, 2007

Reading: DTW book, Chap 13

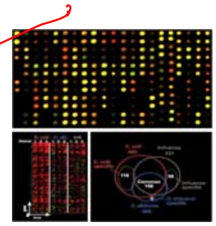
Phenotypical Traits

- Body measures:



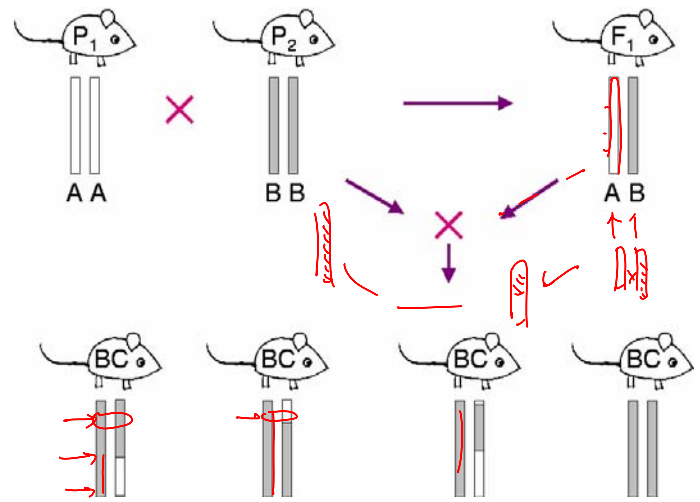
- 

Disease susceptibility and drug response

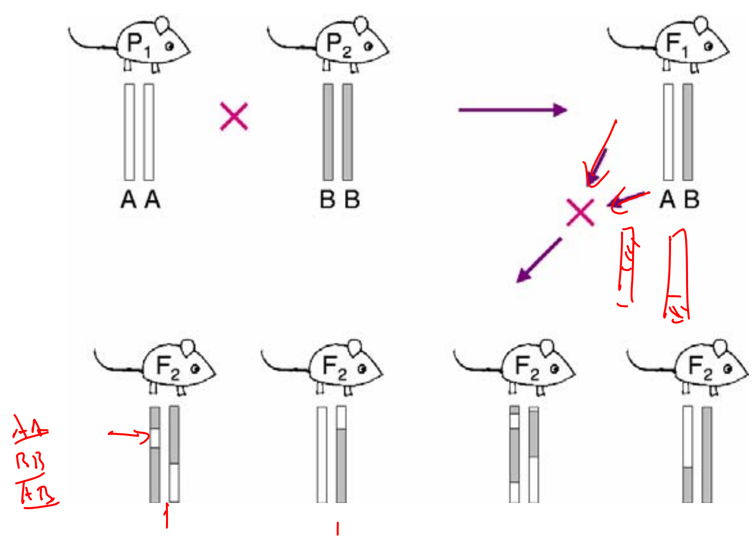


- Gene expression (microarray)

