Introduction to QTL mapping in experimental crosses

Karl W Broman

Outline

- Experiments and data
- Models
- LOD thresholds
- ANOVA at marker loci
- Interval mapping
- Marker error
- Covariates
- Non-normal traits
- Selection bias
- Genotyping errors
- Errors in the map
- Selection bias
- How many markers/mice?
- Power to detect QTLs

http://biosun01.biostat.jhsph.edu/~kbroman

The Johns Hopkins University
Department of Biostatistics

in experimental crosses
With each of the parental strains, individuals are genetically identical.

Environmental variation may or may not be consistent with genotypic identity.

Phenotypes: Commonal variation. Individuals genetic as well as environmental variation may contribute to variation in the phenotype. For the backcross generation, genotype may contribute to variation in the phenotype. Within each of the parental strains, individuals are genetically identical.

Phenotypes: Data and Goals

Goals:

• Estimate QTL effects.
• Form confidence intervals for QTL locations.
• Identify the (or at least one) genomic regions (QTLs) that contribute to variation in the phenotype.

Genetic map: (for a backcross) $x_{ij} = 1/0$ if mouse i is BB/AB at marker j $y_{ij} = 1$ if phenotype for mouse i

Phenotype distributions

Phenotypes: Genotypes:

Parental strains

F1 generation

Backcross generation

Data and Goals
We assume: Mendel’s rules

No crossover interference

We assume: Mendel’s rules

Models: Recombination

Models: Genotype Phenotype

\[
(\ell | x \sim I + f^{\ell} x \sim I + f^{\ell} x = \ell + f^{\ell} x)
\]

\[
\begin{align*}
\text{Markov chain:} & \quad \ell_p \\
\text{is the genetic distance in Morgans:} & \quad \ell = \ell_p \\
\text{recombination fraction} & \quad \ell = (x | 0 = I + f^{\ell} x) \sim I + f^{\ell} x \\
\text{form a Markov chain with transition probabilities:} & \quad \ell / I = (x | 0 = f^{\ell} x) \sim I + f^{\ell} x
\end{align*}
\]
The simplest method: ANOVA

\[ (1 - 2) \nabla = (\nabla \tau + \nabla \varphi) - (\nabla \varphi - \nabla \tau) \]

\[ \nabla \tau + \nabla \varphi = \nabla \varphi + \nabla \tau \]

Consider a marker linked to the QTL, with \( t = \text{Recom.frac.} \).

Consider the case of a single QTL with effect \( \nabla = \nabla \text{QTL} = \nabla \text{QTL} \).

\[ \text{Phenotype at D1M30} \]

\[ \text{Phenotype at D2M99} \]

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

\[ \text{Genotype at D1M30} \]

\[ \text{Genotype at D2M99} \]
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.
- Only considers one QTL at a time.
- Suffers in low density scans.
- OTL localization.
- Imperfect information about QTL genotype.

---

### OTL genotype

<table>
<thead>
<tr>
<th>QTL genotype</th>
<th>AB</th>
<th>AB</th>
<th>BB</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotypes at linked markers, $P_T(z = 1</td>
<td>\text{marker data})$:</td>
<td>( \frac{\alpha - 1}{\alpha + \beta} )</td>
<td>( \frac{\alpha}{\alpha + \beta} )</td>
<td>( \frac{\beta}{\alpha + \beta} )</td>
</tr>
</tbody>
</table>

- Assume a single QTL model.
- Each position in the genome, one at a time, is tested as the putative QTL.

### Interval mapping (IM)

Lander & Botstein (1989)

\[
\begin{align*}
    (AB) & \text{ AB} \\
    (AB) & \text{ AB} \\
    (AB) & \text{ AB} \\
    (AB) & \text{ AB} \\
    (AB) & \text{ AB} \\
    (AB) & \text{ AB} \\
    (AB) & \text{ AB} \\
    (AB) & \text{ AB} \\
    (AB) & \text{ AB} \\
\end{align*}
\]
The normal mixtures

The components of the mixtures.

The dashed curves correspond to
markers.

Ional on the genotypes at the two
abortions of the phenotype condi-
marker.

The figure at right show the dis-
Two markers separated by 20 cm.

