Module 17: Advanced QTL Mapping
Zhao-Bang Zeng, Brian S. Yandell
Presentation Schedule

Monday
Unit pages title
1 ZBZ 2-24 Multiple interval mapping (MIM)
2 ZBZ 25-36 Score statistic to aid MIM model selection

lunch
A BSY 75-93 Overview: multiple QTL approaches
B BSY 94-103 R/qtl software demo

Tuesday
Unit pages title
C BSY 114-136 Bayesian interval mapping (BIM)
D BSY 103-113 R/qtlbim software demo

lunch
E BSY 137-147 Data examples in detail
148-157 Multiple traits and co-mapping
3 ZBZ 37-57 Multiple trait analysis

Wednesday
4 ZBZ 58-74 Gene expression QTL (eQTL) analysis
F BSY 158-170 eQTL Tools
171-204 Causal networking
Multiple Interval Mapping

Zhao-Bang Zeng
Department of Statistics
North Carolina State University

Purpose of Multiple Interval Mapping

- Simultaneously search and map multiple QTL: more powerful and accurate, though computationally intensive.
- Study QTL epistasis: Epistasis is a context-dependent phenomenon. It is important to study the epistasis in its totality (as much as possible) jointly with QTL main effects.
- Comprehensive estimation of genetic architecture of quantitative traits: number, positions, effects and interactions of QTL; Partition of genetic variance due to each QTL effect; ...
- Produce efficient prediction function for marker-assisted selection, pharmacogenetics jointly with co-factors, ....

It combines QTL mapping analysis with the study of genetic architecture of quantitative traits and marker-assisted selection.
QTL Mapping

Trait Phenotypic Values: Y

Genetic Model: $P(Y|G, \theta)$

QTL: G

$P(G|X, \lambda)$

Markers: X

Likehood of Data: $P(Y, X) = P(Y|X) P(X) = \sum_G P(Y|G, \theta) P(G|X, \lambda) P(X|\delta)$
Statistical Framework

- Data:
  \( Y = \) Quantitative trait phenotype
  \( X = \) Molecular marker genotypes

- Joint probability of data:
  \[
P(Y, X) = P(Y|X)P(X)
  \]

  \( P(X) \): Marker analysis, including segregation and linkage analysis, linkage phase inference, and linkage map construction.

  \( P(Y|X) \): QTL analysis which is to analyze the relationship between markers and trait through QTL, and through this analysis to infer the genetic structure (or architecture) of quantitative traits, such as the number, genomic position, allelic frequencies, effects and interaction of QTL.

Marker Likelihood Analysis

The likelihood of marker data can be symbolically denoted as
\[
P(X|\gamma, \phi, \omega)
\]

- \( \gamma \) = recombination frequencies between markers
- \( \phi \) = marker linkage phases
- \( \omega \) = marker linkage order

Through this likelihood analysis, we can infer \( \gamma, \phi, \omega \).
**QTL Analysis**

QTL analysis contains two parts:

\[ P(Y|X) = \sum_G P(Y|G)P(G|X) \]

\( P(G|X) \) involves the segregation analysis of QTL given marker genotypes and is a function of QTL position \( \lambda \).

\( P(Y|G) \) is a link function between QTL genotypes \( G \) and trait phenotypes \( Y \), and can be modeled as a function of QTL effect parameters \( \theta \), such as the additive, dominance and epistatic effects of QTL and any other parameters that link QTL genotypes to trait phenotypes.

Together, \( \lambda \) and \( \theta \) represent the genetic architecture parameters of quantitative traits.

Thus, \( P(Y|X, \theta, \lambda) = \sum_G P(Y|G, \theta)P(G|X, \lambda) \)

---

**MIM model**

For \( m \) putative QTL, the multiple interval mapping model (for a backcross population) is defined by

\[ y_i = \mu + \sum_{r=1}^{m} \alpha_r x_{ir}^* + \sum_{r \neq s \in [1, \ldots, m]} \beta_{rs}(x_{ir}^*x_{is}^*) + e_i \]

where

- \( y_i \) is the phenotypic value of individual \( i \);
- \( i \) indexes individuals of the sample: \( i = 1, 2, \ldots, n \);
- \( \mu \) is the mean of the model;
- \( \alpha_r \) is the marginal effect of putative QTL \( r \);
- \( x_{ir}^* \) is a coded variable denoting the genotype of putative QTL \( r \) (defined by 1/2 or -1/2 for the two genotypes), which is unobserved but can be inferred from maker data in sense of
probability;

- $\beta_{rs}$ is the epistatic effect between putative QTL $r$ and $s$;
- $r \neq s \in \{1, \cdots, m\}$ denotes a subset of QTL pairs that each shows a significant epistatic effect to avoid the over-parameterization that could result when using all pairs;
- $m$ is the number of putative QTL chosen based on either their significant marginal effects or significant epistatic effects;
- $t$ is the number of significant pairwise epistatic effects;
- $e_i$ is a residual effect of the model assumed to be normally distributed with mean zero and variance $\sigma^2$.

**Likelihood**

The likelihood function of the data given the model is a mixture of normal distributions

\[
L(E, \mu, \sigma^2) = \prod_{i=1}^{n} \left[ \sum_{j=1}^{2m} p_{ij} \phi(y_i | \mu + G_j E, \sigma^2) \right]
\]

- $p_{ij}$ is the probability of each multilocus genotype conditional on marker data;
- $E$ is a vector of QTL parameters ($\alpha$’s and $\beta$’s);
- $G_j$ is a vector specifying the configuration of $x^*$’s associated with each $\alpha$ and $\beta$ for the $j$th QTL genotype;
- $\phi(y | \mu, \sigma^2)$ denotes a normal density function.
Calculate \( P(G|X) \)

**Example:** For a backcross population, let a QTL \( (g) \) be located between two markers \((x_1 \text{ and } x_2)\). Let the recombination frequency between \(x_1\) and \(g\) be \(r_1\), that between \(g\) and \(x_2\) be \(r_2\), and between \(x_1\) and \(x_2\) be \(r_{12}\). The conditional probability \( P(g|x_1, x_2) \) is:

<table>
<thead>
<tr>
<th>Marker genotype, ( x_1, x_2 )</th>
<th>QTL genotype, ( g )</th>
<th>Freq</th>
<th>( 1 )</th>
<th>( 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 11 )</td>
<td>( \frac{1-r_{12}}{2} )</td>
<td>( \frac{(1-r_1)(1-r_2)}{1-r_{12}} )</td>
<td>( 1 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( 10 )</td>
<td>( \frac{r_{12}}{2} )</td>
<td>( \frac{(1-r_1)r_2}{r_{12}} )</td>
<td>( 1 - \lambda )</td>
<td>( \frac{r_1(1-r_2)}{r_{12}} )</td>
</tr>
<tr>
<td>( 01 )</td>
<td>( \frac{r_{12}}{2} )</td>
<td>( \frac{r_1(1-r_2)}{r_{12}} )</td>
<td>( \lambda )</td>
<td>( \frac{(1-r_1)r_2}{r_{12}} )</td>
</tr>
<tr>
<td>( 00 )</td>
<td>( \frac{1-r_{12}}{2} )</td>
<td>( \frac{r_2}{1-r_{12}} )</td>
<td>( 0 )</td>
<td>( \frac{(1-r_1)(1-r_2)}{1-r_{12}} )</td>
</tr>
</tbody>
</table>

where \( \lambda = \frac{r_1}{r_{12}} \)

For multiple QTL in multiple marker intervals

\[
P(G|X, \lambda) = \prod_{r=1}^{m} P(G_r|X, \lambda_r)
\]

For more complicated cases, a hidden Markov model is generally used for the calculation.
EM algorithm
Take derivative of $\log L$ with respect to each model parameter ($\mu$, $E_r$, $\sigma^2$) and equate the derivatives to zero, we can obtain a series of functions with the following algorithm (called EM algorithm).

EM is an iterative procedure involving an E-step (Expectation) and M-step (Maximization) in each iteration. In the $[t+1]$th iteration, E-step:

$$\pi_{ij}^{[t+1]} = \frac{p_{ij} \phi(y_i | \mu^{[t]} + D_j E^{[t]}, \sigma^2[j])}{\sum_{j=1}^{2^m} p_{ij} \phi(y_i | \mu^{[t]} + D_j E^{[t]}, \sigma^2[j])}$$

M-step:

$$E_r^{[t+1]} = \sum_i \sum_j \pi_{ij}^{[t+1]} G_{jr} [y_i - \mu^{[t]}] - \sum_{s=1}^{r-1} G_{js} E_s^{[t+1]} - \sum_{s=r+1}^{m+t} G_{js} E_s^{[t]}]$$

$$\mu^{[t+1]} = \frac{1}{n} \sum_i \left( y_i - \sum_j \sum_r \pi_{ij}^{[t+1]} G_{jr} E_r^{[t+1]} \right)$$

$$\sigma^2^{[t+1]} = \frac{1}{n} \left[ \sum_i (y_i - \mu^{[t+1]})^2 - 2 \sum_i (y_i - \mu^{[t+1]}) \sum_j \sum_r \pi_{ij}^{[t+1]} G_{jr} E_r^{[t+1]} \right] + \sum_r \sum_s \sum_i \sum_j \pi_{ij}^{[t+1]} G_{jr} G_{js} E_r^{[t+1]} E_s^{[t+1]}$$
**Conditional QTL genotype probability** $\pi_{ij}$

Probability of QTL genotype conditional on marker genotype:

$$P(G_j|X_i) = p_{ij}$$

Conditional density of trait phenotype given QTL genotype:

$$P(Y_i|G_j) = \phi(y_i|\mu + G_jE, \sigma^2)$$

Probability of QTL genotype conditional on marker genotype and trait phenotype:

$$P(G_j|X_i, Y_i) = \pi_{ij} = \frac{P(G_j|X_i)P(Y_i|G_j)}{\sum_G P(G_j|X_i)P(Y_i|G_j)}$$

$$= \frac{p_{ij}\phi(y_i|\mu + G_jE, \sigma^2)}{\sum_j p_{ij}\phi(y_i|\mu + G_jE, \sigma^2)}$$

---

**Advantage and disadvantage of EM algorithm**

- **Advantages:** Generally stable i.e. likelihood increases steadily and do not diverge; easy to implement; easy to generalize to more complicated situation (such as multiple QTL with epistasis)
- **Disadvantages:** Very slow to converge to optimum
GEM-NR algorithm

Speed up the converging process while preserving the property of stability. Combine EM (generalized EM) with a quadratic algorithm in the maximization step (Newton-Ralphson algorithm).

In the E-step, still calculate the conditional probability ($\pi$'s) of QTL genotypes given current estimates of model parameters.

The M-step is obtained by

$$
\theta^{(t+1)} = \theta^{(t)} - \alpha^{(t)} \frac{\partial Q(\theta|\theta^{(t)})}{\partial \theta} \bigg|_{\theta^{(t)}} - \frac{1}{2} \left( \theta^{(t+1)} - \theta^{(t)} \right)' \frac{\partial^2 Q(\theta|\theta^{(t)})}{\partial \theta \partial \theta'} \bigg|_{\theta^{(t)}} \left( \theta^{(t+1)} - \theta^{(t)} \right)
$$

where $Q(\theta|\theta^{(t)})$ is the expected complete-data loglikelihood conditional on the current estimates of parameter values.

Comments on the algorithms

- GEM-NR algorithm (Generalized EM with Newton-Ralphson algorithm):

$$
Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) = (\theta^{(t+1)} - \theta^{(t)} - \theta^{(t)})' \frac{\partial Q(\theta|\theta^{(t)})}{\partial \theta} \bigg|_{\theta^{(t)}} - \frac{1}{2} (\theta^{(t+1)} - \theta^{(t)})' \frac{\partial^2 Q(\theta|\theta^{(t)})}{\partial \theta \partial \theta'} \bigg|_{\theta^{(t)}} (\theta^{(t+1)} - \theta^{(t)})
$$

- EM algorithm (Expectation-Maximization algorithm):

$$
Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) = (\theta^{(t+1)} - \theta^{(t)})' \frac{\partial Q(\theta|\theta^{(t)})}{\partial \theta} \bigg|_{\theta^{(t)}}
$$

- NR algorithm (Newton-Ralphson algorithm):

$$
l(\theta^{(t+1)}) - l(\theta^{(t)}) = (\theta^{(t+1)} - \theta^{(t)})' \frac{\partial l(\theta)}{\partial \theta} \bigg|_{\theta^{(t)}} - \frac{1}{2} (\theta^{(t+1)} - \theta^{(t)})' \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta'} \bigg|_{\theta^{(t)}} (\theta^{(t+1)} - \theta^{(t)})
$$
Dealing with many QTL

- $m$ QTL $\rightarrow 2^m$ possible mixture components.
- This can be prohibitive for efficient numerical analysis.
- But most genotypes have negligible probabilities.
- Can we skip these evaluations?

Practical implementation of MIM algorithm:
Select a subset of “significant” mixture components for each individual for evaluation: (1) set any $p_{ij} < \delta$ (= 0.005) to zero (drop them); (2) Sum of “significant” $p_{ij} > 0.95$ (adjust $\delta$ if needed); (3) normalize “significant” probs: $\sum_j p_{ij} = 1$.

Number of “significant” components $\sim$ 10-100, depending on marker density, number and position of QTL. It has negligible loss of accuracy of likelihood evaluation as compared to no selection.

Conditional likelihood ratio test
Test for each QTL effect $E_r$ conditional on other QTL effects:

$$LOD = \log_{10} \frac{L(\text{all } E_s \neq 0)}{L(E_r = 0, \text{ all other } E_s \neq 0)}$$

It can proceed as above if we have positions of $m$ putative QTL and selected $m + t$ QTL effects.

How do we search for multiple QTL?
How do we decide how many QTL to include?
How do we select best genetic model? (number, positions, gene action, epistasis)

Criterion: fit data well in some sense.
Interactive model selection procedure in Window QTL Cartographer

1. Initial model: select New Model to use an automatical stepwise selection procedure, CIM or stepwise marker selection for initial model selection.

2. Search for new QTL: select Refine Model => Search for New QTL => Search for QTL to scan the genome for new QTL and determine whether to accept QTL based on the selected criterion.

3. Search for QTL epistasis: select Refine Model => Search for New QTL => Search for Epistasis to search for epistatic effects among identified QTL based on the selected criterion (better to use a lower criterion such as AIC).

4. Re-evaluate QTL effects: select Refine Model => Testing for Existing QTL to re-evaluate the significance of each QTL effect in the model based on the selected criterion. This procedure can remove non-significant effects from the model.

5. Optimize QTL positions: select Refine Model => Optimizing QTL Position to optimize QTL position estimates in the current model. QTL position is optimized one by one in a sequential order.

6. Procedure 2 to 5 can be repeated if needed.

7. Selection criterion: Quite a few provided (such as BIC and AIC).

8. Display graphic result: select Refine Model => MIM Model Summary => Graphic Result File to calculate and display the likelihood profile for each QTL.

9. Show MIM estimates: select Refine Model => MIM Model Summary => Model Summary File to show the MIM output result file. Information includes position, likelihood ratio and
effect of each QTL, epistatic effects of QTL, partition of the variance explained by QTL (due to main and interaction effects), estimate of genotypic values of individuals based on the model.

Model selection criteria

- Akaike information criterion (AIC): minimize $-2(\log L_k - k)$.
- $C_p$ method: minimize adjusted $R^2$.
- Bayes information criterion (BIC): minimize $-2(\log L_k - kc(n)/2)$ with $c(n) = \log(n)$, or $c(n) = 2\log(\log n)$, or other penalty function.
- Final prediction error (FPE) method: minimize prediction error.
- Delete-one cross-validation, Delete-$d$ cross-validation, and generalized cross-validation: different ways to implement FPE.
- Bootstrap model selection: use bootstrap resampling to implement FPE.
- Minimizing posterior predictive loss: similar to FPE in concept.
Residual permutation test

Suppose that we want to test the hypotheses: $H_0$: $k$ QTL vs. $H_1$ $(k + 1)$ QTL. Assume that the $k$ QTL of $H_0$ is contained in the model of $(k + 1)$ QTL of $H_1$.

Let $\hat{y}_{i|H_0}$ and $\hat{y}_{i|H_1}$ be the estimated genotypic value of individual $i$ under $H_0$ and $H_1$, respectively.

**Residual permutation sample:** Let $\hat{\epsilon}_i = y_i - \hat{y}_{i|H_0}$ and $\bar{\epsilon} = \sum_i \hat{\epsilon}_i / n$. To generate a residual permutation sample $\{(X_i^*, Y_i^*)\}$, we first generate a random sample of residuals $\{\epsilon_i^*\}$ from $\{\hat{\epsilon}_j, j = 1, \cdots, n\}$ without replacement, and define $X_i^* = X_i$ and $Y_i^* = \hat{y}_{i|H_0} + \epsilon_i^*$.

Residual permutation test is performed as follows:

1. Draw a residual permutation sample $\{(X_i^*, Y_i^*)\}$.
2. Search for the best position in the genome (other than the positions of the $k$ QTL) for the hypothetical $k + 1$ QTL and perform the likelihood ratio test for the hypotheses.
3. Repeat step 1 and 2 for a predetermined number of times to obtain an empirical bootstrap distribution of the test statistic, $T^*$.
4. Reject $H_0$ if the test statistic in the original data exceeds $\hat{T}_\alpha$, where $\hat{T}_\alpha$ is the $(1 - \alpha)$th quantile of the bootstrap distribution of $T^*$. 

Slide 25

Slide 26
Estimating the variance explained by QTL

Variance explained by QTL effect $E_r$ can be estimated as

$$\hat{\sigma}^2_{E_r} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{2^m} \hat{\pi}_{ij} (G_{jr} - \bar{G}_r)^2 \hat{E}_r^2$$

Covariance explained by QTL effect $E_r$ and $E_s$ is

$$\hat{\sigma}_{E_r, E_s} = \frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{2^m} \hat{\pi}_{ij} (G_{jr} - \bar{G}_r)(G_{js} - \bar{G}_s) \hat{E}_r \hat{E}_s$$

Thus the total genetic variance explained by QTL is

$$\hat{\sigma}^2_g = \sum_r \hat{\sigma}^2_{E_r} + \sum_{r \neq s} \hat{\sigma}_{E_r, E_s}$$

It is convenient and may also be informative to combine the variance due to each QTL effect with half of the covariances between this QTL effect and other effects and report it as the variance component associated with this QTL effect

$$\hat{\sigma}^2_r = \hat{\sigma}^2_{E_r} + \frac{1}{2} \sum_{s \neq r} \hat{\sigma}_{E_r, E_s}$$
Estimation of genotypic values

The genotypic value of an individual can be estimated as:

\[ \hat{y}_i = \hat{\mu} + \sum_{j=1}^{m+t} \sum_{r=1}^{m+t} \hat{\pi}_{ij} G_{jr} \hat{E}_r \]

To predict the genotypic values of quantitative traits based on marker information only (e.g. in cross-prediction; early selection), we need to use

\[ \hat{y}_i = \hat{\mu} + \sum_{j=1}^{m+t} \sum_{r=1}^{m+t} \hat{\pi}_{ij} G_{jr} \hat{E}_r \]

as \( \hat{\pi}_{ij} \) is a function of phenotype \( \hat{y}_i \) which is unavailable in early selection.

These can be used for marker-assisted selection.

MIM example: Genetic architecture of wing size of *Drosophila melanogaster* on chromosome 2 (Ken Weber):

- Population: 701 recombinant inbred lines originating from a cross between high and low selected lines on wing size. Only QTL on chromosome 2 are segregating in the population, and other chromosomes are identical for all RIL.
- Trait: wing size measured in radian in an allometric analysis.
- 10 QTL are identified by MIM analysis. There is a good agreement between the sum of estimated additive effects of QTL and the observed parental genotype difference.
- There are some significant additive by additive interaction effects between QTL. The interaction pattern is complex.
- Together, 10 additive and 14 additive by additive QTL effects explain 95% of the total variance in the population.
### Statistics of several models in the backward stepwise elimination process for selecting significant QTL epistatic effects

<table>
<thead>
<tr>
<th>Model</th>
<th>QTL #</th>
<th>Epis #</th>
<th>2ln(Likelihood)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_{45}$</td>
<td>10</td>
<td>45</td>
<td>7094</td>
<td>0.952</td>
</tr>
<tr>
<td>$M_{14}$</td>
<td>10</td>
<td>14</td>
<td>7088</td>
<td>0.954</td>
</tr>
<tr>
<td>$M_{13}$</td>
<td>10</td>
<td>13</td>
<td>7075</td>
<td>0.954</td>
</tr>
<tr>
<td>$M_{5}$</td>
<td>10</td>
<td>5</td>
<td>7058</td>
<td>0.953</td>
</tr>
<tr>
<td>$M_{4}$</td>
<td>10</td>
<td>4</td>
<td>7041</td>
<td>0.953</td>
</tr>
<tr>
<td>$M_{3}$</td>
<td>10</td>
<td>3</td>
<td>7008</td>
<td>0.952</td>
</tr>
<tr>
<td>$M_{0}$</td>
<td>10</td>
<td>0</td>
<td>6997</td>
<td>0.947</td>
</tr>
</tbody>
</table>

### Estimates of QTL positions and additive effects

<table>
<thead>
<tr>
<th>QTL</th>
<th>Posi (cM)</th>
<th>Effect ($\times 10^{-2}$)</th>
<th>Effect (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1.35</td>
<td>20.9</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>0.78</td>
<td>12.2</td>
</tr>
<tr>
<td>3</td>
<td>37</td>
<td>0.32</td>
<td>4.9</td>
</tr>
<tr>
<td>4</td>
<td>47</td>
<td>0.54</td>
<td>8.4</td>
</tr>
<tr>
<td>5</td>
<td>54</td>
<td>0.86</td>
<td>13.4</td>
</tr>
<tr>
<td>6</td>
<td>58</td>
<td>0.64</td>
<td>9.9</td>
</tr>
<tr>
<td>7</td>
<td>72</td>
<td>0.35</td>
<td>5.4</td>
</tr>
<tr>
<td>8</td>
<td>88</td>
<td>0.63</td>
<td>9.8</td>
</tr>
<tr>
<td>9</td>
<td>98</td>
<td>0.55</td>
<td>8.5</td>
</tr>
<tr>
<td>10</td>
<td>108</td>
<td>0.37</td>
<td>5.8</td>
</tr>
</tbody>
</table>

Total 99.1
### Estimates of QTL additive by additive interaction effects

<table>
<thead>
<tr>
<th>QTL pair</th>
<th>Effect ($\times 10^{-2}$)</th>
<th>QTL pair</th>
<th>Effect ($\times 10^{-2}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1&amp;2)</td>
<td>1.49</td>
<td>(2&amp;8)</td>
<td>-0.22</td>
</tr>
<tr>
<td>(1&amp;3)</td>
<td>-1.10</td>
<td>(3&amp;4)</td>
<td>-1.08</td>
</tr>
<tr>
<td>(1&amp;5)</td>
<td>-0.52</td>
<td>(3&amp;7)</td>
<td>0.13</td>
</tr>
<tr>
<td>(1&amp;9)</td>
<td>-0.17</td>
<td>(4&amp;6)</td>
<td>0.33</td>
</tr>
<tr>
<td>(2&amp;4)</td>
<td>1.48</td>
<td>(8&amp;9)</td>
<td>0.73</td>
</tr>
<tr>
<td>(2&amp;5)</td>
<td>0.27</td>
<td>(8&amp;10)</td>
<td>-0.80</td>
</tr>
<tr>
<td>(2&amp;6)</td>
<td>-0.36</td>
<td>(9&amp;10)</td>
<td>0.89</td>
</tr>
</tbody>
</table>