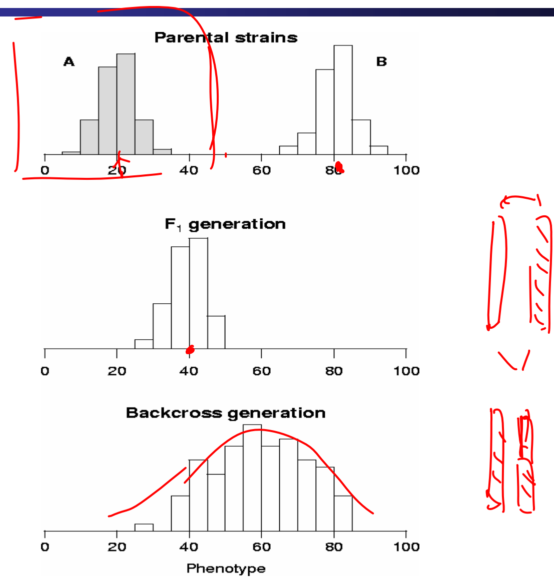
Backcross experiment



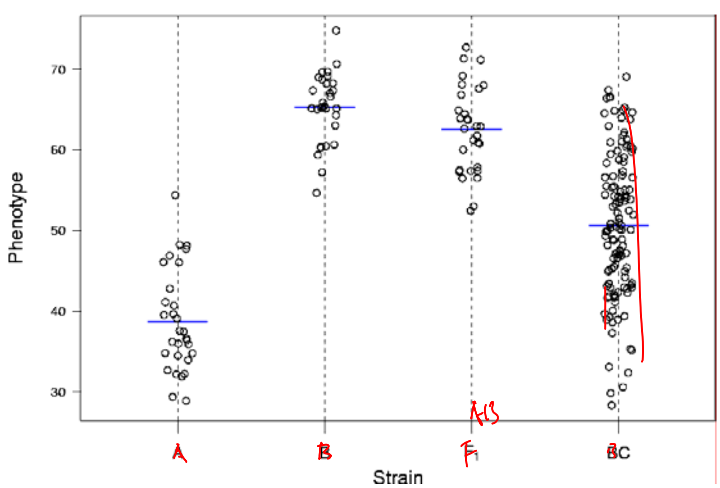
F_2 intercross experiment



Trait distributions: a classical view

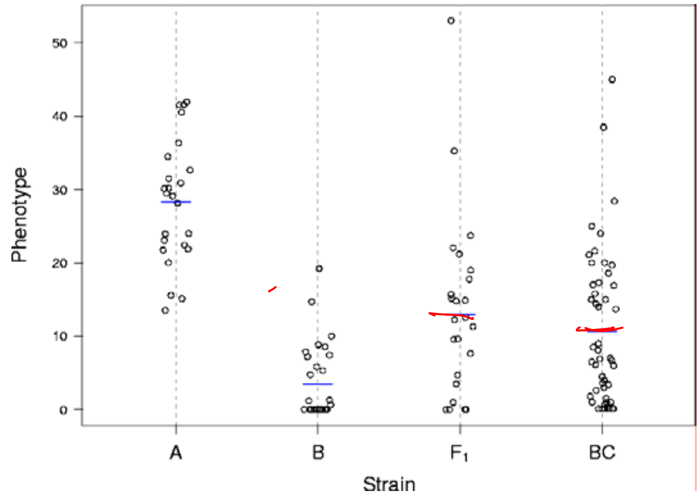


Another representation of a trait distribution



Note the equivalent of dominance in our trait distributions.

A second example



Note the approximate additivity in our trait distributions here.

QTL mapping

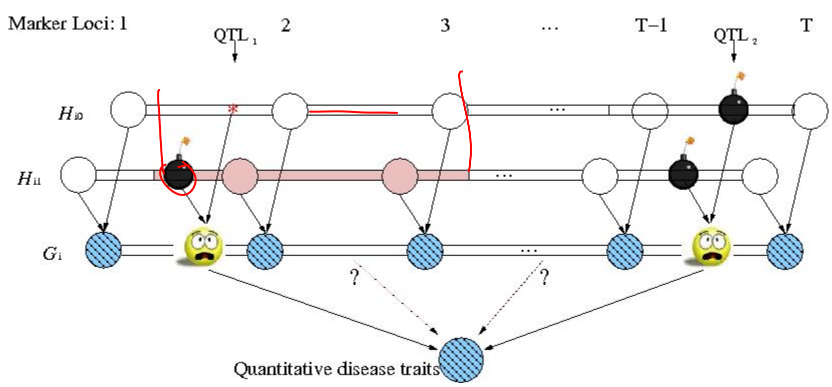
AA AB

0/1 BC

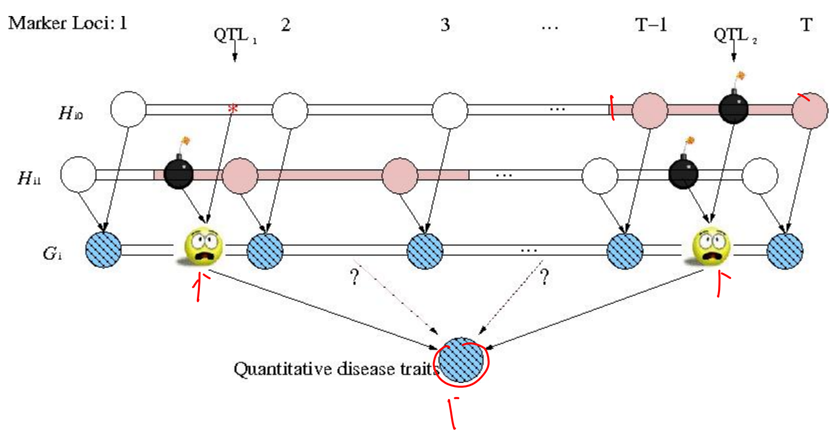
1 0 1 0

- 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)



Models: Recombination

- We assume no chromatid or crossover interference.

