QTL Mapping I: Overview and using Inbred Lines
Key idea: Looking for marker-trait associations in collections of relatives

If (say) the mean trait value for marker genotype MM is statistically different from that for genotype mm, then the M/m marker is linked to a QTL

One can use a random collection of such markers spanning a genome (a genomic scan) to search for QTLs
Experimental Design: Crosses

P1 x P2
F1
F1 x F1
F2
F2 design

B1
Backcross design

F1

B2
Backcross design

Fk
Advanced intercross Design (AIC, AICk)

RILs = Recombinant inbred lines (selfed F1s)
Experimental Designs: Marker Analysis

Single marker analysis

Flanking marker analysis (interval mapping)

Composite interval mapping
  Interval mapping plus additional markers

Multipoint mapping
  Uses all markers on a chromosome simultaneously
Conditional Probabilities of QTL Genotypes

The basic building block for all QTL methods is
$\Pr(Q_k | M_j)$ --- the probability of QTL genotype
$Q_k$ given the marker genotype is $M_j$.

$$\Pr(Q_k | M_j) = \frac{\Pr(Q_kM_j)}{\Pr(M_j)}$$

Consider a QTL linked to a marker (recombination Fraction = $c$). Cross MMQQ x mmqq. In the F1, all
gametes are MQ and mq

In the F2, $\text{freq}(MQ) = \text{freq}(mq) = (1-c)/2,$
$\text{freq}(mQ) = \text{freq}(Mq) = c/2$
Hence, $\Pr(MMQQ) = \Pr(MQ)\Pr(MQ) = (1-c)^2/4$

$\Pr(MMQq) = 2\Pr(MQ)\Pr(Mq) = 2c(1-c)/4$

$\Pr(MMqq) = \Pr(Mq)\Pr(Mq) = c^2/4$

Why the 2? MQ from father, Mq from mother, OR MQ from mother, Mq from father

Since $\Pr(MM) = 1/4$, the conditional probabilities become

$\Pr(QQ \mid MM) = \Pr(MMQQ)/\Pr(MM) = (1-c)^2$

$\Pr(Qq \mid MM) = \Pr(MMQq)/\Pr(MM) = 2c(1-c)$

$\Pr(qq \mid MM) = \Pr(MMqq)/\Pr(MM) = c^2$

How do we use these?
Expected Marker Means

The expected trait mean for marker genotype $M_j$ is just

$$\mu_{M_j} = \sum_{k=1}^{N} \mu_{Q_k} \Pr(Q_k | M_j)$$

For example, if $QQ = 2a$, $Qq = a(1+k)$, $qq = 0$, then in the F2 of an MMQQ/mmqq cross,

$$\frac{\mu_{MM} - \mu_{mm}}{2} = a(1 - 2c)$$

• If the trait mean is significantly different for the genotypes at a marker locus, it is linked to a QTL
• A small MM-mm difference could be (i) a tightly-linked QTL of small effect or (ii) loose linkage to a large QTL
2 Marker loci

Suppose the cross is $M_1M_1QQM_2M_2 \times m_1m_1qqm_2m_2$

Genetic map

No interference: $c_{12} = c_1 + c_2 - 2c_1c_2$

Complete interference: $c_{12} = c_1 + c_2$

In F2, $Pr(M_1QM_2) = (1-c_1)(1-c_2)$

$Pr(M_1Qm_2) = (1-c_1)c_2$  \hspace{1cm}  $Pr(m_1QM_2) = (1-c_1)c_2$

Likewise, $Pr(M_1M_2) = 1-c_{12} = 1-c_1 + c_2$
A little bookkeeping gives

\[
\Pr(QQ \mid M_1 M_1 M_2 M_2) = \frac{(1 - c_1)^2(1 - c_2)^2}{(1 - c_{12})^2}
\]

\[
\Pr(Qq \mid M_1 M_1 M_2 M_2) = \frac{2c_1 c_2(1 - c_1)(1 - c_2)}{(1 - c_{12})^2}
\]

\[
\Pr(qq \mid M_1 M_1 M_2 M_2) = \frac{c_1^2 c_2^2}{(1 - c_{12})^2}
\]

\[
\frac{\mu_{M_1 M_1 M_2 M_2} - \mu_{m_1 m_1 m_2 m_2}}{2} = a - \left(\frac{1 - c_1 - c_2}{1 - c_1 - c_2 + 2c_1 c_2}\right)
\]

\[
\simeq a (1 - 2c_1 c_2)
\]

This is essentially a for even modest linkage
\[ c_1 = \frac{1}{2} \left( 1 - \frac{\mu_{M_1 M_1} - \mu_{m_1 m_1}}{2a} \right) \]

\[ \approx \frac{1}{2} \left( 1 - \frac{\mu_{M_1 M_1} - \mu_{m_1 m_1}}{\mu_{M_1 M_1 M_2 M_2} - \mu_{m_1 m_1 m_2 m_2}} \right) \]

Hence, \( a \) and \( c \) can be estimated from the mean values of flanking marker genotypes.
Linear Models for QTL Detection

The use of differences in the mean trait value for different marker genotypes to detect a QTL and estimate its effects is a use of linear models.