### Estimated variances and covariances of the QTL additive effects in percentage of the total phenotypic variance

<table>
<thead>
<tr>
<th>QTL 1 2 3 4 5 6 7 8 9 10 Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 16.6 7.0 0.9 0.2 -1.5 -1.1 -0.7 -1.6 -2.2 -1.7 15.6</td>
</tr>
<tr>
<td>2 7.0 5.6 1.1 1.1 0.6 0.3 -0.1 -0.6 -1.2 -1.0 12.8</td>
</tr>
<tr>
<td>3 0.9 1.1 0.9 1.2 1.4 1.0 0.3 0.1 -0.2 -0.2 6.4</td>
</tr>
<tr>
<td>4 0.2 1.1 1.2 2.7 3.5 2.4 0.8 0.6 -0.2 -0.3 11.8</td>
</tr>
<tr>
<td>5 -1.5 0.6 1.4 3.5 6.8 4.8 1.6 1.1 -0.2 -0.5 17.4</td>
</tr>
<tr>
<td>6 -1.1 0.3 1.0 2.4 4.8 3.7 1.2 0.9 -0.1 -0.3 12.9</td>
</tr>
<tr>
<td>7 -0.7 -0.1 0.3 0.8 1.6 1.2 1.1 1.2 0.5 0.2 6.2</td>
</tr>
<tr>
<td>8 -1.6 -0.6 0.1 0.6 1.1 0.9 1.2 3.6 2.2 1.2 8.7</td>
</tr>
<tr>
<td>9 -2.2 -1.2 -0.2 -0.2 -0.2 -0.1 0.5 2.2 2.8 1.6 3.0</td>
</tr>
<tr>
<td>10 -1.7 -1.0 -0.2 -0.3 -0.5 -0.3 0.2 1.2 1.6 1.3 0.3</td>
</tr>
<tr>
<td>Total 95.1</td>
</tr>
</tbody>
</table>
### Estimated variances and covariances of the QTL additive by additive interaction effects in percentage of the total phenotypic variance

<table>
<thead>
<tr>
<th>QTLs</th>
<th>1&amp;2</th>
<th>1&amp;3</th>
<th>1&amp;5</th>
<th>1&amp;9</th>
<th>2&amp;4</th>
<th>2&amp;5</th>
<th>2&amp;6</th>
<th>2&amp;8</th>
<th>3&amp;4</th>
<th>3&amp;7</th>
<th>4&amp;6</th>
<th>8&amp;9</th>
<th>8&amp;10</th>
<th>9&amp;10</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1&amp;2</td>
<td>2.4</td>
<td>-1.2</td>
<td>-0.4</td>
<td>0.0</td>
<td>-0.9</td>
<td>-0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>-0.1</td>
<td>0.2</td>
<td>-0.1</td>
<td>-0.6</td>
</tr>
<tr>
<td>1&amp;3</td>
<td>-1.2</td>
<td>2.7</td>
<td>0.8</td>
<td>0.0</td>
<td>-1.6</td>
<td>-0.2</td>
<td>0.3</td>
<td>0.0</td>
<td>-0.4</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>-0.3</td>
<td>0.1</td>
<td>0.8</td>
</tr>
<tr>
<td>1&amp;5</td>
<td>-0.4</td>
<td>0.8</td>
<td>0.6</td>
<td>0.0</td>
<td>-1.1</td>
<td>-0.2</td>
<td>0.3</td>
<td>0.0</td>
<td>0.2</td>
<td>0.0</td>
<td>-0.1</td>
<td>0.1</td>
<td>-0.2</td>
<td>0.0</td>
<td>0.3</td>
</tr>
<tr>
<td>1&amp;9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>2&amp;4</td>
<td>-0.9</td>
<td>-1.6</td>
<td>-1.1</td>
<td>0.0</td>
<td>4.7</td>
<td>0.7</td>
<td>-0.9</td>
<td>-0.2</td>
<td>-1.0</td>
<td>0.0</td>
<td>-0.2</td>
<td>0.2</td>
<td>-0.3</td>
<td>-0.1</td>
<td>-0.4</td>
</tr>
<tr>
<td>2&amp;5</td>
<td>-0.2</td>
<td>-0.2</td>
<td>-0.2</td>
<td>0.0</td>
<td>0.7</td>
<td>0.2</td>
<td>-0.2</td>
<td>0.0</td>
<td>-0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>-0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>-0.1</td>
</tr>
<tr>
<td>2&amp;6</td>
<td>0.2</td>
<td>0.3</td>
<td>0.3</td>
<td>0.0</td>
<td>-0.9</td>
<td>-0.2</td>
<td>0.3</td>
<td>0.0</td>
<td>0.2</td>
<td>0.0</td>
<td>-0.1</td>
<td>0.1</td>
<td>-0.1</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>2&amp;8</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-0.2</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>-0.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>3&amp;4</td>
<td>0.2</td>
<td>-0.4</td>
<td>0.2</td>
<td>0.0</td>
<td>-1.0</td>
<td>-0.2</td>
<td>0.2</td>
<td>0.0</td>
<td>1.2</td>
<td>-0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>3&amp;7</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>4&amp;6</td>
<td>0.0</td>
<td>0.1</td>
<td>-0.1</td>
<td>0.0</td>
<td>-0.2</td>
<td>0.0</td>
<td>-0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>8&amp;9</td>
<td>-0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>0.0</td>
<td>-0.2</td>
<td>-0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.6</td>
<td>-0.6</td>
<td>-0.1</td>
<td>-0.1</td>
</tr>
<tr>
<td>8&amp;10</td>
<td>0.2</td>
<td>-0.3</td>
<td>-0.2</td>
<td>0.0</td>
<td>0.3</td>
<td>0.1</td>
<td>-0.1</td>
<td>-0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-0.6</td>
<td>1.0</td>
<td>-0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>9&amp;10</td>
<td>-0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>-0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-0.1</td>
<td>-0.4</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**Total**  0.3
Use of score statistic to aid model selection for mapping multiple QTL

Zhao-Bang Zeng
Departments of Statistics & Genetics
Bioinformatics Research Center
North Carolina State University
Email: zeng@stat.ncsu.edu

Multiple Interval Mapping (MIM)

We developed MIM in 1999 (Kao et al. 1999; Zeng et al. 1999) and implemented it in Windows QTL Cartographer (Wang et al. 2001-2007).

MIM is a model selection-based method for identifying multiple QTL (mostly sequentially), and based on the identification to estimate the parameters of genetic architecture including epistasis.

A few model selection criteria, such as BIC, were implemented. However, it is still not quite clear what appropriate criteria should be used for model selection.

The criteria should take into account a number of experimental factors, such as genome size, genetic map density, informativeness of markers, and proportion of missing data.
MIM Model and Likelihood

Model (for \( m \) putative QTL in a backcross population):

\[
y_i = \mu + \sum_{k=1}^{m} a_k x_{ik} + \sum_{k \neq l \in \{1, \ldots, m\}} \delta_{kl} \gamma_{kl} x_{ik} x_{il} + \epsilon_i.
\]

where \( x_{ik} \) is unobserved QTL genotype with known conditional probability from genetic markers and \( \epsilon_i \sim N(0, 1) \). Likelihood:

\[
L(\theta; v) = \prod_{i=1}^{n} \sum_{j=1}^{2^m} P(G_j|X_i) P(Y_i|G_j) = \prod_{i=1}^{n} \left[ \sum_{j=1}^{2^m} p_{ij} \phi(y_i|\mu + G_j E, \sigma^2) \right]
\]

\[
l(\theta; v) = \sum_{i=1}^{n} l_i(\theta; v) = \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{2^m} p_{ij} \phi(y_i|\mu_j, \sigma^2) \right\}
\]

We have worked out an efficient algorithm that combines generalized EM and Newton-Raphson method (GEM-NR) to maximize the likelihood for complex genetic models.

Model Search and Model Selection

There are several ways to do model selection with MIM. One way is to perform sequential search, adding one parameter at a time.

- Starting with no QTL, scan the genome and compute test statistics for adding one QTL at a time. The putative QTL position with the maximum test statistic is added to the model if the statistic exceeds a specified threshold. This process is repeated until no more QTL is found.

- Then we search for adding parameters for epistasis between pairs of QTL identified.

- The model can be refined iteratively by dropping insignificant parameters and searching for new parameters if necessary.

The key ingredient in the model selection is choosing an appropriate test statistic and corresponding threshold value.
Score Statistic

Zou et al. (2004 Genetics 168:2307-2316) proposed using score statistic to test QTL effect and a resampling procedure for determining the appropriate threshold.

Suppose we have identified the \( m - 1 \) QTL with parameters \( \eta \) and want to test for adding the \( m \)th QTL with parameter \( \beta \).

Let \( U(d) \) denote the score function for \( \beta \), at genomic position \( d \), evaluated at \( \beta = 0 \) and \( \hat{\eta} \).

\[
\hat{U}_i(d) = U_{\beta,i}(0, \hat{\eta}; d) = \left( \frac{\partial^2 l(0, \hat{\eta}; d)}{\partial \beta \partial \eta} \right) \left( \frac{\partial^2 l(0, \hat{\eta}; d)}{\partial \eta^2} \right)^{-1} U_{\eta,i}(0, \hat{\eta}; d)
\]

\[
U_{\beta,i}(\beta, \eta; d) = \frac{\partial l_i(\beta, \eta; d)}{\partial \beta}
\]

\[
U_{\eta,i}(\beta, \eta; d) = \frac{\partial l_i(\beta, \eta; d)}{\partial \eta} = \left( \frac{\partial l_i}{\partial \theta_1}, \cdots, \frac{\partial l_i}{\partial \theta_{m-1}}, \frac{\partial l_i}{\partial \mu}, \frac{\partial l_i}{\partial \sigma^2} \right)'
\]

\[
\hat{U}(d) = \sum_{i=1}^{n} \hat{U}_i(d)
\]

The score statistic for \( H_0: \beta = 0 \) against \( H_1: \beta \neq 0 \) at location \( d \) is

\[
W(d) = \hat{U}'(d)\hat{V}^{-1}(d)\hat{U}(d)
\]

where \( \hat{V}(d) = \sum_{i=1}^{n} \hat{U}_i(d)\hat{U}_i'(d) \).
Resampling with Score Statistic
An efficient way to simulate conditional null distribution

1. Generate $G_i, i = 1, 2, \cdots, n$ from $N(0, 1)$.

2. Calculate $U^*(d) = \sum_{i=1}^{n} \hat{U}_i(d) G_i$, $W^*(d) = U^*(d) \tilde{V}^{-1} U^*(d)$, and $S^* = \max_d W^*(d)$.

3. Repeat step 1 and 2 for $N$ times to find $S_k^*$ for $k = 1, \cdots, N$.

4. Compute the $100(1 - \alpha)^{th}$ percentile of $\{S_k^* : k = 1, \cdots, N\}$ to determine the threshold value.

5. Accept the position being tested as identifying a new QTL if the observed score statistic for the position exceeds the threshold value.
Simulation

Case I:
• 9 chromosomes each with 12 markers in 10cM interval
• 8 QTL with equal effect on different chromosomes
• Heritability = 0.4
• 1000 times score statistics bootstraps
• 1000 Replications

Case II:
• 9 chromosomes each with 12 markers in 10cM interval
• 8 QTL with different effects and some linkage
• Heritability = 0.4
• 1000 times score statistics bootstraps
• 1000 Replications

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>6</th>
<th>7</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positions (cM)</td>
<td>27.4</td>
<td>90.3</td>
<td>49.0</td>
<td>32.5</td>
<td>88.9</td>
<td>9.3</td>
<td>70.7</td>
<td>63.2</td>
</tr>
<tr>
<td>QTL effect</td>
<td>0.503</td>
<td>0.670</td>
<td>0.798</td>
<td>0.590</td>
<td>-0.503</td>
<td>0.710</td>
<td>0.255</td>
<td>0.399</td>
</tr>
</tbody>
</table>
Q-Q plot of score resampling test statistic vs. likelihood ratio with permutation at null (no QTL) (Case I)

Comparison of thresholds of score statistic and Likelihood ratio with permutation (Case I)
Statistical power of detecting QTL for Case I

Statistical power of detecting QTL for Case II
False discovery rate (FDR) of QTL identification by using different genome-wide significance level and different LOD support interval for defining true or false identification (case I and II)

<table>
<thead>
<tr>
<th>α level</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOD-1.0</td>
<td>I</td>
<td>0.090</td>
<td>0.094</td>
<td>0.100</td>
<td>0.105</td>
<td>0.110</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>0.090</td>
<td>0.107</td>
<td>0.112</td>
<td>0.119</td>
<td>0.127</td>
</tr>
<tr>
<td>LOD-1.5</td>
<td>I</td>
<td>0.042</td>
<td>0.046</td>
<td>0.051</td>
<td>0.056</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>0.045</td>
<td>0.052</td>
<td>0.057</td>
<td>0.063</td>
<td>0.071</td>
</tr>
<tr>
<td>LOD-2.0</td>
<td>I</td>
<td>0.023</td>
<td>0.028</td>
<td>0.033</td>
<td>0.038</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>0.023</td>
<td>0.029</td>
<td>0.034</td>
<td>0.038</td>
<td>0.043</td>
</tr>
</tbody>
</table>

Confidence of LOD support interval (% of true QTL inside the interval, case II)

<table>
<thead>
<tr>
<th>α level</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOD</td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
</tr>
</tbody>
</table>

| QTL-1   | 0.879| 0.952| 0.985| 0.887| 0.954| 0.984| 0.889| 0.955| 0.979|
| QTL-2   | 0.911| 0.960| 0.981| 0.917| 0.962| 0.981| 0.921| 0.967| 0.982|
| QTL-3   | 0.938| 0.969| 0.981| 0.939| 0.967| 0.976| 0.930| 0.959| 0.970|
| QTL-4   | 0.926| 0.978| 0.993| 0.922| 0.976| 0.989| 0.916| 0.973| 0.985|
| QTL-5   | 0.929| 0.970| 0.987| 0.923| 0.970| 0.985| 0.925| 0.966| 0.980|
| QTL-6   | 0.938| 0.971| 0.984| 0.935| 0.970| 0.983| 0.934| 0.968| 0.980|
| QTL-7   | 0.709| 0.842| 0.923| 0.673| 0.828| 0.918| 0.686| 0.825| 0.920|
| QTL-8   | 0.800| 0.911| 0.953| 0.806| 0.905| 0.956| 0.818| 0.907| 0.956|

Average 0.879 0.944 0.973 0.875 0.942 0.972 0.877 0.940 0.969
## Confidence of LOD support interval

(% of true QTL inside the interval, case II)

<table>
<thead>
<tr>
<th>α level</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOD</td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>QTL-1</td>
<td>0.888</td>
<td>0.954</td>
<td>0.978</td>
</tr>
<tr>
<td>QTL-2</td>
<td>0.924</td>
<td>0.968</td>
<td>0.983</td>
</tr>
<tr>
<td>QTL-3</td>
<td>0.928</td>
<td>0.954</td>
<td>0.969</td>
</tr>
<tr>
<td>QTL-4</td>
<td>0.914</td>
<td>0.970</td>
<td>0.983</td>
</tr>
<tr>
<td>QTL-5</td>
<td>0.921</td>
<td>0.964</td>
<td>0.981</td>
</tr>
<tr>
<td>QTL-6</td>
<td>0.923</td>
<td>0.961</td>
<td>0.972</td>
</tr>
<tr>
<td>QTL-7</td>
<td>0.689</td>
<td>0.829</td>
<td>0.924</td>
</tr>
<tr>
<td>QTL-8</td>
<td>0.818</td>
<td>0.904</td>
<td>0.958</td>
</tr>
<tr>
<td>Average</td>
<td>0.876</td>
<td>0.938</td>
<td>0.969</td>
</tr>
</tbody>
</table>

## LOD support interval width (cM)

(case II)

<table>
<thead>
<tr>
<th>α level</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOD</td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>QTL-1</td>
<td>18.9</td>
<td>26.3</td>
<td>34.8</td>
</tr>
<tr>
<td>QTL-2</td>
<td>13.1</td>
<td>17.4</td>
<td>21.9</td>
</tr>
<tr>
<td>QTL-3</td>
<td>11.7</td>
<td>15.1</td>
<td>18.6</td>
</tr>
<tr>
<td>QTL-4</td>
<td>17.4</td>
<td>24.0</td>
<td>32.3</td>
</tr>
<tr>
<td>QTL-5</td>
<td>17.4</td>
<td>23.7</td>
<td>33.0</td>
</tr>
<tr>
<td>QTL-6</td>
<td>12.5</td>
<td>15.9</td>
<td>19.3</td>
</tr>
<tr>
<td>QTL-7</td>
<td>22.9</td>
<td>33.7</td>
<td>47.8</td>
</tr>
<tr>
<td>QTL-8</td>
<td>22.4</td>
<td>33.5</td>
<td>46.7</td>
</tr>
<tr>
<td>Average</td>
<td>17.0</td>
<td>23.7</td>
<td>31.8</td>
</tr>
</tbody>
</table>
### LOD support interval width (cM)
*(case II)*

<table>
<thead>
<tr>
<th>α level</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOD 1.0</td>
<td>20.9</td>
<td>21.0</td>
<td>21.5</td>
</tr>
<tr>
<td>LOD 1.5</td>
<td>29.2</td>
<td>29.4</td>
<td>29.9</td>
</tr>
<tr>
<td>LOD 2.0</td>
<td>38.7</td>
<td>39.3</td>
<td>40.0</td>
</tr>
<tr>
<td>QTL-1</td>
<td>13.4</td>
<td>13.7</td>
<td>13.7</td>
</tr>
<tr>
<td>QTL-2</td>
<td>17.9</td>
<td>18.1</td>
<td>18.2</td>
</tr>
<tr>
<td>QTL-3</td>
<td>22.4</td>
<td>22.7</td>
<td>22.9</td>
</tr>
<tr>
<td>QTL-4</td>
<td>17.7</td>
<td>17.5</td>
<td>17.6</td>
</tr>
<tr>
<td>QTL-5</td>
<td>18.0</td>
<td>18.0</td>
<td>18.1</td>
</tr>
<tr>
<td>QTL-6</td>
<td>24.9</td>
<td>25.0</td>
<td>25.3</td>
</tr>
<tr>
<td>QTL-7</td>
<td>33.9</td>
<td>33.4</td>
<td>33.3</td>
</tr>
<tr>
<td>QTL-8</td>
<td>11.9</td>
<td>12.0</td>
<td>12.8</td>
</tr>
<tr>
<td>QTL-9</td>
<td>15.4</td>
<td>15.6</td>
<td>16.9</td>
</tr>
<tr>
<td>QTL-10</td>
<td>19.1</td>
<td>19.3</td>
<td>21.5</td>
</tr>
<tr>
<td>QTL-11</td>
<td>17.5</td>
<td>18.2</td>
<td>22.9</td>
</tr>
<tr>
<td>QTL-12</td>
<td>22.7</td>
<td>22.9</td>
<td>31.8</td>
</tr>
<tr>
<td>QTL-13</td>
<td>17.6</td>
<td>18.2</td>
<td>33.3</td>
</tr>
<tr>
<td>QTL-14</td>
<td>24.3</td>
<td>25.3</td>
<td>36.4</td>
</tr>
<tr>
<td>QTL-15</td>
<td>43.6</td>
<td>46.9</td>
<td>65.4</td>
</tr>
<tr>
<td>QTL-16</td>
<td>27.1</td>
<td>29.9</td>
<td>39.3</td>
</tr>
<tr>
<td>QTL-17</td>
<td>40.0</td>
<td>43.3</td>
<td>56.9</td>
</tr>
<tr>
<td>Average</td>
<td>18.1</td>
<td>25.6</td>
<td>35.0</td>
</tr>
</tbody>
</table>

### QTL parameter estimation
*(averaged over detected replicates for each QTL, Case II)*

<table>
<thead>
<tr>
<th>QTL</th>
<th>Position</th>
<th>Mean Estimate</th>
<th>SD</th>
<th>Effect</th>
<th>Mean Estimate</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27.4</td>
<td>27.4</td>
<td>7.47</td>
<td>0.503</td>
<td>0.557</td>
<td>0.112</td>
</tr>
<tr>
<td>2</td>
<td>90.3</td>
<td>89.6</td>
<td>4.59</td>
<td>0.670</td>
<td>0.716</td>
<td>0.141</td>
</tr>
<tr>
<td>3</td>
<td>49.0</td>
<td>49.1</td>
<td>3.76</td>
<td>0.798</td>
<td>0.816</td>
<td>0.139</td>
</tr>
<tr>
<td>4</td>
<td>32.5</td>
<td>31.5</td>
<td>6.22</td>
<td>0.590</td>
<td>0.629</td>
<td>0.141</td>
</tr>
<tr>
<td>5</td>
<td>88.9</td>
<td>89.3</td>
<td>6.43</td>
<td>-0.503</td>
<td>-0.572</td>
<td>0.126</td>
</tr>
<tr>
<td>6</td>
<td>9.3</td>
<td>9.9</td>
<td>4.04</td>
<td>0.710</td>
<td>0.732</td>
<td>0.128</td>
</tr>
<tr>
<td>7</td>
<td>70.7</td>
<td>68.8</td>
<td>16.76</td>
<td>0.255</td>
<td>0.463</td>
<td>0.093</td>
</tr>
<tr>
<td>8</td>
<td>63.2</td>
<td>62.7</td>
<td>12.91</td>
<td>0.399</td>
<td>0.485</td>
<td>0.093</td>
</tr>
</tbody>
</table>
General Conclusion

- For mapping multiple QTL, sequential search via MIM offers a promising solution.
- It is appropriate to use model- and data-based criterion to aid for the model selection, such as score statistic and its resampling threshold. It takes the data and model complexity into account and computationally is very efficient. However, the threshold does not seem to change in different cycle of model search process, and may need to be calculated only once to save more computational time.

General Conclusion

- It is too conservative to use genome-wide significance level 0.05 for model selection. With significance level 0.1 to 0.2, FDR is still controlled approximately at 0.05. If higher FDR, such as 0.1, can be tolerated, we can even use significance level 0.2 to 0.4 for model selection to increase statistical power to detect relatively weaker QTL.
- It is more appropriate to use 1.5-LOD, rather than 1-LOD, support interval to report likely region of QTL. For moderate and strong QTL, 1.5-LOD support interval can give approximately 95% confidence interval for QTL position.
Multiple Trait Analysis: Genetic Basis of Trait Correlation

Zhao-Bang Zeng
Department of Statistics
North Carolina State University
Email: zeng@stat.ncsu.edu
URL: http://statgen.ncsu.edu/zeng/zeng.html

Why multiple traits/environments?
- Most QTL mapping experiments record observations on multiple traits.
- Different attributes of a general biological character: For example
  - Size of an area (e.g. wing size): size vs. shape.
  - Body weight: different parts of the weight.
  - Fitness components: viability vs. fecundity.
- Same trait at different developmental stages: Developmental aspects of the character.
- Same trait at different character states: growth in tropic area vs. growth in temporal area, grain yield in NC vs. Iowa. This is the issue of genotype by environment interaction, and can be analyzed by multiple trait analysis.
Why analyzing multiple traits/environments?

By taking into account the correlated structure of multiple traits, the joint analysis on multiple traits for mapping QTL can

- Improve the statistical power to detect QTL,
- Improve the resolution to estimate QTL positions and effects,
- Provide formal procedures to test a number of biologically interesting hypotheses concerning the nature of genetic correlations between different traits, such as
  - Joint mapping QTL (testing and estimating QTL affecting multiple traits)
  - Testing pleiotropy of QTL
  - Testing QTL × environment interaction
  - Testing whether significant effects at a genome region on multiple traits is due to pleiotropy of the same QTL or close linkage

• Provide a comprehensive estimation about the genetic architecture of quantitative traits including the structure of genetic correlations between traits.
Two types of data structures

We consider two types of data structures for multiple trait analysis.