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

$\Rightarrow$ 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 $\rightarrow$ Phenotype

- Let y = phenotype, $\sim N(\mu, \sigma^2)$
  g = whole genome genotype

$\mu(\text{—})$
$\sigma^2(\text{—})$

- 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

  $y|g \sim N(\mu_g, \sigma^2)$

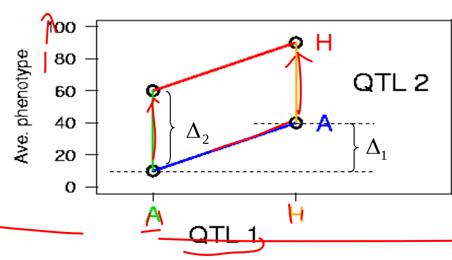
- **Additivity**:

  $\mu(g_1,\ldots g_p) = \mu + \sum \Delta_j g_j$ (g_j = 0/1 for BC)

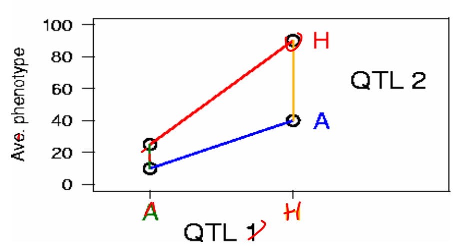
- **Epistasis**: Any deviations from additivity.

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

Additivity, or non-additivity (BC)



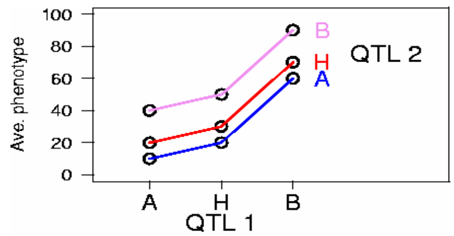
The effect of QTL 1 is the same, irrespective of the genotype of QTL 2, and vice versa.

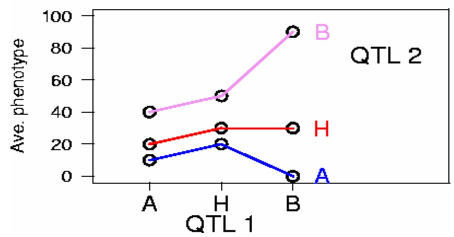


Epistatic QTLs

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

Additivity or non-additivity: F2

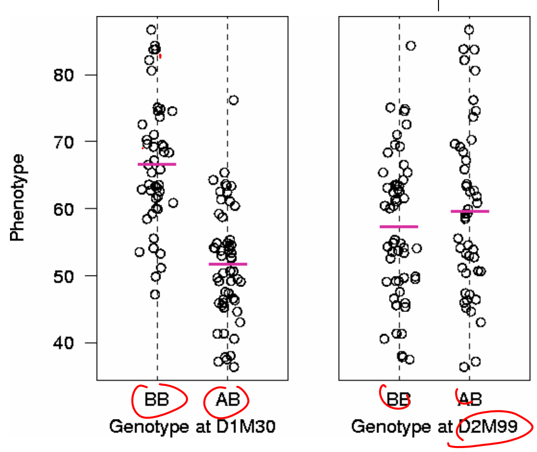




The simplest method: ANOVA

- 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)