Let

where

\[ \log \mathcal{P}(\hat{\theta} | \text{marker data}) = \frac{1}{2} \log \mathcal{P}(\hat{\theta} | \text{marker data}) \]

\[ = \frac{1}{2} \log \mathcal{P}(\hat{\theta} | \text{marker data}) \]

\[ \sim \left\{ \begin{array}{ll}
\hat{\theta} & \text{if } \mathcal{P}(\hat{\theta} | \text{marker data}) = 1
\end{array} \right. \]
No QTL model: The phenotypes are independent and identically distributed (iid) 

\[
\text{LOD scores}
\]

Start with \( \hat{\theta} \) iterate the E & M steps until convergence. 

The algorithm:

M step: 

\[
\frac{\left(\varphi^{(d)}\right) f^{(d-1)} + \left(\varphi^{(d)} \varphi^{(d)} \nabla + \left(\varphi^{(d)} \right) f^{(d)} \right)}{\left(\varphi^{(d)} + \varphi^{(d)} \nabla + \left(\varphi^{(d)} \right) f^{(d)} \right) f^{(d)}}
\]

E step:

\[
\left(\varphi^{(d)} \nabla \nabla^{(d)} \right) f^{(d)} \left| \right. \hat{\theta} = \text{marker data.}
\]

Dempster et al. (1977)
An example LOD curve.
Interval mapping

Advantages

• Investigates interactions between QTLs (epistasis).
• Separates linked QTLs.
• Reduces residual variation.

Disadvantages

• Only considers one QTL at a time.
• Difficult to generalize.
• Requires specialized software.
• Time-consuming and computation.

Multiple QTL methods

Why consider multiple QTLs at once?

Advantages

• Provides pretty graphs.
• Determines the effects of QTLs.
• Gives improved estimates of QTL positions between markers.
• Allows examination of interactions between QTLs (epistasis).
• Takes proper account of missing data.
Large LOD scores indicate evidence for the presence of a QTL.

Null distribution of the LOD score

- Mutation (randomization) tests.
- Estimating the threshold: simulations, analytical calculations, per-
- LOD threshold.

- Genome-wide. The 95th percentile of this distribution serves as a genome-wide
- We seek the distribution of the maximum LOD score, genome-
- genome-wide.

- Multiple putative QTL locations.
- Key point: We must make some adjustment for our examination of
- hypotheses of no QTL.
- We consider the distribution of the LOD score under the null

- How large is large?

Large LOD thresholds
Permutation tests

Value: conditions on observed phenotypes, marker density, and pattern of missing data. doesn’t rely on normally assumption or asymptotics.

We can’t look at all possible permutations, but a random set of 1000 is feasible.

The 95th %ile of LOD is a genome-wide LOD threshold.

We wish to compare the observed LOD to the distribution of LOD.

Calculate LOD = max(LOD)*z. (z)

Permute/shuffle the phenotypes; keep the genotype data intact.

LOD support intervals

Plot of LOD vs chromosome position (cM). Values: conditions on observed phenotypes, marker density, and pattern of missing data. doesn’t rely on normally assumption or asymptotics.

The 95th %ile of LOD is a genome-wide LOD threshold.

We wish to compare the observed LOD to the distribution of LOD.

Calculate LOD = max(LOD)*z. (z)

Permute/shuffle the phenotypes; keep the genotype data intact.
The power to detect a QTL is the chance that its LOD score exceeds the genome-wide threshold.

### Power to detect QTLs

<table>
<thead>
<tr>
<th>QTL Effect</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2%</td>
</tr>
<tr>
<td>8</td>
<td>41%</td>
</tr>
<tr>
<td>11</td>
<td>97%</td>
</tr>
</tbody>
</table>

### Note: The figures should be taken with a grain of salt.

OTL effect, and luck!

- Depends on number of mice, size of QTL effect.
- Not necessarily increased precision.
- More detailed genotype information.

### More Markers:

- Reduced sampling variation.
- More recombination breakpoints.

### More Mice:

- Dashed: 1 cm spacing
- Solid: 10 cm spacing
- Top: $n = 100$
- Bottom: $n = 200$

### How many markers/mice?

- More markers/mice:
  - More detailed genotype information.
  - Not necessarily increased precision (depends on number of mice, size of QTL effect, and luck).

- Dashed curve: distance of LOD score at $n = 200$.
- Solid curve: distance of LOD score at $n = 100$.
- Dotted curve: distance of max LOD score under null hypothesis.