- Design I: Multiple traits are measured on the same individuals. Data matrices may look like the following (for \( m \) traits, \( t \) markers, \( n \) individuals)

\[
Y = \begin{bmatrix}
y_{11} & y_{12} & \cdots & y_{1n} \\
y_{21} & y_{22} & \cdots & y_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
y_{m1} & y_{m2} & \cdots & y_{mn}
\end{bmatrix}
\quad \text{and} \quad
X = \begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1n} \\
x_{21} & x_{22} & \cdots & x_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
x_{t1} & x_{t2} & \cdots & x_{tn}
\end{bmatrix}
\]

- Design II: Multiple traits or trait states are measured on different individuals. Data matrices may look like the following (with one set of traits measured in population one with \( n_1 \) individuals and \( t_1 \) markers, and another set measured in another population with \( n_2 \) individuals and \( t_2 \) markers)

\[
Y_1 = \begin{bmatrix}
y_{11} & y_{12} & \cdots & y_{1n_1} \\
y_{21} & y_{22} & \cdots & y_{2n_1} \\
\vdots & \vdots & \ddots & \vdots \\
y_{m_11} & y_{m_12} & \cdots & y_{m_1n_1}
\end{bmatrix}
\quad \text{and} \quad
X_1 = \begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1n_1} \\
x_{21} & x_{22} & \cdots & x_{2n_1} \\
\vdots & \vdots & \ddots & \vdots \\
x_{t_11} & x_{t_12} & \cdots & x_{t_1n_1}
\end{bmatrix}
\]

\[
Y_2 = \begin{bmatrix}
y_{11} & y_{12} & \cdots & y_{1n_2} \\
y_{21} & y_{22} & \cdots & y_{2n_2} \\
\vdots & \vdots & \ddots & \vdots \\
y_{m_21} & y_{m_22} & \cdots & y_{m_2n_2}
\end{bmatrix}
\quad \text{and} \quad
X_2 = \begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1n_2} \\
x_{21} & x_{22} & \cdots & x_{2n_2} \\
\vdots & \vdots & \ddots & \vdots \\
x_{t_21} & x_{t_22} & \cdots & x_{t_2n_2}
\end{bmatrix}
\]
This can represent several situations: for example

- The same traits measured in two backcrosses \((B_1\) and \(B_2\)) on different individuals. In this case a test on QTL×backcross interaction is a test about dominance (and some epistasis as well) of QTL.

- The same trait measured in two sexes: test QTL×sex interaction.

- Different groups of individuals are planted in two or multiple geographic locations.

---

**Model and likelihood**

For \(m\) putative QTL of \(T\) traits in \(S\) environments/populations, the multiple interval mapping model (for a backcross population) is defined by

\[
y_{sti} = \mu_{st} + \sum_{r=1}^{m} \alpha_{str} x_{str}^* + e_{sti}
\]

where

- \(y_{sti}\) is the phenotypic value of trait \(t\) for individual \(i\) in environment/population \(s\);
- \(i\) indexes individuals of the sample: \(i = 1, 2, \cdots, n_s\);
- \(t\) indexes traits: \(t = 1, 2, \cdots, T\);
- \(s\) indexes environments/populations: \(s = 1, 2, \cdots, S\);
- \(\mu_{st}\) is the mean of the model;
- \(\alpha_{str}\) is the effect of putative QTL \(r\) on trait \(t\) in population \(s\),

22
• $x_{ir}^*$ is a coded variable denoting the genotype of putative QTL $r$ (defined by $1/2$ or $-1/2$ for the two genotypes) for individual $i$ in population $s$, which is unobserved but can be inferred from maker data in sense of probability;

• $e_{st}$ is a residual effect of the model assumed to be multivariate normal distributed with mean vector $0$ and variance matrix $V_s$.

Likelihood

The likelihood function of the data given the model is a mixture of normal distributions

$$L = \prod_{s=1}^{S} \prod_{i=1}^{n_s} \sum_{j=1}^{2^n} p_{sij} \phi(y_{si} | \mu_s + E_s D_{sj}, V_s)$$

• $p_{sij}$ is the probability of each multilocus genotype conditional on marker data;

• $E_s$ is a matrix of QTL parameters ($\alpha$’s) for population $s$;

• $D_{sj}$ is a vector specifying the configuration of $x^*$’s associated with each $\alpha$ for the $j$th QTL genotype;

• $\phi(y|\mu, V)$ denotes a multivariate normal density function for $y$ with mean vector $\mu$ and variance matrix $V$. 

23
EM algorithm

E-step:

$$\pi_{sij}^{[t+1]} = \frac{p_{sij} \phi(y_{si}|\mu_s^{[t]} + \mathbf{E}_i^t \mathbf{D}_{sj}, \mathbf{V}_s^{[t]})}{\sum_{j=1}^m p_{sij} \phi(y_{si}|\mu_s^{[t]} + \mathbf{E}_i^t \mathbf{D}_{sj}, \mathbf{V}_s^{[t]})}$$

M-step:

$$\mu_{st}^{[t+1]} = \frac{1}{n_s} \sum_t \left( y_{sti} - \sum_j \sum_{r} \pi_{sij}^{[t+1]} D_{srj} E_{str}^{[t+1]} \right)$$

$$\nu_{suw}^{[t+1]} = \frac{1}{n_u} \sum_t \left[ \sum_j \sum_{r} \pi_{sij}^{[t+1]} (y_{sui} - \mu_{su}^{[t+1]} - \sum_{r} D_{srj} E_{sur}^{[t+1]} - y_{swi} - \mu_{sw}^{[t+1]} - \sum_{r} D_{srj} E_{swr}^{[t+1]} ) \right]$$

Model selection

Model selection can proceed as in MIM. In this case, when a QTL is selected, its effects are fitted and estimated for all traits in all environments or populations, irregardless whether the QTL effect is significant for a particular trait in a particular environment.

1. Initial model: Use multivariate backward stepwise regression on markers to select an initial model.
2. Optimize the estimates of QTL positions based on the currently selected model.
3. Scan the genome to determine the best position for adding a new QTL.
4. Repeat (2) and (3) for a few times to select a few competing models.
5. If epistasis is considered, select significant epistatic terms.
6. Select the final model based on some information criterion.

Hypotheses testing

Given a selected genetic model, we can still test a number of hypotheses. Testing these hypotheses can help us to understand and interprete the genetic architecture of quantitative traits in those environments/populations.

- Joint mapping of QTL: Here we can test the hypothesis

\[ H_0 : \alpha_r = 0 \quad \text{vs.} \quad H_1 : \alpha_r \neq 0 \]

with

\[ LR = -2 \ln \frac{L(\alpha_r = 0, \hat{\mu}, \hat{\alpha}, \hat{V})}{L(\hat{\mu}, \hat{\alpha}, \hat{V})} \]
• QTL × environment interaction:

\[ H_0 : \alpha_1 = \cdots = \alpha_S \text{ or } \alpha_T \]

vs. \[ H_1 : \alpha_1 \neq \cdots \neq \alpha_S \text{ or } \alpha_T \]

with

\[ LR = -2 \ln \frac{L(\alpha_1 = \cdots = \alpha_S, \hat{\alpha}, \hat{V})}{L(\hat{\mu}, \hat{\alpha}, \hat{V})} \]

Rejection of the null hypothesis means that the QTL effects depend on environment, being different in different environments.

Note: For design II (S populations) where the joint likelihood is just a product of the separate likelihoods assuming independent samples, it is for testing this null hypothesis that the joint analysis is required.

---

**How does the joint analysis improve the power and resolution of mapping QTL?**

Joint Analysis vs. Separate Analysis

If we let the likelihood ratios under separate tests be

\[ LR_{S1} \simeq n\beta_1 \]

where \[ \beta_1 = a_1^*/(\sqrt{2}\sigma_{y,z} \cdot X) \]

\[ LR_{S2} \simeq n\beta_2 \]

where \[ \beta_2 = a_2^*/(\sqrt{2}\sigma_{y,z} \cdot X) \]

The likelihood ratio under the joint test can be approximated as

\[ LR_J \simeq n \frac{\beta_1^2 + \beta_2^2 - 2\rho \beta_1 \beta_2}{1 - \rho^2} \]
Some observations:

1. \( LR_J \geq \text{maximum}[LR_{S1}, LR_{S2}] \).
2. If \( \rho = 0 \), \( LR_J \simeq LR_{S1} + LR_{S2} \).
3. If \( \beta_2 = 0 \), \( LR_J \simeq \frac{LR_{S1}}{1-\rho^2} \geq LR_{S1} \).
4. If \( \rho \beta_1 \beta_2 < 0 \), i.e. \( \rho \) and \( \beta_1 \beta_2 \) are in different signs, \( LR_J > LR_{S1} + LR_{S2} \). In this case, \( \text{power}(J) > \text{maximum[power}(S1), \text{power}(S2)) \).

This shows that \( LR_J \) can be significantly higher than \( \text{max}[LR_{S1}, LR_{S2}] \). Of course, the threshold for \( LR_J \) is higher than that for \( LR_{S1} \) and \( LR_{S2} \) as well. However, in general we find that the power of \( LR_J \) is significantly higher than that of \( LR_{S1} \) and \( LR_{S2} \).

---

**A simulation example on joint vs. separate analysis:**

We show a simulation study with three QTL having effects on three quantitative traits, distributed on a chromosome with 150 cM length in an \( F_2 \) population with 150 sample size. We assume markers are observed at every 10cM position. Genetic parameters of QTL and estimates of QTL positions and effects (averaged over 100 replicates with standard deviations in bracket) are shown in Table 3, and phenotypic parameters and estimates in Table 4.
<table>
<thead>
<tr>
<th>QTL Position</th>
<th>Additive effect</th>
<th>Dominance effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trait 1</td>
<td>Trait 2</td>
</tr>
<tr>
<td>Parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 21.0</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2 84.0</td>
<td>-0.30</td>
<td>-1.00</td>
</tr>
<tr>
<td>3 142.0</td>
<td>-1.00</td>
<td>0.30</td>
</tr>
<tr>
<td>Estimates by J-123</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 21.0 (4.3)</td>
<td>1.00 (0.43)</td>
<td>1.00 (0.42)</td>
</tr>
<tr>
<td>2 84.9 (4.7)</td>
<td>-0.24 (0.51)</td>
<td>-1.00 (0.38)</td>
</tr>
<tr>
<td>3 142.4 (3.9)</td>
<td>-1.03 (0.38)</td>
<td>0.27 (0.27)</td>
</tr>
<tr>
<td>Estimates by J-12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 20.9 (5.2)</td>
<td>1.03 (0.43)</td>
<td>1.02 (0.41)</td>
</tr>
<tr>
<td>2 83.7 (9.6)</td>
<td>-0.25 (0.56)</td>
<td>-0.97 (0.43)</td>
</tr>
<tr>
<td>3 141.3 (6.8)</td>
<td>-1.08 (0.36)</td>
<td>0.26 (0.34)</td>
</tr>
<tr>
<td>Estimates by J-13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 22.4 (7.8)</td>
<td>1.05 (0.48)</td>
<td></td>
</tr>
<tr>
<td>2 85.6 (6.0)</td>
<td>-0.26 (0.52)</td>
<td>-1.03 (0.40)</td>
</tr>
<tr>
<td>3 143.1 (2.8)</td>
<td>-1.00 (0.38)</td>
<td>1.03 (0.30)</td>
</tr>
<tr>
<td>Estimates by J-23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 21.8 (6.5)</td>
<td>1.05 (0.41)</td>
<td>0.34 (0.40)</td>
</tr>
<tr>
<td>2 84.9 (4.4)</td>
<td>-1.01 (0.38)</td>
<td>-1.04 (0.37)</td>
</tr>
<tr>
<td>3 142.4 (4.5)</td>
<td>0.28 (0.30)</td>
<td>1.04 (0.35)</td>
</tr>
<tr>
<td>Estimates by S-1, S-2 and S-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 21.9,21.6,25.8 (7.7,5.7,13.6)</td>
<td>1.08 (0.42)</td>
<td>1.12 (0.34)</td>
</tr>
<tr>
<td>2 84.9,84.3,86.0 (17.1,8.7,6.7)</td>
<td>-0.29 (0.69)</td>
<td>-1.03 (0.38)</td>
</tr>
<tr>
<td>3 141,136,143 (6.2,11.7,4.1)</td>
<td>-1.13 (0.31)</td>
<td>0.33 (0.45)</td>
</tr>
</tbody>
</table>
Seven mapping analyses were performed and compared: J-123 (joint mapping on three traits); J-12, J-13, J-23 (joint mapping on two traits); S-1, S-2, S-3 (separate mapping on each trait).

Conclusions:

- Standard deviations of estimates of QTL positions are generally smaller with joint analyses. This shows that joint analysis can significantly improve the resolution on the estimation of QTL position.
- There is some improvement on the estimation of QTL effects, but the improvement does not seem to be very significant.
- The joint analyses also improve significantly on the power to detect QTL (Table 5).

### Table 5: Observed statistical power (proportion of significant replicates over all replicates) of seven methods of QTL mapping from 100 replicates of simulations

<table>
<thead>
<tr>
<th>QTL</th>
<th>J-123</th>
<th>J-12</th>
<th>J-13</th>
<th>J-23</th>
<th>S-1</th>
<th>S-2</th>
<th>S-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.80</td>
<td>0.78</td>
<td>0.51</td>
<td>0.64</td>
<td>0.46</td>
<td>0.64</td>
<td>0.04</td>
</tr>
<tr>
<td>2</td>
<td>0.79</td>
<td>0.37</td>
<td>0.36</td>
<td>0.84</td>
<td>0.00</td>
<td>0.39</td>
<td>0.41</td>
</tr>
<tr>
<td>3</td>
<td>0.89</td>
<td>0.51</td>
<td>0.84</td>
<td>0.64</td>
<td>0.42</td>
<td>0.00</td>
<td>0.64</td>
</tr>
</tbody>
</table>
A practical data analysis with three populations and four trait states (locations):

- Three related backcross populations with a trait evaluated in four locations \( (S = 3, T = 4, n_s = 128) \).
- QTL analysis was performed with four locations and three populations combined in model identification and estimation.
- An initial model was selected based on a combined forward-and-backward stepwise regression on markers. Then, an extensive search and position-optimization analysis was performed repeatedly under multiple interval mapping.
- The final model contains 19 QTL that have significant effects on the trait in different locations and different populations.
- The following figure plots the LOD score profile for the 19 QTL. It depicts the statistical evidence and strength of mapping for each QTL.

The LOD scores for each QTL in each population (with four locations combined) and estimates of QTL effects in each location and population provide statistical evidence and information about the significance of QTL effects in each location and population.

Estimated genetic variances, covariances and genetic correlations provide further information about the genetic architecture of QTL. These estimates can be further partitioned into components due to each individual QTL.

Genetic correlations between different locations provide an overall structure of QTL by location interaction. High genetic correlation reflects lack of QTL by location interaction. Low or negative genetic correlation reflects some or strong QTL by location interaction. It is clear that QTL x Location interaction is prevalent for many QTL in all the three populations.
For example, in population cross 3, QTL effects in location C are in an opposite direction to those in other locations for QTL 1, 7, 8, 11, 14, 15, and 17. Together, these produce the negative genetic correlations between location C and location A, B and D. In population 1, genetic correlation between location B and D is also negative, due to mostly QTL 13, 14, 17 and 18. The genetic correlation between location A and B is also low, due to mostly QTL 4, 17 and 18. The low genetic correlation between location C and D is mostly due to QTL 4, 11 and 14. In population 2, genetic correlations between location C and location B and D are also relatively lower, indicating some QTLxLocation interaction (due to mostly QTL 11 and 19).

This integrated multiple QTL oriented approach for multiple traits, environments and populations provides a comprehensive method to study the details of genetic architecture of quantitative traits in a variety of populations. It also facilitates the comparison and test of genetic basis for multiple traits in multiple environments and populations.

“Total is more than the sum!”
Experiment Design

L₁  L₂  L₃
F₁₂  F₁₃  F₂₃
T₁  T₂
B₁₂ Türkiye
B₁₃ Türkiye
B₂₃ Türkiye
## LOD score 19 putative QTL

<table>
<thead>
<tr>
<th>QTL</th>
<th>Ch:cM</th>
<th>LOD in population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1:0</td>
<td>.73</td>
</tr>
<tr>
<td>2</td>
<td>1:43</td>
<td>.76</td>
</tr>
<tr>
<td>3</td>
<td>1:92</td>
<td><strong>4.12</strong></td>
</tr>
<tr>
<td>4</td>
<td>1:95</td>
<td><strong>4.74</strong></td>
</tr>
<tr>
<td>5</td>
<td>2:2</td>
<td>2.27</td>
</tr>
<tr>
<td>6</td>
<td>3:0</td>
<td>.98</td>
</tr>
<tr>
<td>7</td>
<td>3:65</td>
<td>.03</td>
</tr>
<tr>
<td>8</td>
<td>3:113</td>
<td>1.82</td>
</tr>
<tr>
<td>9</td>
<td>3:135</td>
<td>1.16</td>
</tr>
<tr>
<td>10</td>
<td>4:64</td>
<td><strong>11.06</strong></td>
</tr>
<tr>
<td>11</td>
<td>6:0</td>
<td>2.59</td>
</tr>
<tr>
<td>12</td>
<td>6:21</td>
<td><strong>3.94</strong></td>
</tr>
<tr>
<td>13</td>
<td>7:20</td>
<td><strong>9.83</strong></td>
</tr>
<tr>
<td>14</td>
<td>7:28</td>
<td><strong>10.38</strong></td>
</tr>
<tr>
<td>15</td>
<td>8:15</td>
<td><strong>8.44</strong></td>
</tr>
<tr>
<td>16</td>
<td>9:26</td>
<td>2.32</td>
</tr>
<tr>
<td>17</td>
<td>9:54</td>
<td><strong>4.63</strong></td>
</tr>
<tr>
<td>18</td>
<td>9:58</td>
<td><strong>4.51</strong></td>
</tr>
<tr>
<td>19</td>
<td>10:63</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Estimates of QTL effects in three populations and four locations (A,B,C,D)

<table>
<thead>
<tr>
<th>QTL</th>
<th>Population 1</th>
<th>Population 2</th>
<th>Population 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>1</td>
<td>3.7</td>
<td>-1.4</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>-2.2</td>
<td>-1.3</td>
<td>-3.2</td>
</tr>
<tr>
<td>3</td>
<td>6.7</td>
<td>5.2</td>
<td>1.4</td>
</tr>
<tr>
<td>4</td>
<td>-3.0</td>
<td>-0.9</td>
<td>4.1</td>
</tr>
<tr>
<td>5</td>
<td>1.5</td>
<td>-3.7</td>
<td>-0.5</td>
</tr>
<tr>
<td>6</td>
<td>0.7</td>
<td>-2.1</td>
<td>-2.7</td>
</tr>
<tr>
<td>7</td>
<td>-1.6</td>
<td>0.0</td>
<td>-0.7</td>
</tr>
<tr>
<td>8</td>
<td>-6.3</td>
<td>-1.1</td>
<td>-6.1</td>
</tr>
<tr>
<td>9</td>
<td>-7.6</td>
<td>-2.3</td>
<td>-0.3</td>
</tr>
<tr>
<td>10</td>
<td>17.0</td>
<td>3.9</td>
<td>7.2</td>
</tr>
<tr>
<td>11</td>
<td>-1.0</td>
<td>7.5</td>
<td>-1.3</td>
</tr>
<tr>
<td>12</td>
<td>0.7</td>
<td>-3.9</td>
<td>4.0</td>
</tr>
<tr>
<td>13</td>
<td>9.3</td>
<td>12.1</td>
<td>1.2</td>
</tr>
<tr>
<td>14</td>
<td>-9.3</td>
<td>-13.5</td>
<td>-6.7</td>
</tr>
<tr>
<td>15</td>
<td>4.5</td>
<td>4.2</td>
<td>5.5</td>
</tr>
<tr>
<td>16</td>
<td>-3.0</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>17</td>
<td>-16.4</td>
<td>9.9</td>
<td>-1.9</td>
</tr>
<tr>
<td>18</td>
<td>20.3</td>
<td>-9.3</td>
<td>-0.2</td>
</tr>
<tr>
<td>19</td>
<td>-4.4</td>
<td>-0.9</td>
<td>1.6</td>
</tr>
</tbody>
</table>
Phenotypic mean, variance ($V_p$), proportion of variance explained by the QTL model ($V_g/V_p$) for each location in each population

<table>
<thead>
<tr>
<th>Env</th>
<th>mean</th>
<th>$V_p$</th>
<th>$V_g/V_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>130.99</td>
<td>244.97</td>
<td>0.567</td>
</tr>
<tr>
<td>B</td>
<td>139.46</td>
<td>128.82</td>
<td>0.358</td>
</tr>
<tr>
<td>C</td>
<td>128.45</td>
<td>83.23</td>
<td>0.464</td>
</tr>
<tr>
<td>D</td>
<td>142.03</td>
<td>65.02</td>
<td>0.479</td>
</tr>
<tr>
<td>Population 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>124.54</td>
<td>295.34</td>
<td>0.555</td>
</tr>
<tr>
<td>B</td>
<td>130.49</td>
<td>152.52</td>
<td>0.407</td>
</tr>
<tr>
<td>C</td>
<td>125.57</td>
<td>86.40</td>
<td>0.283</td>
</tr>
<tr>
<td>D</td>
<td>139.05</td>
<td>97.86</td>
<td>0.448</td>
</tr>
<tr>
<td>Population 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>148.49</td>
<td>183.03</td>
<td>0.580</td>
</tr>
<tr>
<td>B</td>
<td>147.40</td>
<td>102.98</td>
<td>0.407</td>
</tr>
<tr>
<td>C</td>
<td>135.22</td>
<td>90.16</td>
<td>0.434</td>
</tr>
<tr>
<td>D</td>
<td>148.99</td>
<td>62.39</td>
<td>0.562</td>
</tr>
<tr>
<td>Env pair</td>
<td>$r_p$</td>
<td>$r_e$</td>
<td>$r_g$</td>
</tr>
<tr>
<td>----------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Population 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A,B</td>
<td>0.33</td>
<td>0.40</td>
<td>0.25</td>
</tr>
<tr>
<td>A,C</td>
<td>0.43</td>
<td>0.28</td>
<td>0.60</td>
</tr>
<tr>
<td>A,D</td>
<td>0.52</td>
<td>0.59</td>
<td>0.42</td>
</tr>
<tr>
<td>B,C</td>
<td>0.24</td>
<td>0.13</td>
<td>0.40</td>
</tr>
<tr>
<td>B,D</td>
<td>0.24</td>
<td>0.53</td>
<td>-0.16</td>
</tr>
<tr>
<td>C,D</td>
<td>0.37</td>
<td>0.56</td>
<td>0.12</td>
</tr>
<tr>
<td>Population 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A,B</td>
<td>0.43</td>
<td>0.26</td>
<td>0.56</td>
</tr>
<tr>
<td>A,C</td>
<td>0.46</td>
<td>0.42</td>
<td>0.66</td>
</tr>
<tr>
<td>A,D</td>
<td>0.46</td>
<td>0.26</td>
<td>0.56</td>
</tr>
<tr>
<td>B,C</td>
<td>0.34</td>
<td>0.34</td>
<td>0.33</td>
</tr>
<tr>
<td>B,D</td>
<td>0.44</td>
<td>0.38</td>
<td>0.51</td>
</tr>
<tr>
<td>C,D</td>
<td>0.46</td>
<td>0.54</td>
<td>0.29</td>
</tr>
<tr>
<td>Population 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A,B</td>
<td>0.25</td>
<td>0.13</td>
<td>0.33</td>
</tr>
<tr>
<td>A,C</td>
<td>0.08</td>
<td>0.49</td>
<td>-0.27</td>
</tr>
<tr>
<td>A,D</td>
<td>0.31</td>
<td>0.02</td>
<td>0.49</td>
</tr>
<tr>
<td>B,C</td>
<td>0.12</td>
<td>0.31</td>
<td>-0.20</td>
</tr>
<tr>
<td>B,D</td>
<td>0.29</td>
<td>0.07</td>
<td>0.55</td>
</tr>
<tr>
<td>C,D</td>
<td>0.13</td>
<td>0.38</td>
<td>-0.15</td>
</tr>
</tbody>
</table>
### Partition of genetic correlation between location A and C in population 3

