Mapping multiple QTL in experimental crosses

Karl W Broman
Biostatistics & Medical Informatics
University of Wisconsin – Madison

www.biostat.wisc.edu/~kbroman

Human vs mouse

Backcross

$P_1 \times \rightarrow \ \ P_2$

$P_1 \times \ \ F_1$

$\rightarrow \ \ BC$

www.daviddeen.com

250 male mice from the backcross (A x B) x B
Blood pressure after two weeks drinking water with 1% NaCl
Goals

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

Statistical structure

```
Markers ←→ QTL
```

```
QTL
```

```
Covariates
```

```
Phenotype
```

The missing data problem: Markers ←→ QTL

The model selection problem: QTL, covariates → phenotype

ANOVA at marker loci

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

Interval mapping

Lander & Botstein (1989)

- Assume a single QTL model.
- Consider each position in the genome, one at a time, as the location of the putative QTL.
- Let $q = 0/1$ if the (unobserved) QTL genotype is BB/AB.
  (Or 0/1/2 if the QTL genotype is AA/AB/BB in an intercross.)

Assume $y | q \sim N(\mu_q, \sigma)$

$\text{Calculate } p_q = \Pr(q | \text{marker data}).$

$y | \text{marker data} \sim \sum_q p_q \phi(y | \mu_q, \sigma)$
The normal mixtures

- Two markers separated by 20 cM, with the QTL closer to the left marker.
- The figure at right shows the distributions of the phenotype conditional on the genotypes at the two markers.
- The dashed curves correspond to the components of the mixtures.

LOD scores

\[
\text{LOD}(\lambda) = \log_{10} \text{likelihood ratio comparing the hypothesis of a QTL at position } \lambda \text{ versus that of no QTL} \\
= \log_{10} \left\{ \frac{\Pr(y|\text{QTL at } \lambda, \hat{\mu}_q, \hat{\sigma}_\lambda)}{\Pr(y|\text{no QTL}, \hat{\mu}, \hat{\sigma})} \right\} \\
\]

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

No QTL model: The phenotypes are iid \(N(\mu, \sigma^2)\).

Permutation test
Permutation results

Genome-wide maximum LOD score

LOD curves

Modeling multiple QTL

- Reduce residual variation → increased power
- Separate linked QTL
- Identify interactions among QTL (epistasis)

Estimated effects

Genotype blood pressure

Chr 1 @ 48 cM

Chr 4 @ 30 cM

Chr 6 @ 24 cM

Chr 15 @ 20 cM

Estimated effects

Chr 1 @ 48 cM

Chr 4 @ 30 cM

Chr 6 @ 24 cM

Chr 15 @ 20 cM
Estimated effects

Hypothesis testing?

- In the past, QTL mapping has been regarded as a task of hypothesis testing.

  Is this a QTL?

  Much of the focus has been on adjusting for test multiplicity.

- It is better to view the problem as one of model selection.

  What set of QTL are well supported?
  Is there evidence for QTL-QTL interactions?

  Model = a defined set of QTL and QTL-QTL interactions (and possibly covariates and QTL-covariate interactions).

Model selection

- Class of models
  - Additive models
  - + pairwise interactions
  - + higher-order interactions
  - Regression trees

- Model fit
  - Maximum likelihood
  - Haley-Knott regression
  - extended Haley-Knott
  - Multiple imputation
  - MCMC

- Model comparison
  - Estimated prediction error
  - AIC, BIC, penalized likelihood
  - Bayes

- Model search
  - Forward selection
  - Backward elimination
  - Stepwise selection
  - Randomized algorithms

Target

- Selection of a model includes two types of errors:
  - Miss important terms (QTLs or interactions)
  - Include extraneous terms

- Unlike in hypothesis testing, we can make both errors at the same time.

- Identify as many correct terms as possible, while controlling the rate of inclusion of extraneous terms.
What is special here?

- Goal: identify the major players
- A continuum of ordinal-valued covariates (the genetic loci)
- Association among the covariates
  - Loci on different chromosomes are independent
  - Along chromosome, a very simple (and known) correlation structure

Additive QTL

Simple situation:
- Dense markers
- Complete genotype data
- No epistasis

\[ y = \mu + \sum \beta_j q_j + \epsilon \]

\[ \text{LOD}_\alpha(\gamma) = \text{LOD}(\gamma) - T |\gamma| \]

0 vs 1 QTL: \( \text{LOD}_\alpha(\emptyset) = 0 \)
\[ \text{LOD}_\alpha(\lambda) = \text{LOD}(\lambda) - T \]

Automation

- Assistance to the masses
- Understanding performance
- Many phenotypes

Experience

- Controls rate of inclusion of extraneous terms
- Forward selection over-selects
- Forward selection followed by backward elimination works as well as MCMC
- Need to define performance criteria
- Need large-scale simulations

Broman & Speed, JRSS B 64:641-656, 2002
Epistasis

\[ y = \mu + \sum \beta_j q_j + \sum \gamma_{jk} q_j q_k + \epsilon \]

\[ \text{LOD}_{\alpha}(\gamma) = \text{LOD}(\gamma) - T_m |\gamma|_m + T_i |\gamma|_i \]

\( T_m = \) as chosen previously

\( T_i = ? \)

Idea 1

Imagine there are two additive QTL and consider a 2d, 2-QTL scan.

\[ T_i = \text{95th percentile of the distribution of} \]
\[ \max_{s, t} \text{LOD}_i(s, t) - \max_{s, t} \text{LOD}_j(s, t) \]

For the mouse genome:

\( T_m = 2.69 \) (BC) or 3.52 (F\(_2\))

\( T_i^\text{H} = 2.62 \) (BC) or 4.28 (F\(_2\))

\( T_i^\text{L} = 1.19 \) (BC) or 2.69 (F\(_2\))

Models as graphs

Idea 2

Imagine there is one QTL and consider a 2d, 2-QTL scan.

\[ T_m + T_i = \text{95th percentile of the distribution of} \]
\[ \max_{s, t} \text{LOD}_i(s, t) - \max \text{LOD}_j(s, t) \]

For the mouse genome:

\( T_m = 2.69 \) (BC) or 3.52 (F\(_2\))

\( T_i^\text{H} = 2.62 \) (BC) or 4.28 (F\(_2\))

\( T_i^\text{L} = 1.19 \) (BC) or 2.69 (F\(_2\))
Results

Add an interaction?

Add another QTL?

Add another QTL?

Add another QTL?

Add another QTL?

Add another QTL?

Add another QTL?
Add a pair of QTL?

\[ T_m = 2.69 \quad T_{Hi} = 2.62 \quad T_i^L = 1.19 \quad T_m + T_{Hi} = 5.31 \quad T_m + T_i^L = 3.88 \]

To do

- Improve search procedures
- Study performance (especially relative to other approaches)
- Measuring model uncertainty
- Measuring uncertainty in QTL location

Summary

- QTL mapping is a model selection problem
- The criterion for comparing models is most important
- We’re focusing on a penalized likelihood method and are close to a practiceable solution

Acknowledgments

Ani Manichaikul  
Johns Hopkins University

Gary Churchill  
Jackson Laboratory

Saunak Sen  
University of California, San Francisco

Terry Speed  
University of California, Berkeley

Brian Yandell  
University of Wisconsin, Madison

Fumihiro Sugiyama  
now at University of Tsukuba, Japan

Bev Paigen  
Jackson Laboratory