QTL_i

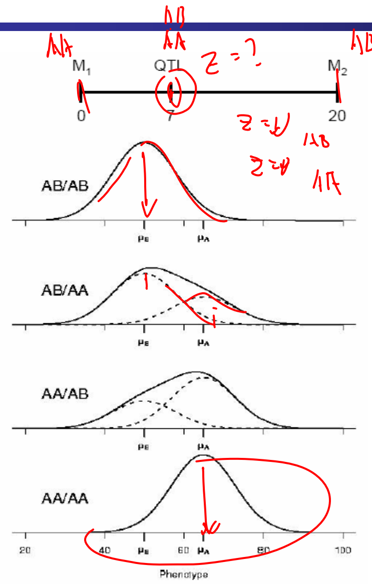
$z_i = \{1\ 0\}$

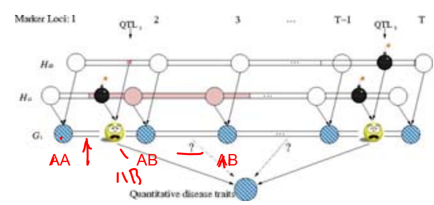
- Consider any one position in the genome as the location for a putative QTL
- For a particular mouse, let z = 1/0 if (unobserved) genotype at QTL is AB/AA
- Calculate Pr(z = 1 | marker data of an interval bracketing the QTL) $z_i = 1, 0$ (X_i)
  - Assume no meiotic interference
  - Need only consider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

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

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

IM: the mixture model





IM: estimation and LOD scores

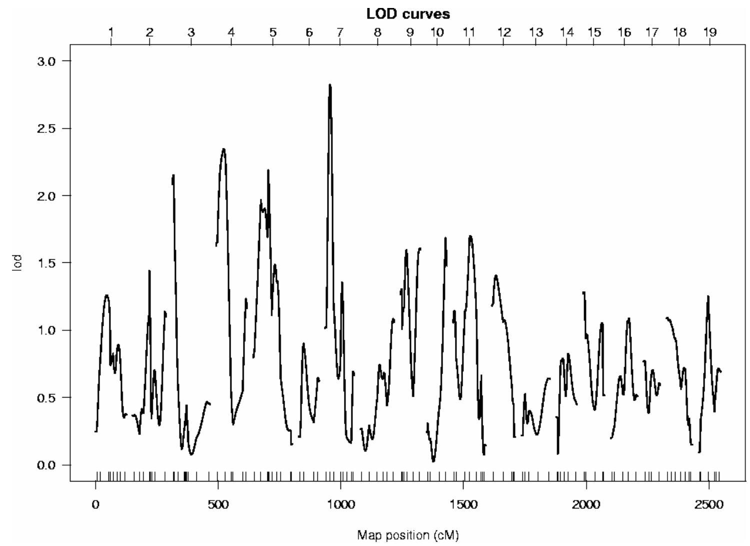
A — QTL — A

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

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

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

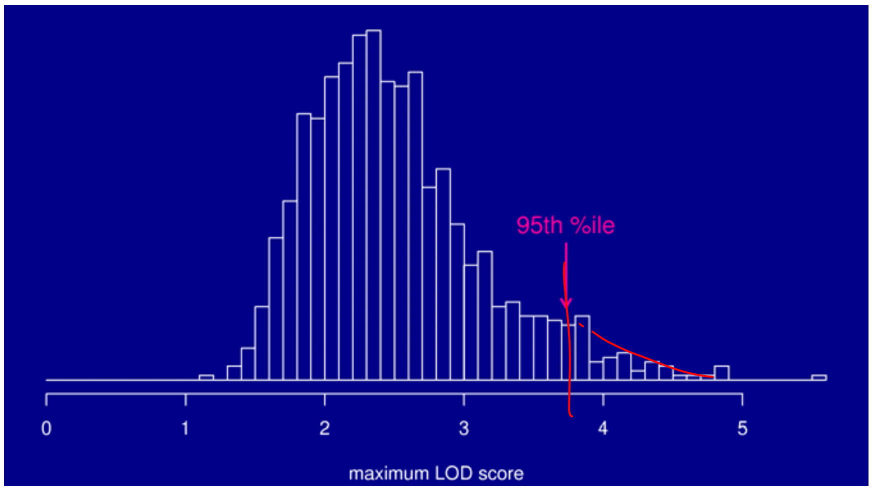
LOD score curves



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

QTL

$p(x|z)$

$P(X) = \sum_z P(X|z)P(z)$

Advantages

- Make proper account of missing data
- Can allow for the presence of genotyping errors

$P(X|z)$

QTL

- 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*

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

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

$f(\longleftarrow \quad)$

Finding QTL as model selection

Select class of models
- Additive models
- Additive plus pairwise interactions
- Regression trees

Compare models (γ)
- $BIC_{\delta}(\gamma) = \log RSS(\gamma) + \gamma(\delta \log n/n)$
- 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

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