<table>
<thead>
<tr>
<th>QTL:C\A</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-.053</td>
<td>.021</td>
<td>.026</td>
<td>-.019</td>
<td>.007</td>
<td>.000</td>
<td>.012</td>
<td>-.023</td>
<td>.024</td>
<td>.005</td>
<td>-.016</td>
<td>-.009</td>
<td>.004</td>
<td>.004</td>
<td>-.007</td>
<td>-.008</td>
<td>-.006</td>
<td>-.003</td>
<td>.006</td>
<td>-.036</td>
</tr>
<tr>
<td>2</td>
<td>.006</td>
<td>-.017</td>
<td>-.010</td>
<td>.008</td>
<td>-.003</td>
<td>.000</td>
<td>.006</td>
<td>-.005</td>
<td>-.002</td>
<td>.012</td>
<td>.003</td>
<td>-.002</td>
<td>-.002</td>
<td>.004</td>
<td>.002</td>
<td>.002</td>
<td>.001</td>
<td>-.002</td>
<td>-.003</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-.040</td>
<td>.056</td>
<td>.258</td>
<td>-.160</td>
<td>.011</td>
<td>.000</td>
<td>.036</td>
<td>-.097</td>
<td>.067</td>
<td>.026</td>
<td>-.099</td>
<td>-.033</td>
<td>.018</td>
<td>.015</td>
<td>-.024</td>
<td>-.020</td>
<td>-.025</td>
<td>-.010</td>
<td>.018</td>
<td>-.003</td>
</tr>
<tr>
<td>4</td>
<td>.025</td>
<td>-.036</td>
<td>-.139</td>
<td>.138</td>
<td>-.011</td>
<td>.000</td>
<td>-.028</td>
<td>.083</td>
<td>-.051</td>
<td>-.020</td>
<td>.065</td>
<td>.023</td>
<td>-.012</td>
<td>-.010</td>
<td>.023</td>
<td>.014</td>
<td>.018</td>
<td>.008</td>
<td>-.018</td>
<td>.072</td>
</tr>
<tr>
<td>5</td>
<td>.004</td>
<td>-.006</td>
<td>-.005</td>
<td>.005</td>
<td>-.018</td>
<td>.000</td>
<td>-.004</td>
<td>.011</td>
<td>-.008</td>
<td>.002</td>
<td>.008</td>
<td>.005</td>
<td>-.003</td>
<td>-.002</td>
<td>.003</td>
<td>.000</td>
<td>.001</td>
<td>-.001</td>
<td>-.012</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-.016</td>
<td>.017</td>
<td>.031</td>
<td>-.027</td>
<td>.016</td>
<td>-.001</td>
<td>.035</td>
<td>-.081</td>
<td>.050</td>
<td>.013</td>
<td>-.052</td>
<td>-.024</td>
<td>.011</td>
<td>.009</td>
<td>-.014</td>
<td>-.004</td>
<td>-.010</td>
<td>-.006</td>
<td>.028</td>
<td>-.026</td>
</tr>
<tr>
<td>7</td>
<td>.023</td>
<td>-.020</td>
<td>-.045</td>
<td>.041</td>
<td>-.013</td>
<td>.001</td>
<td>-.131</td>
<td>.177</td>
<td>-.100</td>
<td>-.019</td>
<td>.086</td>
<td>.026</td>
<td>-.009</td>
<td>-.008</td>
<td>.021</td>
<td>.015</td>
<td>.012</td>
<td>.004</td>
<td>-.007</td>
<td>.053</td>
</tr>
<tr>
<td>8</td>
<td>-.020</td>
<td>.020</td>
<td>.056</td>
<td>-.055</td>
<td>.016</td>
<td>-.001</td>
<td>.080</td>
<td>-.412</td>
<td>.186</td>
<td>.020</td>
<td>-.087</td>
<td>-.034</td>
<td>.013</td>
<td>.011</td>
<td>-.026</td>
<td>-.007</td>
<td>-.025</td>
<td>-.010</td>
<td>.015</td>
<td>-.259</td>
</tr>
<tr>
<td>9</td>
<td>.030</td>
<td>-.025</td>
<td>-.056</td>
<td>.049</td>
<td>-.017</td>
<td>.001</td>
<td>-.066</td>
<td>.274</td>
<td>-.243</td>
<td>-.019</td>
<td>.088</td>
<td>.034</td>
<td>-.013</td>
<td>-.010</td>
<td>.026</td>
<td>.021</td>
<td>.027</td>
<td>.011</td>
<td>-.012</td>
<td>.098</td>
</tr>
<tr>
<td>10</td>
<td>.014</td>
<td>-.022</td>
<td>-.046</td>
<td>.041</td>
<td>-.010</td>
<td>.000</td>
<td>-.026</td>
<td>.062</td>
<td>-.040</td>
<td>-.042</td>
<td>.049</td>
<td>.023</td>
<td>-.012</td>
<td>-.009</td>
<td>.021</td>
<td>.008</td>
<td>.014</td>
<td>.006</td>
<td>-.018</td>
<td>.015</td>
</tr>
<tr>
<td>11</td>
<td>-.015</td>
<td>.040</td>
<td>.062</td>
<td>-.047</td>
<td>.012</td>
<td>.000</td>
<td>.042</td>
<td>-.095</td>
<td>.066</td>
<td>.017</td>
<td>-.296</td>
<td>-.039</td>
<td>.021</td>
<td>.017</td>
<td>-.019</td>
<td>-.018</td>
<td>-.007</td>
<td>-.002</td>
<td>.015</td>
<td>-.243</td>
</tr>
<tr>
<td>12</td>
<td>-.008</td>
<td>.010</td>
<td>.020</td>
<td>-.016</td>
<td>.007</td>
<td>.000</td>
<td>.012</td>
<td>-.035</td>
<td>.024</td>
<td>.008</td>
<td>-.037</td>
<td>-.032</td>
<td>.006</td>
<td>.005</td>
<td>-.008</td>
<td>-.005</td>
<td>-.005</td>
<td>-.002</td>
<td>.007</td>
<td>-.050</td>
</tr>
<tr>
<td>13</td>
<td>-.011</td>
<td>.022</td>
<td>.037</td>
<td>-.028</td>
<td>.013</td>
<td>.000</td>
<td>.014</td>
<td>-.044</td>
<td>.031</td>
<td>.013</td>
<td>-.068</td>
<td>-.019</td>
<td>.049</td>
<td>.017</td>
<td>-.010</td>
<td>-.010</td>
<td>-.016</td>
<td>-.005</td>
<td>.006</td>
<td>-.009</td>
</tr>
<tr>
<td>14</td>
<td>.056</td>
<td>-.081</td>
<td>-.149</td>
<td>.108</td>
<td>-.039</td>
<td>.001</td>
<td>-.064</td>
<td>.189</td>
<td>-.119</td>
<td>-.053</td>
<td>.272</td>
<td>.088</td>
<td>-.085</td>
<td>-.075</td>
<td>.042</td>
<td>.036</td>
<td>.048</td>
<td>.019</td>
<td>-.065</td>
<td>.129</td>
</tr>
<tr>
<td>15</td>
<td>.003</td>
<td>-.006</td>
<td>-.007</td>
<td>.008</td>
<td>-.003</td>
<td>.000</td>
<td>-.005</td>
<td>.014</td>
<td>-.009</td>
<td>-.004</td>
<td>.009</td>
<td>.004</td>
<td>-.001</td>
<td>-.001</td>
<td>.023</td>
<td>.002</td>
<td>.002</td>
<td>.001</td>
<td>-.004</td>
<td>.025</td>
</tr>
<tr>
<td>16</td>
<td>.003</td>
<td>-.002</td>
<td>-.004</td>
<td>.003</td>
<td>.000</td>
<td>.000</td>
<td>-.002</td>
<td>.002</td>
<td>-.005</td>
<td>-.001</td>
<td>.006</td>
<td>.002</td>
<td>-.001</td>
<td>-.001</td>
<td>.001</td>
<td>.009</td>
<td>.003</td>
<td>.001</td>
<td>-.001</td>
<td>.012</td>
</tr>
<tr>
<td>17</td>
<td>-.013</td>
<td>.012</td>
<td>.035</td>
<td>-.028</td>
<td>.002</td>
<td>.000</td>
<td>.013</td>
<td>-.060</td>
<td>.045</td>
<td>.011</td>
<td>-.015</td>
<td>-.011</td>
<td>.011</td>
<td>.007</td>
<td>-.008</td>
<td>-.019</td>
<td>-.063</td>
<td>-.021</td>
<td>.010</td>
<td>-.094</td>
</tr>
<tr>
<td>18</td>
<td>.010</td>
<td>-.008</td>
<td>-.021</td>
<td>.018</td>
<td>.000</td>
<td>.000</td>
<td>-.006</td>
<td>.037</td>
<td>-.027</td>
<td>-.007</td>
<td>.006</td>
<td>.009</td>
<td>-.006</td>
<td>-.004</td>
<td>.005</td>
<td>.009</td>
<td>.031</td>
<td>.020</td>
<td>-.010</td>
<td>.057</td>
</tr>
<tr>
<td>19</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>-.001</td>
<td>.000</td>
<td>.000</td>
<td>-.001</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.004</td>
<td>.002</td>
</tr>
</tbody>
</table>

**Total genetic correlation:** -.273

**Due to pleiotropy (sum of diagonal elements):** -.883

**Due to linkage (sum of off-diagonal elements):** .609
Partition of genetic correlation between location B and D in population 3

<table>
<thead>
<tr>
<th>QTL:D\B</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.104</td>
<td>-.032</td>
<td>-.048</td>
<td>.019</td>
<td>.004</td>
<td>.005</td>
<td>-.003</td>
<td>.018</td>
<td>-.023</td>
<td>.005</td>
<td>-.004</td>
<td>.010</td>
<td>-.028</td>
<td>.004</td>
<td>.008</td>
<td>.008</td>
<td>-.015</td>
<td>-.003</td>
<td>.038</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-.055</td>
<td>.124</td>
<td>.090</td>
<td>-.037</td>
<td>-.009</td>
<td>-.007</td>
<td>.003</td>
<td>-.025</td>
<td>.025</td>
<td>-.010</td>
<td>.013</td>
<td>-.013</td>
<td>-.024</td>
<td>.054</td>
<td>-.010</td>
<td>-.007</td>
<td>-.010</td>
<td>.015</td>
<td>-.004</td>
<td>.121</td>
</tr>
<tr>
<td>3</td>
<td>-.081</td>
<td>.088</td>
<td>.484</td>
<td>-.164</td>
<td>-.007</td>
<td>-.015</td>
<td>.008</td>
<td>-.078</td>
<td>.067</td>
<td>-.024</td>
<td>.023</td>
<td>-.032</td>
<td>-.048</td>
<td>.116</td>
<td>-.015</td>
<td>-.019</td>
<td>-.032</td>
<td>.049</td>
<td>.008</td>
<td>.327</td>
</tr>
<tr>
<td>4</td>
<td>.048</td>
<td>-.055</td>
<td>-.246</td>
<td>.134</td>
<td>.006</td>
<td>.011</td>
<td>-.006</td>
<td>.063</td>
<td>-.048</td>
<td>.018</td>
<td>-.015</td>
<td>.021</td>
<td>.030</td>
<td>-.069</td>
<td>.013</td>
<td>.013</td>
<td>.022</td>
<td>-.035</td>
<td>-.007</td>
<td>-.102</td>
</tr>
<tr>
<td>5</td>
<td>.004</td>
<td>.004</td>
<td>-.002</td>
<td>-.005</td>
<td>-.001</td>
<td>.000</td>
<td>-.004</td>
<td>.004</td>
<td>-.001</td>
<td>.001</td>
<td>-.002</td>
<td>-.003</td>
<td>.005</td>
<td>-.001</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>-.005</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-.014</td>
<td>.012</td>
<td>.026</td>
<td>-.013</td>
<td>-.005</td>
<td>-.021</td>
<td>.003</td>
<td>-.030</td>
<td>.023</td>
<td>-.006</td>
<td>.006</td>
<td>-.011</td>
<td>-.013</td>
<td>.032</td>
<td>-.004</td>
<td>-.002</td>
<td>-.006</td>
<td>.013</td>
<td>.005</td>
<td>-.003</td>
</tr>
<tr>
<td>7</td>
<td>-.032</td>
<td>.022</td>
<td>.058</td>
<td>-.029</td>
<td>-.006</td>
<td>-.015</td>
<td>.019</td>
<td>-.097</td>
<td>.068</td>
<td>-.012</td>
<td>.014</td>
<td>-.017</td>
<td>-.016</td>
<td>.043</td>
<td>-.009</td>
<td>-.010</td>
<td>-.011</td>
<td>.012</td>
<td>.002</td>
<td>-.014</td>
</tr>
<tr>
<td>8</td>
<td>.072</td>
<td>-.057</td>
<td>-.185</td>
<td>.100</td>
<td>.018</td>
<td>.040</td>
<td>-.031</td>
<td>.585</td>
<td>-.329</td>
<td>.033</td>
<td>-.036</td>
<td>.057</td>
<td>.058</td>
<td>-.148</td>
<td>.028</td>
<td>.012</td>
<td>.057</td>
<td>-.086</td>
<td>-.011</td>
<td>.177</td>
</tr>
<tr>
<td>9</td>
<td>-.048</td>
<td>.031</td>
<td>.084</td>
<td>-.040</td>
<td>-.008</td>
<td>-.016</td>
<td>.011</td>
<td>-.174</td>
<td>.193</td>
<td>-.014</td>
<td>.017</td>
<td>-.026</td>
<td>-.027</td>
<td>.062</td>
<td>-.012</td>
<td>-.016</td>
<td>-.028</td>
<td>.041</td>
<td>.004</td>
<td>.033</td>
</tr>
<tr>
<td>10</td>
<td>.014</td>
<td>-.017</td>
<td>-.041</td>
<td>.020</td>
<td>.003</td>
<td>.005</td>
<td>-.003</td>
<td>.023</td>
<td>-.019</td>
<td>.018</td>
<td>-.005</td>
<td>.011</td>
<td>.014</td>
<td>-.034</td>
<td>.006</td>
<td>.004</td>
<td>.009</td>
<td>-.013</td>
<td>-.003</td>
<td>-.008</td>
</tr>
<tr>
<td>11</td>
<td>-.003</td>
<td>-.007</td>
<td>-.013</td>
<td>.005</td>
<td>.001</td>
<td>.002</td>
<td>-.001</td>
<td>.008</td>
<td>-.007</td>
<td>.002</td>
<td>-.008</td>
<td>.004</td>
<td>.006</td>
<td>-.015</td>
<td>.001</td>
<td>.002</td>
<td>.001</td>
<td>-.001</td>
<td>-.001</td>
<td>-.016</td>
</tr>
<tr>
<td>12</td>
<td>-.013</td>
<td>.012</td>
<td>.029</td>
<td>-.013</td>
<td>-.003</td>
<td>-.006</td>
<td>.002</td>
<td>-.022</td>
<td>.019</td>
<td>-.006</td>
<td>.007</td>
<td>-.024</td>
<td>-.012</td>
<td>.032</td>
<td>-.004</td>
<td>-.004</td>
<td>.005</td>
<td>.010</td>
<td>.002</td>
<td>.002</td>
</tr>
<tr>
<td>13</td>
<td>.020</td>
<td>-.029</td>
<td>-.059</td>
<td>.024</td>
<td>.007</td>
<td>.009</td>
<td>-.003</td>
<td>.030</td>
<td>-.026</td>
<td>.010</td>
<td>-.014</td>
<td>.015</td>
<td>.108</td>
<td>-.112</td>
<td>.005</td>
<td>.008</td>
<td>.017</td>
<td>-.023</td>
<td>-.002</td>
<td>-.013</td>
</tr>
<tr>
<td>14</td>
<td>-.024</td>
<td>.027</td>
<td>.059</td>
<td>-.024</td>
<td>-.005</td>
<td>-.009</td>
<td>.003</td>
<td>-.032</td>
<td>.025</td>
<td>-.010</td>
<td>.014</td>
<td>-.018</td>
<td>-.047</td>
<td>.123</td>
<td>-.005</td>
<td>-.008</td>
<td>.013</td>
<td>.020</td>
<td>.006</td>
<td>.081</td>
</tr>
<tr>
<td>15</td>
<td>-.012</td>
<td>.017</td>
<td>.026</td>
<td>-.015</td>
<td>-.003</td>
<td>-.004</td>
<td>.002</td>
<td>-.020</td>
<td>.017</td>
<td>-.006</td>
<td>.004</td>
<td>-.007</td>
<td>-.007</td>
<td>.018</td>
<td>-.025</td>
<td>-.003</td>
<td>-.004</td>
<td>.006</td>
<td>.003</td>
<td>-.013</td>
</tr>
<tr>
<td>16</td>
<td>.017</td>
<td>-.009</td>
<td>-.026</td>
<td>.012</td>
<td>.000</td>
<td>.001</td>
<td>-.002</td>
<td>.007</td>
<td>-.017</td>
<td>.003</td>
<td>-.005</td>
<td>.005</td>
<td>.009</td>
<td>-.019</td>
<td>.002</td>
<td>.029</td>
<td>.012</td>
<td>-.014</td>
<td>-.002</td>
<td>.004</td>
</tr>
<tr>
<td>17</td>
<td>-.025</td>
<td>.017</td>
<td>.059</td>
<td>-.027</td>
<td>-.001</td>
<td>-.006</td>
<td>.003</td>
<td>-.044</td>
<td>.041</td>
<td>-.009</td>
<td>.003</td>
<td>-.010</td>
<td>-.026</td>
<td>.048</td>
<td>-.004</td>
<td>-.017</td>
<td>-.076</td>
<td>.091</td>
<td>.004</td>
<td>.021</td>
</tr>
<tr>
<td>18</td>
<td>.033</td>
<td>-.019</td>
<td>-.062</td>
<td>.030</td>
<td>.000</td>
<td>.009</td>
<td>-.002</td>
<td>.046</td>
<td>-.042</td>
<td>.010</td>
<td>-.002</td>
<td>.013</td>
<td>.024</td>
<td>-.050</td>
<td>.005</td>
<td>.013</td>
<td>.064</td>
<td>-.153</td>
<td>-.007</td>
<td>-.089</td>
</tr>
<tr>
<td>19</td>
<td>-.003</td>
<td>.003</td>
<td>.006</td>
<td>-.004</td>
<td>-.000</td>
<td>-.002</td>
<td>.000</td>
<td>-.003</td>
<td>.002</td>
<td>-.002</td>
<td>.001</td>
<td>-.002</td>
<td>-.001</td>
<td>.009</td>
<td>-.001</td>
<td>-.001</td>
<td>-.002</td>
<td>.004</td>
<td>.008</td>
<td>.010</td>
</tr>
</tbody>
</table>

Total genetic correlation: .551
Due to pleiotropy (sum of diagonal elements): 1.617
Due to linkage (sum of off-diagonal elements): -1.066
Multiple Interval Mapping for eQTL Analysis

Zhao-Bang Zeng
Bioinformatics Research Center
Departments of Statistics & Genetics
North Carolina State University
zeng@stat.ncsu.edu
Goals and Issues of eQTL Mapping Analysis

- Identify and map genomic regions that significantly affect expression levels of different genes
  - Statistical methods and power to map eQTL
  - Justification of mapping procedures and results (e.g. FDR)
  - Epistasis of eQTL
  - Multiple trait analysis
- Through the mapping to identify cis- and trans-regulation of eQTL
- Identify gene expression co-regulation patterns (eQTL hot-spots)
  - Why are they co-regulated? Is there any functional relationship among those co-regulated genes?
- Prioritize candidate genes (from eQTL to genes)
  - By using regulative and functional relationship between candidate genes in eQTL regions and genes whose expressions being regulated, we might be able to prioritize and suggest candidate causal genes for some eQTL.
- Toward to network and pathway analysis

An eQTL study on a yeast hybrid segregant population

Experimental Design and Data

- **Sample**: BY (lab strain), RM (natural strain) and 112 F1 segregants.
- **Markers**: 3312 using yeast oligoarrays
- **Gene expression**: Samples were labeled and hybridized to cDNA microarrays, containing 6215 open reading frames (ORF).
- **Reference design**: Each two-color experiment involved one sample and one reference, with the same BY RNA reference being used for all experiments.
- **Dye swap**: Two hybridizations were carried out for each sample, one with the sample labeled with Cy3 and the reference with Cy5, and one with the fluors reversed; for each gene, the two log ratios were averaged.

Yeast experiment data structure

<table>
<thead>
<tr>
<th>Markers</th>
<th>Expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ind</td>
<td>1 2 3 4 5 6 7 8 9 10 11 ...... 3312</td>
</tr>
<tr>
<td>YU</td>
<td>1 1 1 1 1 1 1 1 1 1 ...... 1</td>
</tr>
<tr>
<td>RM</td>
<td>0 0 0 0 0 0 0 0 0 0 ...... 0</td>
</tr>
<tr>
<td>S1</td>
<td>1 1 1 1 0 0 0 0 0 1 ...... 0</td>
</tr>
<tr>
<td>S2</td>
<td>0 0 0 1 1 1 1 0 0 0 ...... 1</td>
</tr>
<tr>
<td>S3</td>
<td>0 0 0 0 0 0 1 1 1 1 ...... 0</td>
</tr>
<tr>
<td>....</td>
<td>.....</td>
</tr>
<tr>
<td>S112</td>
<td>1 1 1 0 0 0 0 0 1 1 ...... 0</td>
</tr>
</tbody>
</table>

Data: For $l = 1, 2, \ldots, 112+2$

$X_{ij}$ for $j = 1, 2, \ldots, 3312$  $Y_{ik}$ for $k = 1, 2, \ldots, 6215$
The Simplest Analysis – t-test

- For every $j$ (marker) and $k$ (expression trait):
  \[ y_{ik} = \alpha + \beta x_{ij} + e_i \quad \text{for} \quad i = 1, 2, \ldots, n \]
  
  test $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$

- Genome-wide search (multiple test issue): We want to test
  $H_0 : \text{there is no QTL in the genome}$
  $H_1 : \text{there is QTL in the genome}$

  Want to know the distribution and threshold of
  $\max(T(j), j \in \text{genome})$ at the null.

  Solution: permutation test

- Declare those $j$’s with $T(j) > \text{threshold linked to QTL}$

Interval Mapping – smooth the likelihood profile and define a confidence region

- Model:
  \[ y_{ik} = \alpha + \beta x_{il} + e_i \]

  $l$: a location in the genome
  $x_{il}$: missing data, but with known $\Pr(x_{il} = 1$ or 0|markers)

- Analyze the mixture model and calculate
  \[ LOD(l) = \log_{10} \frac{L(\beta \neq 0)_l}{L(\beta = 0)_l} \]

- The genome-wide threshold for $LOD(l)$ can be determined by a permutation test

- 1.5-LOD support interval: The genome region covered by dropping LOD by 1.5 from the peak gives a confidence interval of QTL position
Search for multiple eQTL and epistasis

- Exhaustive 2D (or multi-D) search for 2 or more eQTL
  - There is empirical evidence that sequential search is more powerful than exhaustive 2D search (Storey et al 2005)
- Sequential search (Storey et al version)
- Sequential search (Zou and Zeng version)

Sequential search for eQTL pairs and epistasis
(Storey et al 2005 PLoS Biology)

Storey et al (2005) proposed a two-stage sequential search to detect interactive eQTL:

- For each eTrait, perform one-dimensional genome scan to identify the best position for a putative eQTL.
- Conditional on the 1st putative eQTL position, perform another one-dimensional genome scan to detect the 2nd putative eQTL with a model of both main and interaction effects.
- Final selection on the (interactive) eQTL is based on FDR.
  - “A significant expression trait is called a false discovery if any of the loci selected for that trait is a false positive. That is, a true discovery is an expression trait where all selected loci are truly linked.”
Storey et al (2005)

• “Based on the joint linkage probabilities, we estimate that 2,300 traits (approximately 37%) are jointly linked to two loci, although we cannot identify all of these with high confidence. Of these 2,300 traits, 170 can be identified at a FDR of 10%.”

• “In total, 58 traits demonstrate a \textit{cis} linkage”

Multiple interval mapping for eQTL analysis (Zou and Zeng 2006)

• Model:

$$y_{ik} = \alpha + \sum_t \beta_t x_{it} + \sum_{s < t} \gamma_{st} x_{it} x_{it} + e_{ik}$$

• Sequential search for each eQTL conditional on the significance in the previous cycle for each eTrait.