### Right:

- Significance of the LOD threshold.
- Density of markers.
- Type of cross.
- Number of progeny.
- Size of the QTL effect.

Power depends on the genome-wide threshold.
Selection bias

- The estimated effect of a QTL will vary somewhat from its true effect. Only when the estimated effect is very somewhat from its true effect will the QTL be detected.
- Among those experiments in which the QTL is detected, the estimated QTL effect will, on average, be 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.

Support intervals:
- Perceived length of LOD makes a big difference in perceived length of LOD.
- Does not seem to make much difference.

Map distances:
- A signal should not completely eliminate a signal.
- Causes wiggly LOD curves.

The genetic map: effects of errors

- Greater effects of errors with support intervals.
- Markers in marker order.
- Causes wiggly LOD curves.
- Does not seem to make much difference.
- Error in marker order.
- Map distances.
- Error in marker spacing.
The genetic map: finding problems

- Estimated genetic map
- Pairwise recombination fractions
- Misplaced markers
- Big gaps in the map
- Large recombination fractions

Genotyping errors: effects

- With genotyping errors, individuals are placed in the wrong genotype group.
- With widely spaced markers, there is little effect.
- With dense markers, errors make the LOD curve have more dips.
Identifying genotyping errors

- Look for tight double crossovers. (Crossover interference is often strong.)
- Error LOD scores (Lincoln & Lander 1992)
  - Assuming no interference:
    - Assumed error rate
    - Model for genotyping errors.
  - Strong errors:
    - Crossovers.
    - Look for tight double

Selective genotyping

- Save effort by only typing the most informative individuals.
- Useful in context of a single, inexpensive trait.
- Can’t get all interactions.
- Likely better to also genotype some random individuals.
- Can't get all phenotypes.

Tricky to estimate the effects of QTLs: use IM with all phenotypes.

Phenotype

BBAB

Top and Bottom 10%

Markers

Individuals

Chromosome 5

Position (cM)

Genotyping error LOD scores

Error LOD scores

Strong errors:

Crossovers.

(Chromosomal interference is often strong.)
Covariates

- Examples: treatment, sex, litter, lab, age.
- Treatment, sex, litter, lab, age.
- Transformed residuals (e.g., log, square root)
- Nonparametric approaches (Kruglyak & Lander 1995)

Control residual variation.

Avoid confounding.

Look for QTL environment interactions.

Adjust before interval mapping (IM) versus adjusted interval mapping, with LOD thresholds derived from permutation tests, generally performs just fine anyway.

In reality: we see dichotomous traits, counts, skewed distributions, outliers, and all sorts of odd things.

Standard interval mapping assumes normally distributed residual variation. Thus the phenotype distribution is a mixture of normals.

Non-normal traits

Standard interval mapping assumes normally distributed residuals.

Adjust before interval mapping (IM) versus adjusted interval mapping.

Look for QTL × environment interactions.

Avoid confounding.

Control residual variation.

Examples: treatment, sex, litter, lab, age.
ANOVA software is not available for your preferred model, start with some version of consider transformations or specialy-tailored models. If interval mapping
intermediate works reasonably well even with non-normal traits. But

Interval mapping improves phenomema.

The consideration of covariates may reveal extremely

Incredible phenomema. But don't worry about them too much.

Study your data. Look for errors in the genetic map, genotyping errors and

Intermediary too large.

Estimates of QTL effects are subject to selection bias. Such estimated effects

Study your data. Much more important than more markers. But this depends on the number of mice,

Once you've achieved a 10-cM marker spacing, more mice will probably be
depicting evidence for QTL location.

The LOD curve, re-centered so that its maximum is at 0, is a valuable tool for
the LOD curve, re-centered so that its maximum is at 0, is a valuable tool for
a plot of 1.5-LOD support intervals indicate the plausible location of a QTL. A plot of

Tests are extremely useful for this.

Statistical significance of LOD peaks requires consideration of the maximum
spurious...

The better separation of linked QTLs and are necessary for the investigation of
ANOVA and multiple single-QTL models. Multiple QTL methods allow
data.

Interval mapping improves on ANOVA by allowing inference of QTLs to
To include covariates or accommodate complex models.

ANOVA at marker loci (aka marker regression) is simple and easily extended