Introduction to QTL mapping

Karl Broman
Biostatistics and Medical Informatics
University of Wisconsin – Madison

rqtl.org
kbroman.org
github.com/kbroman
@kwbromanc
Human vs mouse
Backcross

P1 × P2 → F1

P1 × F1 → BC
Intercross

P₁ × P₂

F₁ × F₁

F₂
Phenotype data

Sugiyama et al. (2002) Physiol Genomics 10:5–12
Genetic map

Chromosome Location (cM)
Genotype data
**Goals**

- Identify quantitative trait loci (QTL) (and interactions among QTL)
- Interval estimates of QTL location
- Estimated QTL effects
Statistical structure

The missing data problem: Markers $\leftarrow\rightarrow$ QTL

The model selection problem: QTL, covariates $\rightarrow$ phenotype
• Also known as marker regression.
• Split mice into groups according to genotype at a marker.
• Do a t-test / ANOVA.
• Repeat for each marker.
ANOVA at marker loci

Advantages
- Simple.
- Easily incorporates covariates.
- Easily extended to more complex models.
- Doesn’t require a genetic map.

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

Lander & Botstein (1989)

• Assume a single QTL model.

• Each position in the genome, one at a time, is posited as the putative QTL.

• Let \( q \) = the unobserved QTL genotype
  Assume \( y | q \sim N(\mu_q, \sigma) \)

• We don’t know \( q \), but we can calculate \( \Pr(q | \text{marker data}) \)

• Estimate \( \mu_q, \sigma \) by *maximum likelihood* using an iterative EM algorithm
Calculate \( P_r(q \mid \text{marker data}) \), assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)
Calculate $\Pr(q \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)
Calculate $Pr(q \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)
Calculate $Pr(q \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)
Genotype probabilities

Calculate \( \Pr(q \mid \text{marker data}) \), assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)
Calculate $P_r(q \mid \text{marker data})$, assuming

- No crossover interference
- No genotyping errors

Or use the hidden Markov model (HMM) technology

- To allow for genotyping errors
- To incorporate dominant markers
- (Still assume no crossover interference.)
EM algorithm

Dempster et al. (1977)

E step:
Let \( w_{ij}^{(k)} = \Pr(q_i = j | y_i, \text{marker data}, \hat{\mu}_0^{(k-1)}, \hat{\mu}_1^{(k-1)}, \hat{\sigma}^{(k-1)}) \)

\[
= \frac{p_{ij} f(y_i; \hat{\mu}_j^{(k-1)}, \hat{\sigma}^{(k-1)})}{\sum_j p_{ij} f(y_i; \hat{\mu}_j^{(k-1)}, \hat{\sigma}^{(k-1)})}
\]

M step:
Let \( \hat{\mu}_j^{(k)} = \frac{\sum_i y_i w_{ij}^{(k)}}{\sum_i w_{ij}^{(k)}} \)
\( \hat{\sigma}^{(k)} = \sqrt{\frac{\sum_i \sum_j w_{ij}^{(k)} (y_i - \hat{\mu}_j^{(k)})^2}{n}} \)

The algorithm:
Start with \( w_{ij}^{(1)} = p_{ij} \); iterate the E & M steps until convergence.
LOD scores

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

$$LOD(\lambda) = \log_{10} \text{likelihood ratio comparing the hypothesis of a QTL at position } \lambda \text{ versus that of no QTL}$$

$$= \log_{10} \left\{ \frac{\Pr(y|\text{QTL at } \lambda, \hat{\mu}_0\lambda, \hat{\mu}_1\lambda, \hat{\sigma}_\lambda)}{\Pr(y|\text{no QTL}, \hat{\mu}, \hat{\sigma})} \right\}$$

\(\hat{\mu}_0\lambda, \hat{\mu}_1\lambda, \hat{\sigma}_\lambda\) are the MLEs, assuming a single QTL at position \(\lambda\).

No QTL model: The phenotypes are independent and identically distributed (iid) \(N(\mu, \sigma^2)\).
Results

The diagram shows LOD scores for different chromosomes, with two lines representing body weight and heart weight. The LOD scores are plotted against chromosome numbers, indicating regions of interest for further genetic analysis.
• read.cross()

• summary(), plot()

• nind(), nmar(), totmar(), nchr(), nphe()

• calc.genoprob()

• scanone()
Interval mapping

Advantages

• Takes proper account of missing data.
• Allows examination of positions between markers.
• Gives improved estimates of QTL effects.
• Provides pretty graphs.

Disadvantages

• Increased computation time.
• Requires specialized software.
• Difficult to generalize.
• Only considers one QTL at a time.
Large LOD scores indicate evidence for the presence of a QTL

Question: How large is large?

**LOD threshold** = 95 \%ile of distr’n of max LOD, genome-wide, if there are no QTLs anywhere

**Derivation:**
- Analytical calculations (L & B 1989)
- Simulations (L & B 1989)
- Permutation tests (Churchill & Doerge 1994)
• Null distribution derived by computer simulation of backcross with genome of typical size.

• Dashed curve: distribution of LOD score at any one point.

• Solid curve: distribution of maximum LOD score, genome-wide.
Permutation test

- Individuals
- Genotype (data)
- Markers
- Phenotypes
- LOD scores
- Maximum LOD score
Permutation results

Genome-wide maximum LOD score
• `scanone()` for permutations
LOD support intervals

1.5-LOD support interval
• `lodint()`

• `bayesint()`
Haley-Knott regression

A quick approximation to Interval Mapping.

\[ E(y_i|q_i) = \mu_q \]

\[ E(y_i|M_i) = E[E(y_i|q_i)|M_i] = \sum_j Pr(q=j|M_i)\mu_j \]

\[ = \sum_j p_{ij}\mu_j \]

Regress \( y \) on \( p_i \), pretending the residual variation is normally distributed (with constant variance).

\[ LOD = \frac{n}{2} \log_{10} \left( \frac{RSS_0}{RSS_1} \right) \]
• `scanone()` with `method="hk"`
Haley-Knott results
H-K with selective genotyping
Data diagnostics

- Plot phenotypes
- Look for sample duplicates
- Look for excessive missing data
- Investigate segregation distortion
- Verify genetic maps/marker positions
- Look for genotyping errors
- Look at counts of crossovers

See Ch 3 in the R/qtl book, rqtl.org/book
• The estimated effect of a QTL will vary somewhat from its true effect.
• Only when the estimated effect is large will the QTL be detected.
• Among those experiments in which the QTL is detected, the estimated QTL effect will be, on average, larger than its true effect.
• This is selection bias.
• Selection bias is largest in QTLs with small or moderate effects.
• The true effects of QTLs that we identify are likely smaller than was observed.
Implications

- Estimated % variance explained by identified QTLs
- Repeating an experiment
- Congenics (aka near isogenic lines)
- Marker-assisted selection
Modelling multiple QTL

- Reduce residual variation $\Rightarrow$ increased power
- Separate linked QTL
- Identify interactions among QTL
Epistasis in BC

Additive

Epistatic
Epistasis in $F_2$

Additive

Epistatic
References

  A review for non-statisticians.

  Chapter on QTL mapping.

  The seminal paper.

  LOD thresholds by permutation tests.
References

  Discusses selection bias in estimated QTL effects.

  Two-part model; also discusses binary traits and non-parametric QTL mapping.

  Haley-Knott regression

  Multiple imputation

  Additive and interactive covariates.

  Multiple-QTL model selection with additive QTL.

  Also account for epistasis.