• For each eTrait:
  - In cycle 1, if the max test statistic > threshold, the first eQTL is identified and continue the next step; otherwise stop the search.
  - In cycle t+1, if the conditional max test statistic > threshold, one more eQTL is added and continue the search; otherwise stop.
  - After the search for the main effects, epistatic effects of eQTL are tested based on the threshold and then added to the model.
  - Obtain 1.5-LOD support interval for each identified eQTL
MIM for eQTL analysis

- Threshold is first determined by a permutation test with a controlled type I error rate for the genome scan (e.g. 95 percentile of test statistic in a genome scan under the null).
- Then the threshold is evaluated or adjusted based on the calculation of False Discovery Rate (FDR) in the sequential genome scans for the whole detected eQTL for all the expression traits.

<table>
<thead>
<tr>
<th>Cycle</th>
<th>#Scanned$^1$</th>
<th>#Retained$^2$</th>
<th>#Claimed$^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6195</td>
<td>3367</td>
<td>3354</td>
</tr>
<tr>
<td>2</td>
<td>3367</td>
<td>1617</td>
<td>1242</td>
</tr>
<tr>
<td>3</td>
<td>1617</td>
<td>578</td>
<td>422</td>
</tr>
<tr>
<td>4</td>
<td>578</td>
<td>197</td>
<td>122</td>
</tr>
<tr>
<td>5</td>
<td>197</td>
<td>66</td>
<td>37</td>
</tr>
<tr>
<td>6</td>
<td>66</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

1. # of eTraits in each cycle
2. # of eTraits in the initial genome scans using the 10% genome-wide type I error rate
3. # of eTraits in the final result using the 5% genome-wide type I error rate

With the 5% genome-wide type I error in each genome scan, the False Discovery Rate (FDR) for all the detected eQTL is estimated about 8%.
Two procedure differences between Storey et al and Zou-Zeng

- Storey et al (2005) selected the 2nd eQTL for each trait from the genome based on both main and epistatic effects. This test has two degree of freedom. [1] Both main and epistatic effects
- They attempted to find the 2nd eQTL for each trait no matter how significant the 1st eQTL is. [2] Unrestricted search

Number of two QTL genetic models declared with 10% FDR

<table>
<thead>
<tr>
<th>Methods</th>
<th>Number of eTraits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main effect only: relax [1]</td>
<td>746</td>
</tr>
</tbody>
</table>
A simulation study

- Simulate 620 (10%) traits from the pool of inferred models with two eQTL among 6215 traits.
- To study the effect of epistasis on searching strategies, we simulate 5%, 50% and 95% proportions of genetic models with epistasis.

Simulation result:
Mean number (SD) of 2 QTL model with 10% FDR

<table>
<thead>
<tr>
<th>Interaction proportion</th>
<th>Storey et al method</th>
<th>Main effect only</th>
<th>Restricted search</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>99(37)</td>
<td>238(71)</td>
<td>168(72)</td>
</tr>
<tr>
<td>50%</td>
<td>161(40)</td>
<td>194(39)</td>
<td>236(67)</td>
</tr>
<tr>
<td>95%</td>
<td>226(62)</td>
<td>190(58)</td>
<td>306(64)</td>
</tr>
</tbody>
</table>
The role of threshold in MIM-eQTL

- In the later cycles of genome scans, the search is restricted within the parameter space where the chance to detect strong association is high as we focus on those traits that have shown significant QTL in the previous cycles;
- It serves as a stopping rule to decide how many QTL we can find for each trait.

### Table 4.1: A single gene as the most probably gene for large number of traits

<table>
<thead>
<tr>
<th>gene</th>
<th>max</th>
<th>all</th>
<th>annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>YNL067W</td>
<td>92</td>
<td>294</td>
<td>Protein component of the large (60S) ribosomal subunit</td>
</tr>
<tr>
<td>YNL069C</td>
<td>116</td>
<td>367</td>
<td>N-terminally acetylated protein component of the large (60S) ribosomal subunit</td>
</tr>
<tr>
<td>YBR154C</td>
<td>146</td>
<td>436</td>
<td>RNA polymerase subunit ABC27, common to RNA polymerases I, II, and III; contacts DNA and affects transactivation</td>
</tr>
<tr>
<td>YNL096C</td>
<td>224</td>
<td>325</td>
<td>Protein component of the small (40S) ribosomal subunit</td>
</tr>
<tr>
<td>YOL077C</td>
<td>284</td>
<td>452</td>
<td>Nuclear protein, constituent of 60S pre-ribosomal particles</td>
</tr>
</tbody>
</table>

1. # of overlapping eQTL that the gene is identified as the most probable underlying gene by our Bayesian gene prioritization analysis.
2. Total # of overlapping eQTL that include the gene as a candidate gene.
3. From http://www.yeastgenome.org/

### Acknowledgement

- **NC State University Bioinformatics Research Center**
  - Chris Basten
  - Wei Zou
  - Jessica Maia

- **Funding**
  - NIH GM45344
  - USDA Plant Genome

- **NC State University Forest Biotechnology Group** (for *Eucalyptus* data)
  - Ronald R. Sederoff
  - Matias Kirst

- **U. Washington** (for yeast data)
  - Leonid Kruglyak
Identify network structure from eQTL hotspots
Duarte and Zeng (2009)

• A method for finding a “framework network” that decomposes the network into a series of hierarchical layers through a series of statistical tests of conditional independence or dependence.

• Model:

\[ P(G|T) = \prod_{S \in S} P(S|T) \prod_{G \in G \setminus S} P(G|S) \]

• Example:

Illustration of algorithm steps
An example of discovered network

<table>
<thead>
<tr>
<th>shielded</th>
<th>shacked</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFR1</td>
<td>HHF2 TOS4 YDR306W YER156C</td>
</tr>
<tr>
<td>AGA1</td>
<td>FAR1 PCL2 PRM6</td>
</tr>
<tr>
<td>FIG2</td>
<td>FIG1 PRM1 PRM2</td>
</tr>
<tr>
<td>FUS1</td>
<td>ASG7 BUD14 CHS1 EST4 FUS3 GPA1 HAL1 HYM1 ISP2 KAR5 KNI1 MCM4 NIS1 NTa1 PDS1 PRM1 PRM5 RGD2 SHU1 SMY1 SNL1 SPP1 STH2 STE14 TEC1 UME6 YBP2 YDR124W YDR249C YDR292C YER195C YIL080W YJR028W YJR054W YOR343C</td>
</tr>
<tr>
<td>YHR127W</td>
<td>KAR1 SRP102 YIL158W</td>
</tr>
</tbody>
</table>

Table 4.3: Discovered Network 6: eQTL at Chromosome 8 at 98,513 bp.

- Strong evidence that the eQTL is a deletion in GPA1 Gene.
- GPA1 is the alpha subunit of the G protein coupled to mating factor receptor and is involved in the mating pheromone signal transduction pathway.
- FUS1, AFR1, AGA1, and FIG2 are all associated with the GO biological process "response to pheromone during conjugation with cellular fusion".
- STE12 is a transcription factor that is activated by a MAPK kinase signaling cascade involved in pheromone.
- FUS1 is the first downstream target of STE12.
- 61.1% of genes shielded by FUS1 are targets of the STE12 transcription factor.
- Itself a target of STE12, it may modulate the transcriptional effects of STE12 on other targets.

A model for inferring gene pathway from gene knockout experiment (Aylor and Zeng, 2007)

Classical genetic model and interpretation:

\[ A^+B^+ : y = \mu + \varepsilon \]
\[ A^-B^+ : y = \mu + \beta_A + \varepsilon \]
\[ A^+B^- : y = \mu + \beta_B + \varepsilon \]
\[ A^-B^- : y = \begin{cases} \mu + \beta_A + \varepsilon & \text{if } A \text{ is epistatic to } B \\ \mu + \beta_B + \varepsilon & \text{if } B \text{ is epistatic to } A \end{cases} \]

Mostly applied for lethal phenotype or viability
A quantitative genetic model

\[ y = \mu + \beta_A x_A + \beta_B x_B + \beta_I x_A x_B + \varepsilon \]

Model 1: \( y = \mu + \beta_A + \varepsilon \)
Model 2: \( y = \mu + \beta_B + \varepsilon \)
Model 3: \( y = \mu + \beta_I + \varepsilon \)
Model 4: \( y = \mu + \beta_A + \beta_B + \varepsilon \)
Model 5: \( y = \mu + \beta_A + \beta_I + \varepsilon \)
Model 6: \( y = \mu + \beta_B + \beta_I + \varepsilon \)
Model 7: \( y = \mu + \beta_A + \beta_B + \beta_I + \varepsilon \)
Model 8: \( y = \mu + \varepsilon \)

Gene pathway interpretation

• We considered all combinations of gene order and action within simple ON/OFF models and then predicted the hypothetical effect of deleting genes on each of them.
• There are four points of variation to model for each gene pair relationship.
  – The first is the identity of the upstream gene, i.e. the gene order.
  – Secondly, the upstream gene will turn the downstream gene either on (enhance) or off (repress).
  – Thirdly, the downstream gene can enhance or repress the expression of a target gene for which expression is observed.
  – Lastly, we consider that the upstream gene itself will be enhanced or repressed by some initiating factor such as a developmental cue or environmental perturbation.
Linking quantitative model to pathway interpretation: An example

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Upstream Gene</th>
<th>Gene Action</th>
<th>Target Gene Expression</th>
<th>Regression Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>A(^+) B(^+)</td>
<td>ON</td>
<td>( \alpha \rightarrow B \rightarrow 1 ) off</td>
<td>( \mu + \beta_4 + \beta_5 )</td>
<td></td>
</tr>
<tr>
<td>A(^+) B(^-)</td>
<td>ON</td>
<td>( \alpha \rightarrow X \rightarrow 1 ) off</td>
<td>( \mu + \beta_6 )</td>
<td></td>
</tr>
<tr>
<td>A(^-) B(^+)</td>
<td>ON</td>
<td>( \alpha \rightarrow X \rightarrow 1 ) off</td>
<td>( \mu )</td>
<td></td>
</tr>
<tr>
<td>A(^-) B(^-)</td>
<td>ON</td>
<td>( \alpha \rightarrow X \rightarrow 1 ) off</td>
<td>( \mu )</td>
<td></td>
</tr>
</tbody>
</table>

Another example

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Upstream Gene</th>
<th>Gene Action</th>
<th>Target Gene Expression</th>
<th>Regression Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>A(^+) B(^+)</td>
<td>ON</td>
<td>( \alpha \rightarrow B \rightarrow 2 ) on</td>
<td>( \mu + \beta_7 )</td>
<td></td>
</tr>
<tr>
<td>A(^+) B(^-)</td>
<td>ON</td>
<td>( \alpha \rightarrow X \rightarrow 1 ) on</td>
<td>( \mu + \beta_6 + \beta_8 + \beta_9 )</td>
<td></td>
</tr>
<tr>
<td>A(^-) B(^+)</td>
<td>ON</td>
<td>( \alpha \rightarrow X \rightarrow 1 ) on</td>
<td>( \mu + \beta_6 )</td>
<td></td>
</tr>
<tr>
<td>A(^-) B(^-)</td>
<td>ON</td>
<td>( \alpha \rightarrow X \rightarrow 1 ) on</td>
<td>( \mu )</td>
<td></td>
</tr>
<tr>
<td>A(^+) B(^+)</td>
<td>OFF</td>
<td>( \alpha \rightarrow B \rightarrow 1 ) off</td>
<td>( \mu )</td>
<td></td>
</tr>
<tr>
<td>A(^+) B(^-)</td>
<td>OFF</td>
<td>( \alpha \rightarrow X \rightarrow 1 ) off</td>
<td>( \mu )</td>
<td></td>
</tr>
<tr>
<td>A(^-) B(^+)</td>
<td>OFF</td>
<td>( \alpha \rightarrow X \rightarrow 1 ) off</td>
<td>( \mu )</td>
<td></td>
</tr>
<tr>
<td>A(^-) B(^-)</td>
<td>OFF</td>
<td>( \alpha \rightarrow X \rightarrow 1 ) off</td>
<td>( \mu )</td>
<td></td>
</tr>
</tbody>
</table>
Match quantitative models with pathway interpretations

a. Hierarchical Relationships

<table>
<thead>
<tr>
<th>Upstream Gene</th>
<th>A upstream of B</th>
<th>B upstream of A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ON</td>
<td>OFF</td>
</tr>
<tr>
<td>Repressor</td>
<td>$\mu + \beta_4 + \beta_1$ [5]</td>
<td>$\mu + \beta_8$ [2]</td>
</tr>
<tr>
<td>Enhancer</td>
<td>$\mu + \beta_4 + \beta_8 + \beta_1$ [7]</td>
<td>$\mu$ [8]</td>
</tr>
</tbody>
</table>

b. Non-hierarchical Relationships

<table>
<thead>
<tr>
<th>State of A</th>
<th>ON/ON</th>
<th>ON/OFF</th>
<th>OFF/ON</th>
<th>OFF/OFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enhancer/Enhancer</td>
<td>$\mu - \beta_1$ [3]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enhancer/Repressor Or Repressor/Enhancer</td>
<td>$\mu - \beta_4 - \beta_8$ [4]</td>
<td>$\mu - \beta_4$ [1]</td>
<td>$\mu - \beta_8$ [2]</td>
<td>$\mu$ [8]</td>
</tr>
<tr>
<td>Repressor/Repressor</td>
<td>$\mu - \beta_1$ [3]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

An application: Study of gene pathway in *Dictyostelium* (Van Driessche et al. 2007)

- Upon removal of nutrients, *D. discoideum* executes a developmental program in which single cells aggregate and form multicellular organisms.
- PKA pathway is known and important for the process. The pathway gene single and double knockout strains were created.
- Whole genome gene expression profiles were assayed and used to infer the pathway.
Concluding remarks

- The genetics of complex traits is very complex, and has been a black box.
- Through QTL mapping, a number of genetic components can be identified.
- With the aid of other omics information in a systems oriented study, some genetic pathways and networks that contribute to complex trait variation can also be illuminated.
Overview of Multiple QTL

1. what is the goal of multiple QTL study?
2. gene action and epistasis
3. Bayesian vs. classical QTL
4. QTL model selection
5. QTL software options
1. what is the goal of QTL study?

- uncover underlying biochemistry
  - identify how networks function, break down
  - find useful candidates for (medical) intervention
  - epistasis may play key role
  - statistical goal: maximize number of correctly identified QTL
- basic science/evolution
  - how is the genome organized?
  - identify units of natural selection
  - additive effects may be most important (Wright/Fisher debate)
  - statistical goal: maximize number of correctly identified QTL
- select “elite” individuals
  - predict phenotype (breeding value) using suite of characteristics
    (phenotypes) translated into a few QTL
  - statistical goal: minimize prediction error

cross two inbred lines
→ linkage disequilibrium
→ associations
→ linked segregating QTL
(after Gary Churchill)
problems of single QTL approach

- wrong model: biased view
  - fool yourself: bad guess at locations, effects
  - detect ghost QTL between linked loci
  - miss epistasis completely
- low power
- bad science
  - use best tools for the job
  - maximize scarce research resources
  - leverage already big investment in experiment

advantages of multiple QTL approach

- improve statistical power, precision
  - increase number of QTL detected
  - better estimates of loci: less bias, smaller intervals
- improve inference of complex genetic architecture
  - patterns and individual elements of epistasis
  - appropriate estimates of means, variances, covariances
    - asymptotically unbiased, efficient
  - assess relative contributions of different QTL
- improve estimates of genotypic values
  - less bias (more accurate) and smaller variance (more precise)
  - mean squared error = \( \text{MSE} = (\text{bias})^2 + \text{variance} \)
2. Gene Action and Epistasis

additive, dominant, recessive, general effects of a single QTL (Gary Churchill)
additive effects of two QTL (Gary Churchill)

\[ \mu_q = \mu + \beta_{q1} + \beta_{q2} \]

Epistasis (Gary Churchill)

The allelic state at one locus can mask or uncover the effects of allelic variation at another.

- W. Bateson, 1907.
epistasis in parallel pathways (GAC)

- Z keeps trait value low
- neither $E_1$ nor $E_2$ is rate limiting
- loss of function alleles are segregating from parent A at $E_1$ and from parent B at $E_2$

epistasis in a serial pathway (GAC)

- Z keeps trait value high
- either $E_1$ or $E_2$ is rate limiting
- loss of function alleles are segregating from parent B at $E_1$ or from parent A at $E_2$
epistatic interactions

• model space issues
  – 2-QTL interactions only?
    • or general interactions among multiple QTL?
  – partition of effects
    • Fisher-Cockerham or tree-structured or ?

• model search issues
  – epistasis between significant QTL
    • check all possible pairs when QTL included?
    • allow higher order epistasis?
  – epistasis with non-significant QTL
    • whole genome paired with each significant QTL?
    • pairs of non-significant QTL?

• see papers of Nengjun Yi (2000-7) in Genetics

limits of epistatic inference

• power to detect effects
  – epistatic model sizes grow quickly
    • $|A| = 3^n_{qtl}$ for general interactions
  – power tradeoff
    • depends sample size vs. model size
    • want $n / |A|$ to be fairly large (say > 5)
    • 3 QTL, $n = 100$ F2: $n / |A| \approx 4$ with $n = 100$

• rare genotypes may not be observed
  – $aa/BB$ & $AA/bb$ rare for linked loci
  – empty cells mess up balance
    • adjusted tests (type III) are wrong
  – confounds main effects & interactions

<table>
<thead>
<tr>
<th></th>
<th>$aa$</th>
<th>$aA$</th>
<th>$AA$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$bb$</td>
<td>6</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>$bB$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$BB$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2 linked QTL
empty cell
with $n = 100$
limits of multiple QTL?

• limits of statistical inference
  – power depends on sample size, heritability, environmental variation
  – “best” model balances fit to data and complexity (model size)
  – genetic linkage = correlated estimates of gene effects

• limits of biological utility
  – sampling: only see some patterns with many QTL
  – marker assisted selection (Bernardo 2001 Crop Sci)
    • 10 QTL ok, 50 QTL are too many
    • phenotype better predictor than genotype when too many QTL
    • increasing sample size may not give multiple QTL any advantage
  – hard to select many QTL simultaneously
    • 3^m possible genotypes to choose from

QTL below detection level?

• problem of selection bias
  – QTL of modest effect only detected sometimes
  – effects overestimated when detected
  – repeat studies may fail to detect these QTL

• think of probability of detecting QTL
  – avoids sharp in/out dichotomy
  – avoid pitfalls of one “best” model
  – examine “better” models with more probable QTL

• rethink formal approach for QTL
  – directly allow uncertainty in genetic architecture
  – QTL model selection over genetic architecture
3. Bayesian vs. classical QTL study

- classical study
  - maximize over unknown effects
  - test for detection of QTL at loci
  - model selection in stepwise fashion
- Bayesian study
  - average over unknown effects
  - estimate chance of detecting QTL
  - sample all possible models
- both approaches
  - average over missing QTL genotypes
  - scan over possible loci

Bayesian idea

- Reverend Thomas Bayes (1702-1761)
  - part-time mathematician
  - buried in Bunhill Cemetery, Moongate, London
  - famous paper in 1763 *Phil Trans Roy Soc London*
  - was Bayes the first with this idea? (Laplace?)
- basic idea (from Bayes’ original example)
  - two billiard balls tossed at random (uniform) on table
  - where is first ball if the second is to its left?
    - prior: anywhere on the table
    - posterior: more likely toward right end of table
QTL model selection: key players

- observed measurements
  - \( y \) = phenotypic trait
  - \( m \) = markers & linkage map
  - \( i = \) individual index (1,…,n)
- missing data
  - missing marker data
  - \( q \) = QT genotypes
    - alleles QQ, Qq, or qq at locus
- unknown quantities
  - \( \lambda \) = QT locus (or loci)
  - \( \mu \) = phenotype model parameters
  - \( \gamma \) = QTL model/genetic architecture

- \( \text{pr}(q|m,\lambda,\gamma) \) genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for \( q \) given \( m \)
- \( \text{pr}(y|q,\mu,\gamma) \) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters \( \mu \) (could be non-parametric)

Bayes posterior vs. maximum likelihood

- LOD: classical Log ODds
  - maximize likelihood over effects \( \mu \)
  - R/qtl scanone/scantwo: method = “em”
- LPD: Bayesian Log Posterior Density
  - average posterior over effects \( \mu \)
  - R/qtl scanone/scantwo: method = “imp”

\[
\text{LOD}(\lambda) = \log_{10}\{\max_{\mu} \text{pr}(y|m,\mu,\lambda)\} + c
\]

\[
\text{LPD}(\lambda) = \log_{10}\{\int \text{pr}(y|m,\mu,\lambda)\text{pr}(\mu)d\mu\} + C
\]

likelihood mixes over missing QTL genotypes:

\[
\text{pr}(y|m,\mu,\lambda) = \sum_{q} \text{pr}(y|q,\mu)\text{pr}(q|m,\lambda)
\]
LOD & LPD: 1 QTL
n.ind = 100, 1 cM marker spacing

LOD & LPD: 1 QTL
n.ind = 100, 10 cM marker spacing
marginal LOD or LPD

• compare two genetic architectures ($\gamma_2, \gamma_1$) at each locus
  – with ($\gamma_2$) or without ($\gamma_1$) another QTL at locus $\lambda$
    • preserve model hierarchy (e.g. drop any epistasis with QTL at $\lambda$)
  – with ($\gamma_2$) or without ($\gamma_1$) epistasis with QTL at locus $\lambda$
  – $\gamma_2$ contains $\gamma_1$ as a sub-architecture
• allow for multiple QTL besides locus being scanned
  – architectures $\gamma_1$ and $\gamma_2$ may have QTL at several other loci
  – use marginal LOD, LPD or other diagnostic
  – posterior, Bayes factor, heritability

$$LOD(\lambda \mid \gamma_2) = LOD(\lambda \mid \gamma_1)$$
$$LPD(\lambda \mid \gamma_2) = LPD(\lambda \mid \gamma_1)$$

LPD: 1 QTL vs. multi-QTL
marginal contribution to LPD from QTL at $\lambda$

![Diagram showing marginal contribution to LPD from QTL at $\lambda$.]
substitution effect: 1 QTL vs. multi-QTL
single QTL effect vs. marginal effect from QTL at $\lambda$

why use a Bayesian approach?

- first, do both classical and Bayesian
  - always nice to have a separate validation
  - each approach has its strengths and weaknesses
- classical approach works quite well
  - selects large effect QTL easily
  - directly builds on regression ideas for model selection
- Bayesian approach is comprehensive
  - samples most probable genetic architectures
  - formalizes model selection within one framework
  - readily (!) extends to more complicated problems
4. QTL model selection

- select class of models
  - see earlier slides above
- decide how to compare models
  - (Bayesian interval mapping talk later)
- search model space
  - (Bayesian interval mapping talk later)
- assess performance of procedure
  - see Kao (2000), Broman and Speed (2002)
  - Manichaukul, Moon, Yandell, Broman (in prep)
  - be wary of HK regression assessments

pragmatics of multiple QTL

- evaluate some objective for model given data
  - classical likelihood
  - Bayesian posterior
- search over possible genetic architectures (models)
  - number and positions of loci
  - gene action: additive, dominance, epistasis
- estimate “features” of model
  - means, variances & covariances, confidence regions
  - marginal or conditional distributions
- art of model selection
  - how select “best” or “better” model(s)?
  - how to search over useful subset of possible models?
comparing models

• balance model fit against model complexity
  – want to fit data well (maximum likelihood)
  – without getting too complicated a model

<table>
<thead>
<tr>
<th>fit model</th>
<th>smaller model</th>
<th>bigger model</th>
</tr>
</thead>
<tbody>
<tr>
<td>estimate phenotype</td>
<td>miss key features</td>
<td>fits better</td>
</tr>
<tr>
<td>predict new data</td>
<td>may be biased</td>
<td>no bias</td>
</tr>
<tr>
<td>interpret model</td>
<td>may be biased</td>
<td>no bias</td>
</tr>
<tr>
<td>estimate effects</td>
<td>easier</td>
<td>more complicated</td>
</tr>
<tr>
<td></td>
<td>low variance</td>
<td>high variance</td>
</tr>
</tbody>
</table>

Bayesian model averaging

• average summaries over multiple architectures
• avoid selection of “best” model
• focus on “better” models
• examples in data talk later
5. QTL software options

- methods
  - approximate QTL by markers
  - exact multiple QTL interval mapping

- software platforms
  - MapMaker/QTL (obsolete)
  - QTLCart (statgen.ncsu.edu/qtlcart)
  - R/qtl (www.rqtl.org)
  - R/qtlbim (www.qtlbim.org)
approximate QTL methods

• marker regression
  – locus & effect confounded
  – lose power with missing data
• Haley-Knott (least squares) regression
  – correct mean, wrong variance
  – biased by pattern of missing data (Kao 2000)
• extended HK regression
  – correct mean and variance
  – minimizes bias issue (R/qtl “ehk” method)
• composite interval mapping (QTLCart)
  – use markers to approximate other QTL
  – properties depend on marker spacing, missing data

exact QTL methods

• interval mapping (Lander, Botstein 1989)
  – scan whole genome for single QTL
  – bias for linked QTL, low power
• multiple interval mapping (Kao, Zeng, Teasdale 1999)
  – sequential scan of all QTL
  – stepwise model selection
• multiple imputation (Sen, Churchill 2001)
  – fill in (impute) missing genotypes along genome
  – average over multiple imputations
• Bayesian interval mapping (Yi et al. 2005)
  – sample most likely models
  – marginal scans conditional on other QTL
QLT software platforms

- QTLCart (statgen.ncsu.edu/qtlcart)
  - includes features of original MapMaker/QTL
    - not designed for building a linkage map
  - easy to use Windows version WinQTLCart
  - based on Lander-Botstein maximum likelihood LOD
    - extended to marker cofactors (CIM) and multiple QTL (MIM)
    - epistasis, some covariates (GxE)
    - stepwise model selection using information criteria
  - some multiple trait options
  - OK graphics
- R/qtl (www.rqtl.org)
  - includes functionality of classical interval mapping
  - many useful tools to check genotype data, build linkage maps
  - excellent graphics
  - several methods for 1-QTL and 2-QTL mapping
    - epistasis, covariates (GxE)
  - tools available for multiple QTL model selection

Bayesian QTL software options

- Bayesian Haley-Knott approximation: no epistasis
  - Berry C (1998)
    - R/bql (www.r-project.org contributed package)
- multiple imputation: epistasis, mostly 1-2 QTL but some multi-QTL
  - Sen and Churchill (2000)
    - matlab/pseudomarker (www.jax.org/staff/churchill/labsite/software)
  - Broman et al. (2003)
    - R/qtl (www.rqtl.org)
- Bayesian interval mapping via MCMC: no epistasis
    - R/bim (www.r-project.org contributed package)
    - WinQTLCart/bmapqtl (statgen.ncsu.edu/qtlcart)
  - Stephens & Fisch (1998): no code release
  - Sillanpää Arjas (1998)
    - multimapper (www.mi.helsinki.fi/~mjs)
- Bayesian interval mapping via MCMC: epistasis
  - Yandell et al. (2007)
    - R/qtlbim (www.qtlbim.org)
- Bayesian shrinkage: no epistasis
  - Wang et al. Xu (2005): no code release
R/qtlbim: www.qtlbim.org

• Properties
  – cross-compatible with R/qtl
  – new MCMC algorithms
    • Gibbs with loci indicators; no reversible jump
  – epistasis, fixed & random covariates, GxE
  – extensive graphics

• Software history
  – initially designed (Satagopan Yandell 1996)
  – major revision and extension (Gaffney 2001)
  – R/bim to CRAN (Wu, Gaffney, Jin, Yandell 2003)
  – R/qtlbim to CRAN (Yi, Yandell et al. 2006)

• Publications
  – Yi et al. (2005); Yandell et al. (2007); ...

many thanks

U AL Birmingham
  Nengjun Yi
  Tapan Mehta
  Samprit Banerjee
  Daniel Shriner
  Ram Venkataraman
  David Allison

Jackson Labs
  Gary Churchill
  Hao Wu
  Hyuna Yang
  Randy von Smith

UW-Madison Stats
  Alan Attie
  Jonathan Stoehr
  Hong Lan
  Susie Clee
  Jessica Byers
  Mark Gray-Keller
  Tom Osborn
  David Butruille
  Marcio Ferrera
  Josh Udahl
  Pablo Quijada
  Yandell lab
  Jaya Satagopan
  Fei Zou
  Patrick Gaffney
  Chunfang Jin
  Elias Chaibub
  W Whipple Neely
  Jee Young Moon
  Elias Chaibub
  Michael Newton
  Karl Broman
  Christina Kendzierski
  Daniel Gianola
  Liang Li
  Daniel Sorensen

USDA Hatch, NIH/NIDDK (Attie), NIH/R01s (Yi, Broman)
R/qtl & R/qtlbim Tutorials

- R statistical graphics & language system
- R/qtl tutorial
  - R/qtl web site: www.rqtl.org
  - Tutorial: www.rqtl.org/tutorials/rqtltour.pdf
  - R code: www.stat.wisc.edu/~yandell/qtlbim/rqtltour.R
  - url.show("http://www.stat.wisc.edu/~yandell/qtlbim/rqtltour.R")
- R/qtlbim tutorial
  - R/qtlbim web site: www.qtlbim.org
  - Tutorial and R code:
    - www.stat.wisc.edu/~yandell/qtlbim/rqtlbimtour.pdf
    - www.stat.wisc.edu/~yandell/qtlbim/qtlbimtour.R

R/qtl tutorial (www.rqtl.org)

```r
> library(qtl)
> data(hyper)
> summary(hyper)

Backcross

  No. individuals: 250
  No. phenotypes: 2
  Percent phenotyped: 100 100

  No. chromosomes: 20
    Autosomes: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
    X chr: X

  Total markers: 174
  No. markers: 22 8 6 20 14 11 7 6 5 5 14 5 5 11 6 12 4 4 4
  Percent genotyped: 47.7
  Genotypes (%): AA:50.2 AB:49.8

> plot(hyper)
> plot.missing(hyper, reorder = TRUE)
```
R/qtl: find genotyping errors

> hyper <- calc.errorlod(hyper, error.prob=0.01)
> top.errorlod(hyper)

<table>
<thead>
<tr>
<th>chr</th>
<th>id</th>
<th>marker</th>
<th>errorlod</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>118</td>
<td>D1Mit14</td>
<td>8.372794</td>
</tr>
<tr>
<td>1</td>
<td>162</td>
<td>D1Mit14</td>
<td>8.372794</td>
</tr>
<tr>
<td>1</td>
<td>170</td>
<td>D1Mit14</td>
<td>8.372794</td>
</tr>
<tr>
<td>1</td>
<td>159</td>
<td>D1Mit14</td>
<td>8.350341</td>
</tr>
<tr>
<td>1</td>
<td>73</td>
<td>D1Mit14</td>
<td>6.165395</td>
</tr>
<tr>
<td>1</td>
<td>65</td>
<td>D1Mit14</td>
<td>6.165395</td>
</tr>
<tr>
<td>1</td>
<td>88</td>
<td>D1Mit14</td>
<td>6.165395</td>
</tr>
<tr>
<td>1</td>
<td>184</td>
<td>D1Mit14</td>
<td>6.151606</td>
</tr>
<tr>
<td>1</td>
<td>241</td>
<td>D1Mit14</td>
<td>6.151606</td>
</tr>
</tbody>
</table>

... 16 1 215  D1Mit267 5.822192  
17 1 108  D1Mit267 5.822192  
18 1 138  D1Mit267 5.822192  
19 1 226  D1Mit267 5.822192  
20 1 199  D1Mit267 5.819250  
21 1 84   D1Mit267 5.808400  

> plot.geno(hyper, chr=1, ind=c(117:119,137:139,157:184))

QTL 2: Tutorial Seattle SISG: Yandell © 2010
R/qtl: 1 QTL interval mapping

```r
> hyper <- calc.genoprob(hyper, step=1, error.prob=0.01)
> out.em <- scanone(hyper)
> out.hk <- scanone(hyper, method="hk")
> summary(out.em, threshold=3)
  chr  pos  lod
  cl.loc45  1 48.3 3.52
  D4Mit164  4 29.5 8.02

> summary(out.hk, threshold=3)
  chr  pos  lod
  cl.loc45  1 48.3 3.55
  D4Mit164  4 29.5 8.09

> plot(out.em, chr = c(1,4,6,15))
> plot(out.hk, chr = c(1,4,6,15), add = TRUE, lty = 2)
```

black = EM  
blue = HK  

note bias where marker data are missing systematically
R/qtl: permutation threshold

```r
do operm.hk <- scanone(hyper, method="hk", n.perm=1000)
Doing permutation in batch mode ...
> summary(operm.hk, alpha=c(0.01,0.05))
LOD thresholds (1000 permutations)
   lod
1%  3.79
5%  2.78

> summary(out.hk, perms=operm.hk, alpha=0.05, pvalues=TRUE)
   chr  pos lod  pval
1   1 48.3 3.55 0.015
2   4 29.5 8.09 0.000
```
R/qtl: 2 QTL scan

```r
> hyper <- calc.genoprob(hyper, step=5, error.prob=0.01)
> out2.hk <- scantwo(hyper, method="hk")

--Running scanone
--Running scantwo
    (1,1)
    (1,2)
    ...
    (19,19)
    (19,X)
    (X,X)
> summary(out2.hk, thresholds=c(6.0, 4.7, 4.4, 4.7, 2.6))
	pos1f pos2f lod.full lod.fv1 lod.int
  c1 :c4   68.3  30.0    14.13    6.51   0.225
  c2 :c19  47.7  0.0    14.71    6.71   3.458
  c3 :c15  37.2  42.2    6.10    5.01   0.226
  c6 :c15  60.0  20.5    7.17    5.08   0.226
  c9 :c18  67.0  37.2    6.31    4.79   4.083
  c12:19  1.1  40.0    6.48    4.79   4.090

> plot(out2.hk, chr=c(1,4,6,15))

```

upper triangle/left scale: epistasis LOD
lower triangle/right scale: 2-QTL LOD
Effect & Interaction Plots

## Effect plots and interaction plot.

```r
hyper <- sim.geno(hyper, step=2, n.draws=16, error.prob=0.01)
effectplot(hyper, pheno.col = 1, mname1 = "D1Mit334")
effectplot(hyper, pheno.col = 1, mname1 = "D4Mit164")
markers <- find.marker(hyper, chr = c(6,15), pos = c(70,20))
effectplot(hyper, pheno.col = 1,
    mname1 = markers[1], mname2 = markers[2])
effectplot(hyper, pheno.col = 1,
    mname1 = markers[2], mname2 = markers[1])
```

## Strip plot of data (phenotype by genotype).

```r
plot.pxg(hyper, "D1Mit334")
plot.pxg(hyper, "D4Mit164")
plot.pxg(hyper, markers)
```
R/qtl: ANOVA imputation at QTL

> hyper <- sim.geno(hyper, steps=2, n.draws=16, error.prob=0.01)
> qtl <- makeqtl(hyper, chr = c(1, 1, 4, 6, 15), pos = c(50, 76, 30, 70, 20))
> my.formula <- y ~ Q1 + Q2 + Q3 + Q4 + Q5 + Q4:Q5
> out.fitqtl <- fitqtl(hyper, pheno.col = 1, qtl, formula = my.formula)
> summary(out.fitqtl)

Full model result
----------------------------------
Model formula is:  y ~ Q1 + Q2 + Q3 + Q4 + Q5 + Q4:Q5

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>LOD</th>
<th>%var</th>
<th>Pvalue(Chi2)</th>
<th>Pvalue(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>5789.089</td>
<td>964.84822</td>
<td>21.54994</td>
<td>32.76422</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Error</td>
<td>243</td>
<td>11879.847</td>
<td>48.88826</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>249</td>
<td>17668.936</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Drop one QTL at a time ANOVA table:
----------------------------------

<table>
<thead>
<tr>
<th>Chr</th>
<th>df</th>
<th>SS</th>
<th>LOD</th>
<th>%var</th>
<th>Pvalue(Chi2)</th>
<th>Pvalue(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1650</td>
<td>1</td>
<td>297.149</td>
<td>1.341</td>
<td>1.682</td>
<td>6.078</td>
<td>0.01438 *</td>
</tr>
<tr>
<td>1376</td>
<td>1</td>
<td>520.664</td>
<td>2.329</td>
<td>2.947</td>
<td>10.650</td>
<td>0.00126 **</td>
</tr>
<tr>
<td>6030</td>
<td>1</td>
<td>2842.089</td>
<td>11.644</td>
<td>16.085</td>
<td>58.134</td>
<td>5.50e-13 ***</td>
</tr>
<tr>
<td>15020</td>
<td>2</td>
<td>1083.842</td>
<td>4.740</td>
<td>6.134</td>
<td>11.085</td>
<td>2.47e-05 ***</td>
</tr>
<tr>
<td>6670:15820</td>
<td>1</td>
<td>955.268</td>
<td>4.199</td>
<td>5.406</td>
<td>19.540</td>
<td>1.49e-05 ***</td>
</tr>
</tbody>
</table>

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 1
selected R/qtl publications
www.stat.wisc.edu/~yandell/statgen

- www.rqtl.org
- tutorials and code at web site
  - www.rqtl.org/tutorials
- Broman et al. (2003 Bioinformatics)
  - R/qtl introduction
- Broman (2001 Lab Animal)
  - nice overview of QTL issues
- Broman & Sen 2009 book (Springer)

R/qtlbim (www.qtlbim.org)
- cross-compatible with R/qtl
- model selection for genetic architecture
  - epistasis, fixed & random covariates, GxE
  - samples multiple genetic architectures
  - examines summaries over nested models
- extensive graphics

> url.show("http://www.stat.wisc.edu/~yandell/qtlbim/rqtlbimtour.R")
R/qtlbim: tutorial
(www.stat.wisc.edu/~yandell/qtlbim)

> data(hyper)
## Drop X chromosome (for now).
> hyper <- subset(hyper, chr=1:19)
> hyper <- qb.genoprob(hyper, step=2)
## This is the time-consuming step:
> qbHyper <- qb.mcmc(hyper, pheno.col = 1)
## Here we get stored samples.
> data(qbHyper)
> summary(qbHyper)

R/qtlbim: initial summaries

> summary(qbHyper)

Bayesian model selection QTL mapping object qbHyper on cross object hyper
had 3000 iterations recorded at each 40 steps with 1200 burn-in steps.

Diagnostic summaries:

<table>
<thead>
<tr>
<th></th>
<th>nqtl</th>
<th>mean</th>
<th>envvar</th>
<th>varaa</th>
<th>varaa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>2.00</td>
<td>97.42</td>
<td>28.07</td>
<td>5.11</td>
<td>0.00</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>5.00</td>
<td>101.00</td>
<td>44.33</td>
<td>17.01</td>
<td>1.64</td>
</tr>
<tr>
<td>Median</td>
<td>7.00</td>
<td>101.30</td>
<td>48.57</td>
<td>20.06</td>
<td>4.58</td>
</tr>
<tr>
<td>Mean</td>
<td>6.54</td>
<td>101.30</td>
<td>48.80</td>
<td>20.31</td>
<td>5.32</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>8.00</td>
<td>101.70</td>
<td>53.11</td>
<td>23.48</td>
<td>7.86</td>
</tr>
<tr>
<td>Max</td>
<td>13.00</td>
<td>103.90</td>
<td>74.03</td>
<td>51.73</td>
<td>34.94</td>
</tr>
</tbody>
</table>

Percentages for number of QTL detected:

<table>
<thead>
<tr>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
</tbody>
</table>

Percentages for number of epistatic pairs detected:

<table>
<thead>
<tr>
<th>pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>29</td>
</tr>
</tbody>
</table>

Percentages for common epistatic pairs:

|  6.15 |  4.15 |  4.6 |  1.7 |  1.15 |  1.4 |  1.6 |  4.9 |  1.15 |  1.17 |  1.3 |  5.11 |  1.2 |  7.15 |  1.1 |
|  63 |  18 |  10 |  6 |  6 |  5 |  4 |  4 |  3 |  3 |  3 |  2 |  2 |  2 |  2 |

> plot(qb.diag(qbHyper, items = c("herit", "envvar")))
diagnostic summaries

R/qtlbim: 1-D (not 1-QTL!) scan

```r
> one <- qb.scanone(qbHyper, chr = c(1,4,6,15), type = "LPD")
> summary(one)

LPD of bp for main, epistasis, sum

<table>
<thead>
<tr>
<th>n.qtl</th>
<th>pos</th>
<th>m.pos</th>
<th>e.pos</th>
<th>main</th>
<th>epistasis</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>1.331</td>
<td>64.5</td>
<td>64.5</td>
<td>67.8</td>
<td>6.10</td>
<td>6.27</td>
</tr>
<tr>
<td>c4</td>
<td>1.377</td>
<td>29.5</td>
<td>29.5</td>
<td>29.5</td>
<td>11.49</td>
<td>11.61</td>
</tr>
<tr>
<td>c6</td>
<td>0.838</td>
<td>59.0</td>
<td>59.0</td>
<td>59.0</td>
<td>3.99</td>
<td>6.265</td>
</tr>
<tr>
<td>c15</td>
<td>0.961</td>
<td>17.5</td>
<td>17.5</td>
<td>17.5</td>
<td>1.30</td>
<td>6.325</td>
</tr>
</tbody>
</table>

> plot(one, scan = "main")
> plot(out.em, chr=c(1,4,6,15), add = TRUE, lty = 2)
> plot(one, scan = "epistasis")
```
1-QTL LOD vs. marginal LPD

most probable patterns

```r
> summary(qb.BayesFactor(qbHyper, item = "pattern"))

<table>
<thead>
<tr>
<th>nqtl posterior</th>
<th>prior</th>
<th>bf</th>
<th>bfse</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,4,6,15,6:15</td>
<td>5</td>
<td>0.03400 2.71e-05 24.30 2.360</td>
<td></td>
</tr>
<tr>
<td>1,4,6,15,6:15</td>
<td>6</td>
<td>0.00467 2.71e-05 24.30 2.360</td>
<td></td>
</tr>
<tr>
<td>1,1,4,6,6,15,6:15</td>
<td>6</td>
<td>0.00600 9.05e-06 12.80 3.020</td>
<td></td>
</tr>
<tr>
<td>1,1,4,5,6,15,6:15</td>
<td>7</td>
<td>0.00267 4.11e-06 12.60 4.450</td>
<td></td>
</tr>
<tr>
<td>1,4,6,15,15,6:15</td>
<td>6</td>
<td>0.00300 4.96e-06 11.70 3.910</td>
<td></td>
</tr>
<tr>
<td>1,4,6,6,6,15,6:15</td>
<td>6</td>
<td>0.00300 5.81e-06 10.00 3.330</td>
<td></td>
</tr>
<tr>
<td>1,2,4,6,15,6:15</td>
<td>6</td>
<td>0.00767 1.54e-05  9.66 2.010</td>
<td></td>
</tr>
<tr>
<td>1,4,5,6,15,6:15</td>
<td>6</td>
<td>0.00500 1.28e-05  7.56 1.950</td>
<td></td>
</tr>
<tr>
<td>1,1,4,19,6,15,6:15</td>
<td>7</td>
<td>0.00267 6.98e-06  7.41 2.620</td>
<td></td>
</tr>
<tr>
<td>1,4</td>
<td>2</td>
<td>0.01430 1.51e-05  1.84 0.279</td>
<td></td>
</tr>
<tr>
<td>1,1,2,4</td>
<td>4</td>
<td>0.00300 3.66e-05  1.59 0.529</td>
<td></td>
</tr>
<tr>
<td>1,2,4</td>
<td>3</td>
<td>0.00733 1.03e-04  1.38 0.294</td>
<td></td>
</tr>
<tr>
<td>1,1,4</td>
<td>3</td>
<td>0.00400 6.05e-05  1.28 0.370</td>
<td></td>
</tr>
<tr>
<td>1,4,19</td>
<td>3</td>
<td>0.00300 5.82e-05  1.00 0.333</td>
<td></td>
</tr>
</tbody>
</table>

> plot(qb.BayesFactor(qbHyper, item = "nqtl"))
```
hyper: number of QTL posterior, prior, Bayes factors

what is best estimate of QTL?

- find most probable pattern
  - 1,4,6,15,6:15 has posterior of 3.4%
- estimate locus across all nested patterns
  - Exact pattern seen ~100/3000 samples
  - Nested pattern seen ~2000/3000 samples
- estimate 95% confidence interval using quantiles

```r
> best <- qb.best(qbHyper)
> summary(best)$best

<table>
<thead>
<tr>
<th>chrom</th>
<th>locus</th>
<th>locus.LCL</th>
<th>locus.UCL</th>
<th>n.qtl</th>
</tr>
</thead>
<tbody>
<tr>
<td>247</td>
<td>1</td>
<td>69.9</td>
<td>24.44875</td>
<td>95.7985 0.8026667</td>
</tr>
<tr>
<td>245</td>
<td>4</td>
<td>29.5</td>
<td>14.20000</td>
<td>74.3000 0.8800000</td>
</tr>
<tr>
<td>248</td>
<td>6</td>
<td>59.0</td>
<td>13.83333</td>
<td>66.7000 0.7096667</td>
</tr>
<tr>
<td>246</td>
<td>15</td>
<td>19.5</td>
<td>13.10000</td>
<td>55.7000 0.8450000</td>
</tr>
</tbody>
</table>
```

```r
> plot(best)
```
what patterns are “near” the best?

- size & shade – posterior
- distance between patterns
  - sum of squared attenuation
  - match loci between patterns
  - squared attenuation = \((1-2r)^2\)
  - sq.atten in scale of LOD & LPD
- multidimensional scaling
  - MDS projects distance onto 2-D
  - think mileage between cities

> target <- qb.best(qbHyper)$model[[1]]
> summary(qb.close(qbHyper, target))

<table>
<thead>
<tr>
<th>score by sample number of qtl</th>
<th>Percent</th>
<th>Min. 1st Qu. Median Mean 3rd Qu. Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2 1.437 1.735 1.919 1.834 1.919 2.000</td>
</tr>
<tr>
<td></td>
<td>3 1.351 1.735 1.916 1.900 1.919 2.916</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 1.270 1.916 2.437 2.648 3.574 4.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 1.295 1.919 2.835 2.798 3.611 4.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6 1.257 2.254 3.451 3.029 3.648 4.000</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>13</td>
<td>3.694</td>
<td>3.694 3.694 3.694 3.694 3.694</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>score by sample chromosome pattern</th>
<th>Percent</th>
<th>Min. 1st Qu. Median Mean 3rd Qu. Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>481,4,6,15,6:15</td>
<td>3.4 2.946 3.500 3.630 3.613 3.758 4.000</td>
<td></td>
</tr>
<tr>
<td>281,4</td>
<td>1.4 1.437 1.735 1.919 1.832 1.919 2.000</td>
<td></td>
</tr>
<tr>
<td>581,2,4,6,15,6:15</td>
<td>0.8 3.137 3.536 3.622 3.621 3.777 3.923</td>
<td></td>
</tr>
<tr>
<td>381,2,4</td>
<td>0.7 1.351 1.700 1.821 1.808 1.919 2.000</td>
<td></td>
</tr>
<tr>
<td>581,1,4,6,15,6:15</td>
<td>0.6 3.257 3.484 3.563 3.575 3.698 3.916</td>
<td></td>
</tr>
<tr>
<td>581,4,5,6,15,6:15</td>
<td>0.5 3.237 3.515 3.595 3.622 3.777 3.923</td>
<td></td>
</tr>
<tr>
<td>581,4,6,6,15,6:15</td>
<td>0.5 3.203 3.541 3.646 3.631 3.757 3.835</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

> plot(close)
> plot(close, category = "nqtl")
how close are other patterns?

R/qtlbim: automated QTL selection

```r
> hpd <- qb.hpdone(qbHyper, profile = "2logBF")
> summary(hpd)

<table>
<thead>
<tr>
<th>chr</th>
<th>n.qtl</th>
<th>pos</th>
<th>lo.50%</th>
<th>hi.50%</th>
<th>2logBF</th>
<th>A</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.829</td>
<td>64.5</td>
<td>72.1</td>
<td>6.692</td>
<td>103.611</td>
<td>99.090</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3.228</td>
<td>29.5</td>
<td>31.7</td>
<td>11.169</td>
<td>104.584</td>
<td>98.020</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>1.033</td>
<td>59.0</td>
<td>66.7</td>
<td>6.054</td>
<td>99.637</td>
<td>102.965</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>0.159</td>
<td>17.5</td>
<td>17.5</td>
<td>5.837</td>
<td>101.972</td>
<td>100.702</td>
</tr>
</tbody>
</table>

> plot(hpd)
```
2log(BF) scan with 50% HPD region

R/qtlbim: 2-D \textit{(not 2-QTL)} scans

\begin{verbatim}
> two <- qb.scantwo(qbHyper, chr = c(6,15),
>                   type = "2logBF")
> plot(two)

> plot(two, chr = 6, slice = 15)
> plot(two, chr = 15, slice = 6)

> two.lpd <- qb.scantwo(qbHyper, chr = c(6,15),
>                        type = "LPD")
> plot(two.lpd, chr = 6, slice = 15)
> plot(two.lpd, chr = 15, slice = 6)
\end{verbatim}
2-D plot of 2logBF: chr 6 & 15

1-D Slices of 2-D scans: chr 6 & 15
R/qtlbim: slice of epistasis

> slice <- qb.slicetwo(qbHyper, c(6,15), c(59,19.5))
> summary(slice)

2logBF of bp for epistasis

<table>
<thead>
<tr>
<th>n.qtl</th>
<th>pos</th>
<th>m.pos</th>
<th>e.pos</th>
<th>epistasis</th>
<th>slice</th>
</tr>
</thead>
<tbody>
<tr>
<td>c6</td>
<td>0.838</td>
<td>59.0</td>
<td>59.0</td>
<td>66.7</td>
<td>15.8</td>
</tr>
<tr>
<td>c15</td>
<td>0.961</td>
<td>17.5</td>
<td>17.5</td>
<td>17.5</td>
<td>15.5</td>
</tr>
</tbody>
</table>

cellmean of bp for AA, HA, AH, HH

<table>
<thead>
<tr>
<th>n.qtl</th>
<th>pos</th>
<th>m.pos</th>
<th>AA</th>
<th>HA</th>
<th>AH</th>
<th>HH</th>
<th>slice</th>
</tr>
</thead>
<tbody>
<tr>
<td>c6</td>
<td>0.838</td>
<td>59.0</td>
<td>97.4</td>
<td>105</td>
<td>102</td>
<td>100.8</td>
<td>18.1</td>
</tr>
<tr>
<td>c15</td>
<td>0.961</td>
<td>17.5</td>
<td>99.8</td>
<td>103</td>
<td>104</td>
<td>98.5</td>
<td>60.6</td>
</tr>
</tbody>
</table>

estimate of bp for epistasis

<table>
<thead>
<tr>
<th>n.qtl</th>
<th>pos</th>
<th>m.pos</th>
<th>e.pos</th>
<th>epistasis</th>
<th>slice</th>
</tr>
</thead>
<tbody>
<tr>
<td>c6</td>
<td>0.838</td>
<td>59.0</td>
<td>66.7</td>
<td>-7.86</td>
<td>18.1</td>
</tr>
<tr>
<td>c15</td>
<td>0.961</td>
<td>17.5</td>
<td>17.5</td>
<td>-8.72</td>
<td>60.6</td>
</tr>
</tbody>
</table>

> plot(slice, figs = c("effects", "cellmean", "effectplot"))
selected publications
www.stat.wisc.edu/~yandell/statgen

- www.qtlbim.org
- vignettes in R/qtlbim package
  - overview/comparison of QTL methods
- Yandell et al. (2007 *Bioinformatics*)
  - R/qtlbim introduction
- Yi et al. (2005 *Genetics*, 2007 *Genetics*)
  - methodology of R/qtlbim
Bayesian Interval Mapping

1. Bayesian strategy
2. Markov chain sampling
3. sampling genetic architectures
4. criteria for model selection

QTL model selection: key players

- observed measurements
  - $y$ = phenotypic trait
  - $m$ = markers & linkage map
  - $i$ = individual index (1,…,$n$)
- missing data
  - missing marker data
  - $q$ = QT genotypes
    - alleles QQ, Qq, or qq at locus
- unknown quantities
  - $\lambda$ = QT locus (or loci)
  - $\mu$ = phenotype model parameters
  - $\gamma$ = QTL model/genetic architecture
- $\text{pr}(q|m, \lambda, \gamma)$ genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for $q$ given $m$
- $\text{pr}(y|q, \mu, \gamma)$ phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters $\mu$ (could be non-parametric)

QTL 2: Bayes Seattle SISG: Yandell © 2010
1. Bayesian strategy for QTL study

- augment data \((y,m)\) with missing genotypes \(q\)
- study unknowns \((\mu, \lambda, \gamma)\) given augmented data \((y,m,q)\)
  - find better genetic architectures \(\gamma\)
  - find most likely genomic regions = QTL = \(\lambda\)
  - estimate phenotype parameters = genotype means = \(\mu\)
- sample from posterior in some clever way
  - multiple imputation (Sen Churchill 2002)
  - Markov chain Monte Carlo (MCMC)
    - (Satagopan et al. 1996; Yi et al. 2005, 2007)

\[
\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{constant}}
\]

\[
\text{posterior for } q, \mu, \lambda, \gamma = \frac{\text{phenotype likelihood} \times [\text{prior for } q, \mu, \lambda, \gamma]}{\text{constant}}
\]

\[
\text{pr}(q, \mu, \lambda, \gamma | y, m) = \frac{\text{pr}(y | q, \mu, \gamma) \times [\text{pr}(q | m, \lambda, \gamma) \times \text{pr} \times (\lambda | m, \gamma) \times \text{pr} \times (\gamma)]}{\text{pr}(y | m)}
\]
Bayes posterior for normal data

model

\[ y_i = \mu + e_i \]

environment

\[ e \sim N(0, \sigma^2) \], \( \sigma^2 \) known

likelihood

\[ y \sim N(\mu, \sigma^2) \]

prior

\[ \mu \sim N(\mu_0, \kappa \sigma^2), \kappa \text{ known} \]

posterior:

single individual

\[ \mu \sim N(\mu_0 + b_1(y_1 - \mu_0), b_1 \sigma^2) \]

sample of \( n \) individuals

\[ \mu \sim N(b_n \bar{y}_n + (1 - b_n) \mu_0, b_n \sigma^2 / n) \]

with \( \bar{y}_n = \sum_{i=1}^{n} y_i / n \)

shrinkage factor

(shrinks to 1)

\[ b_n = \frac{\kappa n}{\kappa n + 1} \to 1 \]

what values are the genotypic means?

phenotype model \( \text{pr}(y|q, \mu) \)

QTL 2: Bayes Seattle SISG: Yandell © 2010
Bayes posterior QTL means

posterior centered on sample genotypic mean
but shrunken slightly toward overall mean

phenotype mean:  \[ E(y \mid q) = \mu_q \quad V(y \mid q) = \sigma^2 \]

genotypic prior:  \[ E(\mu_q) = \bar{y}, \quad V(\mu_q) = \kappa \sigma^2 \]

posterior:  \[ E(\mu_q \mid y) = b_q \bar{y}_q + (1 - b_q) \bar{y}, \quad V(\mu_q \mid y) = b_q \sigma^2 / n_q \]

\[ n_q = \text{count}\{q_i = q\} \quad \bar{y}_q = \frac{\text{sum } y_i / n_q}{\text{count } \{q_i \neq q\}} \]

shrinkage:  \[ b_q = \frac{\kappa \eta_q}{\kappa \eta_q + 1} \rightarrow 1 \]

partition genotypic effects on phenotype

• phenotype depends on genotype
• genotypic value partitioned into
  – main effects of single QTL
  – epistasis (interaction) between pairs of QTL

\[ \mu_q = \beta_0 + \beta_q = E(Y; q) \]
\[ \beta_q = \beta(q_2) + \beta(q_2) + \beta(q_1, q_2) \]
partition genotypic variance

• consider same 2 QTL + epistasis

• centering variance \( V(\beta_0) = \kappa_0 \sigma^2 = s^2 \)

• genotypic variance \( V(\beta_q) = \kappa_1 \sigma^2 = \sigma_q^2 = \sigma_1^2 + \sigma_2^2 + \sigma_{12}^2 \)

• heritability \( h_q^2 = \frac{\sigma_q^2}{\sigma_q^2 + \sigma^2} = h_1^2 + h_2^2 + h_{12}^2 \)

posterior mean \( \approx \) LS estimate

\( \beta_q \mid y \sim N(b_q \hat{\beta}_q, b_q C_q \sigma^2) \)

\( \approx N(\hat{\beta}_q, C_q \sigma^2) \)

LS estimate \( \hat{\beta}_q = \sum_i [\sum_j \hat{\beta}(q_{ij})] = \sum_i w_{qi} y_i \)

variance \( V(\hat{\beta}_q) = \sum_i w_{qi}^2 \sigma^2 = C_q \sigma^2 \)

shrinkage \( b_q = \kappa_1 / (\kappa_1 + C_q) \rightarrow 1 \)
\begin{equation}
\text{pr}(q|m, \lambda) \text{ recombination model}
\end{equation}
\begin{equation}
\text{pr}(q|m, \lambda) = \text{pr(geno | map, locus)} \approx \\
\text{pr(geno | flanking markers, locus)}
\end{equation}

\textit{distance along chromosome}
what are likely QTL genotypes \( q \)?

how does phenotype \( y \) improve guess?

what are probabilities for genotype \( q \) between markers?

recombinants AA:AB all 1:1 if ignore \( y \) and if we use \( y \)?

posterior on QTL genotypes \( q \)

• full conditional of \( q \) given data, parameters
  – proportional to prior \( \text{pr}(q \mid m, \lambda) \)
    • weight toward \( q \) that agrees with flanking markers
  – proportional to likelihood \( \text{pr}(y \mid q, \mu) \)
    • weight toward \( q \) with similar phenotype values
  – posterior recombination model balances these two
• this is the E-step of EM computations

\[
\text{pr}(q \mid y, m, \mu, \lambda) = \frac{\text{pr}(y \mid q, \mu) \cdot \text{pr}(q \mid m, \lambda)}{\text{pr}(y \mid m, \mu, \lambda)}
\]
Where are the loci $\lambda$ on the genome?

- prior over genome for QTL positions
  - flat prior = no prior idea of loci
  - or use prior studies to give more weight to some regions
- posterior depends on QTL genotypes $q$
  \[
  \text{pr}(\lambda \mid m, q) = \frac{\text{pr}(\lambda) \cdot \text{pr}(q \mid m, \lambda)}{\text{constant}}
  \]
  - constant determined by averaging
    - over all possible genotypes $q$
    - over all possible loci $\lambda$ on entire map
- no easy way to write down posterior

what is the genetic architecture $\gamma$?

- which positions correspond to QTLs?
  - priors on loci (previous slide)
- which QTL have main effects?
  - priors for presence/absence of main effects
    - same prior for all QTL
    - can put prior on each d.f. (1 for BC, 2 for F2)
- which pairs of QTL have epistatic interactions?
  - prior for presence/absence of epistatic pairs
    - depends on whether 0,1,2 QTL have main effects
    - epistatic effects less probable than main effects
\[ \gamma = \text{genetic architecture:} \]

\text{loci:}
- main QTL
- epistatic pairs

\text{effects:}
- add, dom
- aa, ad, dd

Bayesian priors & posteriors

- augmenting with missing genotypes \( q \)
  - prior is recombination model
  - posterior is (formally) E step of EM algorithm
- sampling phenotype model parameters \( \mu \)
  - prior is “flat” normal at grand mean (no information)
  - posterior shrinks genotypic means toward grand mean
  - (details for unexplained variance omitted here)
- sampling QTL loci \( \lambda \)
  - prior is flat across genome (all loci equally likely)
- sampling QTL genetic architecture model \( \gamma \)
  - number of QTL
    - prior is Poisson with mean from previous IM study
  - genetic architecture of main effects and epistatic interactions
    - priors on epistasis depend on presence/absence of main effects
2. Markov chain sampling

- construct Markov chain around posterior
  - want posterior as stable distribution of Markov chain
  - in practice, the chain tends toward stable distribution
    - initial values may have low posterior probability
    - burn-in period to get chain mixing well
- sample QTL model components from full conditionals
  - sample locus $\lambda$ given $q, \gamma$ (using Metropolis-Hastings step)
  - sample genotypes $q$ given $\lambda, \mu, \gamma$ (using Gibbs sampler)
  - sample effects $\mu$ given $q, y, \gamma$ (using Gibbs sampler)
  - sample QTL model $\gamma$ given $\lambda, \mu, y, q$ (using Gibbs or M-H)

\[
(\lambda, q, \mu, \gamma) \sim \text{pr}(\lambda, q, \mu, \gamma | y, m)
\]

\[
(\lambda, q, \mu, \gamma)_1 \rightarrow (\lambda, q, \mu, \gamma)_2 \rightarrow \cdots \rightarrow (\lambda, q, \mu, \gamma)_N
\]

MCMC sampling of unknowns $(q, \mu, \lambda)$ for given genetic architecture $\gamma$

- Gibbs sampler
  - genotypes $q$
  - effects $\mu$
  - not loci $\lambda$

\[
\begin{align*}
q &\sim \text{pr}(q | y, m, \mu, \lambda) \\
\mu &\sim \frac{\text{pr}(y | q, \mu) \text{pr}(\mu)}{\text{pr}(y | q)} \\
\lambda &\sim \frac{\text{pr}(q | m, \lambda) \text{pr}(\lambda | m)}{\text{pr}(q | m)}
\end{align*}
\]

- Metropolis-Hastings sampler
  - extension of Gibbs sampler
  - does not require normalization
    - $\text{pr}(q | m) = \sum_{\lambda} \text{pr}(q | m, \lambda) \text{pr}(\lambda)$
Gibbs sampler
for two genotypic means

• want to study two correlated effects
  – could sample directly from their bivariate distribution
  – assume correlation $\rho$ is known
• instead use Gibbs sampler:
  – sample each effect from its full conditional given the other
  – pick order of sampling at random
  – repeat many times

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

$$\mu_1 \sim N\left( \rho \mu_2, 1 - \rho^2 \right)$$

$$\mu_2 \sim N\left( \rho \mu_1, 1 - \rho^2 \right)$$
full conditional for locus

• cannot easily sample from locus full conditional
  \[ \text{pr}(\lambda | y, m, \mu, q) = \frac{\text{pr}(\lambda | m, q)}{\text{pr}(q | m, \lambda) \text{pr}(\lambda) / \text{constant}} \]
  \[ = \text{pr}(q | m, \lambda) \text{pr}(\lambda) / \text{constant} \]
  \[ = \text{pr}(q | m, \lambda) \text{pr}(\lambda) / \text{constant} \]

• constant is very difficult to compute explicitly
  – must average over all possible loci \( \lambda \) over genome
  – must do this for every possible genotype \( q \)

• Gibbs sampler will not work in general
  – but can use method based on ratios of probabilities
  – Metropolis-Hastings is extension of Gibbs sampler

Metropolis-Hastings idea

• want to study distribution \( f(\lambda) \)
  – take Monte Carlo samples
    • unless too complicated
    – take samples using ratios of \( f \)
  – Metropolis-Hastings samples:
    – propose new value \( \lambda^* \)
      • near (?) current value \( \lambda \)
      • from some distribution \( g \)
    – accept new value with prob \( a \)
      • Gibbs sampler: \( a = 1 \) always

\[
a = \min\left( 1, \frac{f(\lambda^*) g(\lambda^* - \lambda)}{f(\lambda) g(\lambda - \lambda^*)} \right)
\]
Metropolis-Hastings for locus $\lambda$

- Added twist: occasionally propose from entire genome

Metropolis-Hastings samples

- $N = 200$ samples
- $N = 1000$ samples

QTL 2: Bayes Seattle SISG: Yandell © 2010
3. sampling genetic architectures

• search across genetic architectures $\gamma$ of various sizes
  – allow change in number of QTL
  – allow change in types of epistatic interactions

• methods for search
  – reversible jump MCMC
  – Gibbs sampler with loci indicators

• complexity of epistasis
  – Fisher-Cockerham effects model
  – general multi-QTL interaction & limits of inference

\[ \gamma = 1 \text{ QTL} : Y = \beta_0 + \beta(q_{1}) + e \]
\[ \gamma = 2 \text{ QTL} : Y = \beta_0 + \beta(q_{1}) + \beta(q_{2}) + e \]
geometry of reversible jump

Move Between Models

Reversible Jump Sequence

c21 = 0.7

m=2

m=1

geometry allowing q and λ to change

a short sequence

first 1000 with m<3
collinear QTL = correlated effects

4-week

\[
\begin{array}{c}
\text{effect 2} \\
\text{effect 1}
\end{array}
\]

\[\text{cor} = -0.81\]

8-week

\[
\begin{array}{c}
\text{effect 2} \\
\text{effect 1}
\end{array}
\]

\[\text{cor} = -0.7\]

- linked QTL = collinear genotypes
  - correlated estimates of effects (negative if in coupling phase)
  - sum of linked effects usually fairly constant

QTL 2: Bayes Seattle SISG: Yandell © 2010

- update QTL model \( \gamma \) with probability \( 1-b(\gamma) - d(\gamma) \)
  - update current model using full conditionals
  - sample QTL loci, effects, and genotypes
- add a locus with probability \( b(\gamma) \)
  - propose a new locus along genome
  - innovate new genotypes at locus and phenotype effect
  - decide whether to accept the “birth” of new locus
- drop a locus with probability \( d(\gamma) \)
  - propose dropping one of existing loci
  - decide whether to accept the “death” of locus

QTL 2: Bayes Seattle SISG: Yandell © 2010
Gibbs sampler with loci indicators

- consider only QTL at pseudomarkers
  - every 1-2 cM
  - modest approximation with little bias
- use loci indicators in each pseudomarker
  - $\gamma = 1$ if QTL present
  - $\gamma = 0$ if no QTL present
- Gibbs sampler on loci indicators $\gamma$
  - relatively easy to incorporate epistasis
  - Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005 Genetics)
    - (see earlier work of Nengjun Yi and Ina Hoeschele)

$$\mu_q = \mu + \gamma_1 \beta(q_1) + \gamma_2 \beta(q_2), \quad \gamma_k = 0,1$$

Bayesian shrinkage estimation

- soft loci indicators
  - strength of evidence for $\lambda_j$ depends on $\gamma$
  - $0 \leq \gamma \leq 1$ (grey scale)
  - shrink most $\gamma$s to zero
- Wang et al. (2005 Genetics)
  - Shizhong Xu group at U CA Riverside

$$\mu_q = \beta_0 + \gamma_1 \beta_1(q_1) + \gamma_2 \beta_2(q_1), \quad 0 \leq \gamma_k \leq 1$$
other model selection approaches

- include all potential loci in model
- assume “true” model is “sparse” in some sense
- Sparse partial least squares
  - Chun, Keles (2009 *Genetics*; 2010 *JRSSB*)
- LASSO model selection
  - Foster (2006); Foster Verbyla Pitchford (2007 *JABES*)
  - Xu (2007 *Biometrics*); Yi Xu (2007 *Genetics*)
  - Shi Wahba Wright Klein Klein (2008 Stat & Infer)

4. criteria for model selection
balance fit against complexity

- classical information criteria
  - penalize likelihood $L$ by model size $|\gamma|$
  - $IC = -2 \log L(\gamma \mid y) + \text{penalty}(\gamma)$
  - maximize over unknowns
- Bayes factors
  - marginal posteriors $pr(y \mid \gamma)$
  - average over unknowns
classical information criteria

• start with likelihood $L(\gamma | y, m)$
  – measures fit of architecture ($\gamma$) to phenotype ($y$)
    • given marker data ($m$)
  – genetic architecture ($\gamma$) depends on parameters
    • have to estimate loci ($\mu$) and effects ($\lambda$)
• complexity related to number of parameters
  – $|\gamma|$ = size of genetic architecture
    • BC: $|\gamma| = 1 + n.qtl + n.qtl(n.qtl - 1) = 1 + 4 + 12 = 17$
    • F2: $|\gamma| = 1 + 2n.qtl + 4n.qtl(n.qtl - 1) = 1 + 8 + 48 = 57$

construct information criteria

• construct information criteria
  – balance fit to complexity
  – Akaike $\text{AIC} = -2 \log(L) + 2 |\gamma|$
  – Bayes/Schwartz $\text{BIC} = -2 \log(L) + |\gamma| \log(n)$
  – Broman $\text{BIC}_\delta = -2 \log(L) + \delta |\gamma| \log(n)$
  – general form: $\text{IC} = -2 \log(L) + |\gamma| D(n)$
• compare models
  – hypothesis testing: designed for one comparison
    • $2 \log[LR(\gamma_1, \gamma_2)] = L(y|m, \gamma_2) - L(y|m, \gamma_1)$
  – model selection: penalize complexity
    • $\text{IC}(\gamma_1, \gamma_2) = 2 \log[LR(\gamma_1, \gamma_2)] + (|\gamma_2| - |\gamma_1|) D(n)$
information criteria vs. model size

- WinQTL 2.0
- SCD data on F2
- A=AIC
- 1=BIC(1)
- 2=BIC(2)
- d=BIC(\( \delta \))
- models
  - 1,2,3,4 QTL
    - 2+5+9+2
  - epistasis
    - 2:2 AD

Bayes factors

- ratio of model likelihoods
  - ratio of posterior to prior odds for architectures
  - averaged over unknowns
    \[ B_{12} = \frac{pr(\gamma_1 \mid y,m) / pr(\gamma_2 \mid y,m)}{pr(\gamma_1) / pr(\gamma_2)} = \frac{pr(y \mid m,\gamma_1)}{pr(y \mid m,\gamma_2)} \]
- roughly equivalent to BIC
  - BIC maximizes over unknowns
  - BF averages over unknowns
    \[ -2 \log(B_{12}) = -2 \log(LR) - (|\gamma_2| - |\gamma_1|) \log(n) \]
issues in computing Bayes factors

- $BF$ insensitive to shape of prior on $\gamma$
  - geometric, Poisson, uniform
  - precision improves when prior mimics posterior
- $BF$ sensitivity to prior variance on effects $\theta$
  - prior variance should reflect data variability
  - resolved by using hyper-priors
    - automatic algorithm; no need for user tuning
- easy to compute Bayes factors from samples
  - sample posterior using MCMC
  - posterior $\text{pr}(\gamma | y, m)$ is marginal histogram
Bayes factors & genetic architecture $\gamma$

- $|\gamma| =$ number of QTL
  - prior $\text{pr}(\gamma)$ chosen by user
  - posterior $\text{pr}(\gamma | y, m)$
    - sampled marginal histogram
    - shape affected by prior $\text{pr}(A)$

$$BF_{\gamma_1, \gamma_2} = \frac{\text{pr}(\gamma_1 | y, m) / \text{pr}(\gamma_1)}{\text{pr}(\gamma_2 | y, m) / \text{pr}(\gamma_2)}$$

- pattern of QTL across genome
- gene action and epistasis

BF sensitivity to fixed prior for effects

$$\beta_{ij} \sim N\left(0, \sigma^2 / m \right) \sigma^2_G = h^2 \sigma^2_{total}, h^2 \text{ fixed}$$
BF insensitivity to random effects prior

\[ \beta_{ij} \sim N\left(0, \sigma^2_G / m\right), \sigma^2_G = h^2 \sigma^2_{total} + \frac{1}{2} h^2 \sim \text{Beta}(a, b) \]
examples in detail

• simulation study (after Stephens & Fisch (1998))
• obesity in mice ($n = 421$)
  – epistatic QTLs with no main effects
• expression phenotype (SCD1) in mice ($n = 108$)
  – multiple QTL and epistasis
• mapping two correlated phenotypes
  – Jiang & Zeng 1995 paper
  – *Brassica napus* vernalization
• gonad shape in *Drosophila* spp. (insect) ($n = 1000$)
  – multiple traits reduced by PC
  – many QTL and epistasis

---

simulation with 8 QTL

- simulated F2 intercross, 8 QTL
  - (Stephens, Fisch 1998)
  - $n=200$, heritability = 50%
  - detected 3 QTL
- increase to detect all 8
  - $n=500$, heritability to 97%
loci pattern across genome

- notice which chromosomes have persistent loci
- best pattern found 42% of the time

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Count of 8000</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3371</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>751</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>377</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>218</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>218</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>198</td>
</tr>
</tbody>
</table>

obesity in CAST/Ei BC onto M16i

- 421 mice (Daniel Pomp)
  - (213 male, 208 female)
- 92 microsatellites on 19 chromosomes
  - 1214 cM map
- subcutaneous fat pads
  - pre-adjusted for sex and dam effects
- Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005) *Genetics*
non-epistatic analysis

single QTL LOD profile

multiple QTL Bayes factor profile

posterior profile of main effects in epistatic analysis

main effects & heritability profile

Bayes factor profile
posterior profile of main effects in epistatic analysis

model selection via Bayes factors for epistatic model

number of QTL QTL pattern
scatterplot estimates of epistatic loci

stronger epistatic effects
studying diabetes in an F2

- segregating cross of inbred lines
  - B6.ob x BTBR.ob → F1 → F2
  - selected mice with ob/ob alleles at leptin gene (chr 6)
  - measured and mapped body weight, insulin, glucose at various ages
    (Stoehr et al. 2000 Diabetes)
  - sacrificed at 14 weeks, tissues preserved
- gene expression data
  - Affymetrix microarrays on parental strains, F1
    - key tissues: adipose, liver, muscle, β-cells
    - novel discoveries of differential expression (Nadler et al. 2000 PNAS; Lan et al. 2002 in review; Ntambi et al. 2002 PNAS)
  - RT-PCR on 108 F2 mice liver tissues
    - 15 genes, selected as important in diabetes pathways
    - SCD1, PEPCK, ACO, FAS, GPAT, PPARgamma, PPARalpha, G6Pase, PDI,…

Multiple Interval Mapping (QTLCart)

SCD1: multiple QTL plus epistasis!

![Graph showing LOD scores and effect on chromosomes 2, 5, and 9.](QTL 2: Data Seattle SISG: Yandell © 2010 135)
Bayesian model assessment: number of QTL for SCD1

Bayesian LOD and $h^2$ for SCD1
Bayesian model assessment: chromosome QTL pattern for SCD1

trans-acting QTL for SCD1
(no epistasis yet: see Yi, Xu, Allison 2003)
2-D scan: assumes only 2 QTL!

epistasis
LOD peaks

joint LOD peaks

sub-peaks can be easily overlooked!
epistatic model fit

Cockerham epistatic effects
co-mapping multiple traits

- avoid reductionist approach to biology
  - address physiological/biochemical mechanisms
  - Schmalhausen (1942); Falconer (1952)
- separate close linkage from pleiotropy
  - 1 locus or 2 linked loci?
- identify epistatic interaction or canalization
  - influence of genetic background
- establish QTL x environment interactions
- decompose genetic correlation among traits
- increase power to detect QTL

interplay of pleiotropy & correlation

pleiotropy only
Korol et al. (2001)
correlation only
both
3 correlated traits
(Jiang Zeng 1995)

Ellipses centered on genotypic value
width for nominal frequency
main axis angle environmental correlation
3 QTL, F2
27 genotypes

Note signs of
genetic and
environmental

correlation

pleiotropy or close linkage?

2 traits, 2 qtl/trait
pleiotropy @ 54cM
linkage @ 114,128cM

**Brassica napus**: 2 correlated traits

- 4-week & 8-week vernalization effect
  - log(days to flower)
- genetic cross of
  - Stellar (annual canola)
  - Major (biennial rapeseed)
- 105 F1-derived double haploid (DH) lines
  - homozygous at every locus (QQ or qq)
- 10 molecular markers (RFLPs) on LG9
  - two QTLs inferred on LG9 (now chromosome N2)
  - corroborated by Butruille (1998)
  - exploiting synteny with *Arabidopsis thaliana*

**QTL with GxE or Covariates**

- adjust phenotype by covariate
  - covariate(s) = environment(s) or other trait(s)
- additive covariate
  - covariate adjustment same across genotypes
  - “usual” analysis of covariance (ANCOVA)
- interacting covariate
  - address GxE
  - capture genotype-specific relationship among traits
- another way to think of multiple trait analysis
  - examine single phenotype adjusted for others
R/qtl & covariates

- additive and/or interacting covariates
- test for QTL after adjusting for covariates

```r
## Get Brassica data.
library(qtlbim)
data(Bnapus)
Bnapus <- calc.genoprob(Bnapus, step = 2, error = 0.01)

## Scatterplot of two phenotypes: 4wk & 8wk flower time.
plot(Bnapus$pheno$log10flower4, Bnapus$pheno$log10flower8)

## Unadjusted IM scans of each phenotype.
fl8 <- scanone(Bnapus, find.pheno(Bnapus, "log10flower8"))
fl4 <- scanone(Bnapus, find.pheno(Bnapus, "log10flower4"))
plot(fl4, fl8, chr = "N2", col = rep(1,2), lty = 1:2,
     main = "solid = 4wk, dashed = 8wk", lwd = 4)
```

QTL 2: Data Seattle SISG: Yandell © 2010 151
R/qtl & covariates

- additive and/or interacting covariates
- test for QTL after adjusting for covariates

```r
## IM scan of 8wk adjusted for 4wk.
## Adjustment independent of genotype
fl8.4 <- scanone(Bnapus, find.pheno(Bnapus, "log10flower8"),
                 addcov = Bnapus$pheno$log10flower4)

## IM scan of 8wk adjusted for 4wk.
## Adjustment changes with genotype.
fl8.4 <- scanone(Bnapus, find.pheno(Bnapus, "log10flower8"),
                 intcov = Bnapus$pheno$log10flower4)

plot(fl8, fl8.4a, fl8.4, chr = "N2",
     main = "solid = 8wk, dashed = addcov, dotted = intcov")
```
## Set up data frame with peak markers, traits.
markers <- c("E38M50.133", "ec2e5a", "wg7f3a")
tmpdata <- data.frame(pull.geno(Bnapus)[,markers])
tmpdata$fl4 <- Bnapus$pheno$log10flower4
tmpdata$fl8 <- Bnapus$pheno$log10flower8

## Scatterplots grouped by marker.
library(lattice)
xyplot(fl8 ~ fl4, tmpdata, group = wg7f3a,
      col = "black", pch = 3:4, cex = 2, type = c("p","r"),
      xlab = "log10(4wk flower time)",
      ylab = "log10(8wk flower time)",
      main = "marker at 47cM")

xyplot(fl8 ~ fl4, tmpdata, group = E38M50.133,
      col = "black", pch = 3:4, cex = 2, type = c("p","r"),
      xlab = "log10(4wk flower time)",
      ylab = "log10(8wk flower time)",
      main = "marker at 80cM")

scatterplot adjusted for covariate
R/qtlbim and GxE

- similar idea to R/qtl
  - fixed and random additive covariates
  - GxE with fixed covariate
- multiple trait analysis tools coming soon
  - theory & code mostly in place
  - properties under study
  - expect in R/qtlbim later this year
  - Samprit Banerjee (N Yi, advisor)

Reducing many phenotypes to 1

- *Drosophila mauritiana* x *D. simulans*
  - reciprocal backcrosses, ~500 per bc
- response is “shape” of reproductive piece
  - trace edge, convert to Fourier series
  - reduce dimension: first principal component
- many linked loci
  - brief comparison of CIM, MIM, BIM
PC for two correlated phenotypes

shape phenotype via PC

Figure 5.—A plot of the first two principal components of the Fourier coefficients from posterior lobe outlines. Many individuals from each of the six genetic classes are represented. Each point represents an average of scores from the left and right sides of an individual (with a few exceptions for which the score is from one side only). The percentage of variation in the Fourier coefficients accounted for by each principal component is given in parentheses. Liu et al. (1996) Genetics
shape phenotype in BC study indexed by PC1

Zeng et al. (2000)  
CIM vs. MIM

composite interval mapping  
(Liu et al. 1996)

narrow peaks  
miss some QTL

multiple interval mapping  
(Zeng et al. 2000)

triangular peaks

both conditional 1-D scans  
fixing all other "QTL"
CIM, MIM and IM pairscan

multiple QTL: CIM, MIM and BIM
Computational Infrastructure for Systems Genetics Analysis
Brian Yandell, UW-Madison

high-throughput analysis of systems data enable biologists & analysts to share tools

UW-Madison: Yandell, Attie, Broman, Kendziorski
Jackson Labs: Churchill
U Groningen: Jansen, Swertz
UC-Denver: Tabakoff
LabKey: Igra

eQTL Tools Seattle SISG: Yandell © 2010

www.stat.wisc.edu/~yandell/statgen
byandell@wisc.edu

- UW-Madison
  - Alan Attie
  - Christina Kendziorski
  - Karl Broman
  - Mark Keller
  - Andrew Broman
  - Aimee Broman
  - YounJeong Choi
  - Elias Chaibub Neto
  - Jee Young Moon
  - John Dawson
  - Ping Wang
  - NIH Grants DK58037, DK66369, GM74244, GM69430, EY18869

- Jackson Labs (HTDAS)
  - Gary Churchill
  - Ricardo Verdugo
  - Keith Sheppard

- UC-Denver (PhenoGen)
  - Boris Tabakoff
  - Cheryl Hornbaker
  - Laura Saba
  - Paula Hoffman

- Labkey Software
  - Mark Igra

- U Groningen (XGA)
  - Rilsert Jansen
  - Morris Swertz
  - Pjotr Pins
  - Danny Arends

- Broad Institute
  - Jill Mesirov
  - Michael Reich
experimental context

- B6 x BTBR obese mouse cross
  - model for diabetes and obesity
  - 500+ mice from intercross (F2)
  - collaboration with Rosetta/Merck
- genotypes
  - 5K SNP Affymetrix mouse chip
  - care in curating genotypes! (map version, errors, …)
- phenotypes
  - clinical phenotypes (>100 / mouse)
  - gene expression traits (>40,000 / mouse / tissue)
  - other molecular phenotypes

how does one filter traits?

- want to reduce to “manageable” set
  - 10/100/1000: depends on needs/tools
  - How many can the biologist handle?
- how can we create such sets?
  - data-driven procedures
    - correlation-based modules
      - Zhang & Horvath 2005 SAGMB, Keller et al. 2008 Genome Res
      - Li et al. 2006 Hum Mol Gen
    - mapping-based focus on genome region
  - function-driven selection with database tools
    - GO, KEGG, etc
    - Incomplete knowledge leads to bias
    - random sample
why build Web eQTL tools?

• common storage/maintainence of data
  – one well-curated copy
  – central repository
  – reduce errors, ensure analysis on same data
• automate commonly used methods
  – biologist gets immediate feedback
  – statistician can focus on new methods
  – codify standard choices

how does one build tools?

• no one solution for all situations
• use existing tools wherever possible
  – new tools take time and care to build!
  – downloaded databases must be updated regularly
• human component is key
  – need informatics expertise
  – need continual dialog with biologists
• build bridges (interfaces) between tools
  – Web interface uses PHP
  – commands are created dynamically for R
• continually rethink & redesign organization
perspectives for building a community where disease data and models are shared

Benefits of wider access to datasets and models:
1- catalyze new insights on disease & methods
2- enable deeper comparison of methods & results

Lessons Learned:
1- need quick feedback between biologists & analysts
2- involve biologists early in development
3- repeated use of pipelines leads to documented learning from experience
   increased rigor in methods

Challenges Ahead:
1- stitching together components as coherent system
2- ramping up to ever larger molecular datasets

Swertz & Jansen (2007)
collaborative portal (LabKey)

systems genetics portal (PhenoGen)

view results (R graphics, GenomeSpace tools)

iterate many times

get data (GEO, Sage)

run pipeline (CLIO, XGAP, HTDAS)

analysis pipeline acts on objects (extends concept of GenePattern)
pipeline is composed of many steps

combine datasets

compare methods

casual model selection choices
in context of larger, unknown network

causal

reactive

correlated

uncorrelated
BxH ApoE-/- chr 2: causal architecture

12 causal calls

hotspot

BxH ApoE-/- causal network for transcription factor Pscdbp

causal trait

work of Elias Chaibub Neto
collaborative portal (LabKey) → systems genetics portal (PhenoGen) → get data (GEO, Sage) → run pipeline (CLIO, XGAP, HTD AS) → view results (R graphics, GenomeSpace tools)

develop analysis methods & algorithms → iterate many times → view results

update periodically → run pipeline

input → pipeline → output

setting s → pipeline

preserve history → pipeline

package → raw code → R&D

check s → pipeline
Model/View/Controller (MVC) software architecture

- isolate domain logic from input and presentation
- permit independent development, testing, maintenance

![Diagram of MVC architecture](image)

EQTTL Tools Seattle SISG: Yandell © 2010
automated R script

```r
library('B6BTBR07')

out <- multtrait(cross.name='B6BTBR07',
                 filename = 'scanone_1214952578.csv',
                 category = 'islet', chr = c(17),
                 threshold.level = 0.05, sex = 'both',)

sink('scanone_1214952578.txt')
print(summary(out))
sink()

bitmap('scanone_1214952578%03d.bmp',
        height = 12, width = 16, res = 72, pointsize = 20)
plot(out, use.cM = TRUE)
dev.off()
```

eQTL Tools Seattle SISG: Yandell © 2010
Inferring Causal Phenotype Networks

Elias Chaibub Neto & Brian S. Yandell

UW-Madison

June 2010
outline

• QTL-driven directed graphs
  – Assume QTLs known, network unknown
  – Infer links (edges) between pairs of phenotypes (nodes)
    • Based on partial correlation
  – Infer causal direction for edges
  – Chaibub et al. (2008 Genetics)
  – Software R/qdg available on CRAN

• Causal graphical models in systems genetics
  – QTLs unknown, network unknown
  – Infer both genetic architecture (QTLs) and pathways (networks)
  – Chaibub et al. (2010 Ann Appl Statist)
  – Software R/qtlNet (www.stat.wisc.edu/~yandell/sysgen/qtlNet)
QTL-driven directed graphs

• See edited slides by Elias Chaibub Neto
  – BIOCOMP 2008 talk
Introduction

- Our objective is to learn metabolic pathways from data.

- We represent these pathways by directed networks composed by transcripts, metabolites and clinical traits.

- These phenotypes are quantitative in nature, and can be analyzed using quantitative genetics techniques.
In particular, we use Quantitative Trait Loci (QTL) mapping methods to identify genomic regions affecting the phenotypes. Since variations in the genotypes (QTLs) cause variations in the phenotypes, but not the other way around, we can unambiguously determine the causal direction

\[ \text{QTL} \Rightarrow \text{phenotype} \]

Knowing that a QTL causally affects a phenotype will allow us to infer causal direction between phenotypes.
PC algorithm

- Causal discovery algorithm developed by Spirtes et al 1993.

- It is composed of two parts:
  1. Infers the skeleton of the causal model.
  2. Partially orient the graph (orient some but not all edges).

- We are only interested in the first part (the “PC skeleton algorithm”). We do not use the PC algorithm to edge orientation (we use the QDG algorithm instead).
Step 1 (PC skeleton algorithm)

Suppose the true network describing the causal relationships between six transcripts is:

\[ y_1 \rightarrow y_2 \rightarrow y_6 \rightarrow y_3 \rightarrow y_5 \rightarrow y_4 \]

The PC-algorithm starts with the complete undirected graph:

\[ y_1 \leftrightarrow y_2 \leftrightarrow y_6 \leftrightarrow y_3 \leftrightarrow y_5 \leftrightarrow y_4 \]

and progressively eliminates edges based on conditional independence tests.
Step 1 (PC skeleton algorithm)

The algorithm performs several rounds of conditional independence tests of increasing order.

It starts with all zero order tests, then performs all first order, second order . . .

- Notation: $\perp \perp \equiv$ independence. We read $i \perp \perp j \mid k$ as $i$ is conditionally independent from $j$ given $k$.

- Remark: in the Gaussian case zero partial correlation implies conditional independence, thus

  \[
  i \perp j \mid k \iff \text{cor}(i, j \mid k) = 0 \implies \text{drop } (i, j) \text{ edge}
  \]
Example (order 1)

\[ A(1) \setminus 2 = \{2, 4, 5, 6\} \]

\[ 1 \perp 3 \mid 2 \]

\[ \text{vs} \]

\[ 1 \not\perp 3 \mid 2 \]

\[ y_2 \text{ d-separates } y_1 \text{ from } y_3 \]

\[ 1 \perp 3 \mid 2 \]

\[ \text{drop edge} \]

\[ \text{move to next edge} \]
Example (order 1)

The algorithm then moves to second order conditional independence tests.

After all first order conditional independence tests.
Example (order 2)

\[ A(1) \setminus 4 = \{ 2, 5, 6 \} \]

\[ 1 \perp 4 \mid 2, 5 \]

vs

\[ 1 \not\perp 4 \mid 2, 5 \]
Example (order 2)

\[ A(1) \setminus 4 = \{2, 5, 6\} \]

1 \( \perp \perp \) 4 \( | \) 2, 5  
vs  
1 \( \not\perp \not\perp \) 4 \( | \) 2, 5

\((y_2, y_5)\) d-separate \(y_1\) from \(y_4\)

\[ 1 \perp \perp 4 \mid 2, 5 \]

drop edge
move to next edge
After all second order conditional independence tests.

The algorithm than moves to third order, fourth order . . .

It stops when for each pair \((i, j)\) the cardinality of

\[ A(i) \setminus j \]

is smaller then the order of the algorithm.
Consider two traits $y_1$ and $y_2$. Our problem is to decide between models:

\[ M_1 : \xrightarrow{\text{y}_1} \xrightarrow{\text{y}_2} \quad M_2 : \xrightarrow{\text{y}_1} \xleftarrow{\text{y}_2} \]

Problem: the above models are likelihood equivalent,

\[ f(y_1)f(y_2 \mid y_1) = f(y_1, y_2) = f(y_2)f(y_1 \mid y_2). \]
However, models

\[
\begin{align*}
q_{11} & \rightarrow y_1 & q_{21} & \rightarrow y_2 & q_{11} & \rightarrow y_1 & q_{21} & \rightarrow y_2 \\
q_{1k} & \rightarrow \ldots & y_1 & \rightarrow y_2 & q_{2l} & \rightarrow \ldots & q_{1k} & \rightarrow \ldots \\
\end{align*}
\]

are *not* likelihood equivalent because

\[
\begin{align*}
f(q_1)f(y_1 \mid q_1)f(y_2 \mid y_1, q_2)f(q_2) & \\
\neq & \\
f(q_2)f(y_2 \mid q_2)f(y_1 \mid y_2, q_1)f(q_1)
\end{align*}
\]
We perform model selection using a direction LOD score

$$LOD = \log_{10} \left\{ \frac{\prod_{i=1}^{n} f(y_{1i} \mid q_{1i}) f(y_{2i} \mid y_{1i}, q_{2i})}{\prod_{i=1}^{n} f(y_{2i} \mid q_{2i}) f(y_{1i} \mid y_{2i}, q_{1i})} \right\}$$

where $f()$ represents the predictive density, that is, the sampling model with parameters replaced by the corresponding maximum likelihood estimates.
QDG stands for QTL-directed dependency graph. The QDG algorithm is composed of 7 steps:

1. Get the causal skeleton (with the PC skeleton algorithm).
2. Use QTLs to orient the edges in the skeleton.
3. Choose a random ordering of edges, and
4. Recompute orientations incorporating causal phenotypes in the models (update the causal model according to changes in directions).
5. Repeat 4 iteratively until no more edges change direction (the resulting graph is one solution).
6. Repeat steps 3, 4, and 5 many times and store all different solutions.
7. Score all solutions and select the graph with best score.
Step 2

Now suppose that for each transcript we have a set of e-QTLs

Given the QTLs we can distinguish causal direction:

\[ q_1 \rightarrow y_1 \rightarrow y_2 \leftarrow q_2 \]

\[ q_1 \rightarrow y_1 \leftarrow y_2 \rightarrow q_2 \]

\[ q_1 \rightarrow y_1 \leftarrow y_2 \rightarrow q_2 \]

\[ q_5 \rightarrow y_5 \rightarrow y_6 \leftarrow q_6 \]

\[ q_5 \rightarrow y_5 \leftarrow y_6 \rightarrow q_6 \]
Steps 4 and 5 (first iteration)
Steps 4 and 5 (first iteration)
Suppose the updated causal model after the first iteration ($DG_1$) is

Since some arrows changed direction ($DG_1 \neq DG_0$), the algorithm goes for another round of re-computations.
Directing edges without QTLs

- In general, we need to have at least one QTL per pair of phenotypes to infer causal direction.

- In some situations, however, we may be able to infer causal direction for a pair of phenotypes without QTLs. Eg.

- So both QTLs and phenotypes play important roles in the orientation process.
We cannot infer direction when the phenotypes have exactly the same set of QTLs and causal phenotypes

Since

\[ f(y_1 \mid y_3, q) f(y_2 \mid y_1, y_3, q) = f(y_1 \mid y_2, y_3, q) f(y_2 \mid y_3, q) \]
Our simulations showed good performance with toy cyclic graphs, though.

The spurious edges in graph (c) were detected at low rates.

QDG approach cannot detect reciprocal interactions. In graph (c) it orients the edge 2→5 in the direction with higher strength.
We generated 100 data sets according to this network.

Parameters were chosen in a range close to values estimated from real data.

<table>
<thead>
<tr>
<th>n</th>
<th>60</th>
<th>300</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDR</td>
<td>94.53</td>
<td>95.18</td>
<td>91.22</td>
</tr>
<tr>
<td>TPR</td>
<td>52.07</td>
<td>87.33</td>
<td>93.64</td>
</tr>
<tr>
<td>CD</td>
<td>83.65</td>
<td>98.58</td>
<td>99.63</td>
</tr>
</tbody>
</table>

100 nodes, 107 edges

2 or 3 QTLs per phenotype (not shown)
We constructed a network from metabolites and transcripts involved in liver metabolism.

We validated this network with *in vitro* experiments (Ferrara et al 2008). Four out of six predictions were confirmed.
The *qdg* R package is available at CRAN.

References:

Acknowledgements

Co-authors:
- Alan D. Attie
- Brian S. Yandell
- Christine T. Ferrara

Funding:
- CNPq Brazil
- NIH grants DK66369, DK58037 and DK06639
To break the connections (brk) that affect direction of an edge, we permute the corresponding pair of nodes (and their common covariates) as a block.

- In panel (a) we permute \((y_1, y_2, x)\) as a block breaking the connections with \(z, q_1\) and \(q_2\);
- In panel (b) we incorrectly keep \(z\) in the permutation block.
A strong QTL directly affecting an upstream trait may also be (incorrectly) detected as a QTL for a downstream phenotype.

To resolve this situation we apply a generalization of Schadt et al. 2005 allowing for multiple QTLs.

Model (a) supports both traits being directly affected by the common QTL $q$. Model (b) implies that $q$ directly affects $y_1$ but should not be included as a QTL of phenotype $y_2$. Model (c) supports the reverse situation.
causal graphical models in systems genetics


• Related references

• Jointly infer unknowns of interest
  – genetic architecture
  – causal network
Basic idea of QTLnet

• Genetic architecture given causal network
  – Trait y depends on parents pa(y) in network
  – QTL for y found conditional on pa(y)
    • Parents pa(y) are interacting covariates for QTL scan

• Causal network given genetic architecture
  – Build (adjust) causal network given QTL
MCMC for QTLnet

• Propose new causal network with simple changes to current network
  – Change edge direction
  – Add or drop edge

• Find any new genetic architectures (QTLs)
  – Update phenotypes whose parents pa(y) change in new network

• Compute likelihood for new network and QTL

• Accept or reject new network and QTL
  – Usual Metropolis-Hastings idea
Future work

• Incorporate latent variables
  – Aten et al. Horvath (2008 *BMC Sys Biol*)

• Allow for prior information about network

• Improve algorithm efficiency
  – Ramp up to 1000s of phenotypes

• Extend to outbred crosses, humans