One-way ANOVA.

Value of trait in kth individual of marker genotype type i

\[ z_{ik} = \mu + b_i + e_{ik} \]

Effect of marker genotype i on trait value
\[
 z_{ik} = \mu + b_i + e_{ik}
 \]

**Detection:** a QTL is linked to the marker if at least one of the $b_i$ is significantly different from zero

**Estimation:** (QTL effect and position): This requires relating the $b_i$ to the QTL effects and map position
Detecting epistasis

One major advantage of linear models is their flexibility. To test for epistasis between two QTLs, used an ANOVA with an interaction term

\[ z = \mu + a_i + b_k + d_{ik} + e \]

- Effect from marker genotype at first marker set (can be > 1 loci)
- Effect from marker genotype at second marker set
- Interaction between marker genotypes i in 1st marker set and k in 2nd marker set
Detecting epistasis

\[ z = \mu + a_i + b_k + d_{ik} + e \]

- At least one of the \( a_i \) significantly different from 0
  ---- QTL linked to first marker set
- At least one of the \( b_k \) significantly different from 0
  ---- QTL linked to second marker set
- At least one of the \( d_{ik} \) significantly different from 0
  ---- interactions between QTL in sets 1 and two

Problem: Huge number of potential interaction terms
(order \( m^2 \), where \( m \) = number of markers)
Maximum Likelihood Methods

ML methods use the entire distribution of the data, not just the marker genotype means.

More powerful than linear models, but not as flexible in extending solutions (new analysis required for each model).

Basic likelihood function:

\[ \ell(z \mid M_j) = \sum_{k=1}^{N} \varphi(z, \mu_{Q_k}, \sigma^2) \Pr(Q_k \mid M_j) \]

This is a mixture model
Maximum Likelihood Methods

\[ \ell(z \mid M_j) = \sum_{k=1}^{N} \varphi(z, \mu_{Q_k}, \sigma^2) \Pr(Q_k \mid M_j) \]

Sum over the N possible linked QTL genotypes

Distribution of trait value given QTL genotype is k is normal with mean \( \mu_{Q_k} \). (QTL effects enter here)

Probability of QTL genotype k given marker genotype j --- genetic map and linkage phase entire here
ML methods combine both detection and estimation of QTL effects/position.

Test for a linked QTL given from the LR test

\[
LR = -2 \ln \frac{\max \ell_r(z)}{\max \ell(z)}
\]

The LR score is often plotted by trying different locations for the QTL (i.e., values of \(c\)) and computing a LOD score for each

\[
LOD(c) = -\log_{10} \left[ \frac{\max \ell_r(z)}{\max \ell(z, c)} \right] = \frac{LR(c)}{2 \ln 10} \approx \frac{LR(c)}{4.61}
\]
A typical QTL map from a likelihood analysis:

- Estimated QTL location
- Support interval
- Significance
- Threshold
Consider interval mapping using the markers $i$ and $i+1$. QTLs linked to these markers, but outside this interval, can contribute (falsely) to estimation of QTL position and effect.

Now suppose we also add the two markers flanking the interval ($i-1$ and $i+2$).
Inclusion of markers $i-1$ and $i+2$ fully account for any linked QTLs to the left of $i-1$ and the right of $i+2$.

Interval mapping + marker cofactors is called **Composite Interval Mapping (CIM)**.

CIM works by adding an additional term to the linear model, 
\[
\sum_{k \neq i, i+1} b_k x_{kj}
\]

CIM also (potentially) includes unlinked markers to account for QTL on other chromosomes.
Power and Precision

While modest sample sizes are sufficient to detect a QTL of modest effect (power), large sample sizes are required to map it with some precision.

With 200-300 $F_2$, a QTL accounting for 5% of total variation can be mapped to a 40$cM$ interval.

Over 10,000 $F_2$ individuals are required to map this QTL to a 1$cM$ interval.
Power and Repeatability: The Beavis Effect

QTLs with low power of detection tend to have their effects overestimated, often very dramatically.

As power of detection increases, the overestimation of detected QTLs becomes far less serious.

This is often called the Beavis Effect, after Bill Beavis who first noticed this in simulation studies.
Model selection

• With (say) 300 markers, we have (potentially) 300 single-marker terms and $300 \times 299/2 = 44,850$ epistatic terms
  - Hence, a model with up to $p = 45,150$ possible parameters
  - $2^p$ possible submodels $= 10^{13.600}$ ouch!

• The issue of Model selection becomes very important.

• How do we find the best model?
  - Stepwise regression approaches
    • Forward selection (add terms one at a time)
    • Backwards selection (delete terms one at a time)
  - Try all models, assess best fit
  - Mixed-model approaches (SSVS)
Model Selection

Model Selection: Use some criteria to choose among a number of candidate models. Weight goodness-of-fit (L, value of the likelihood at the MLEs) vs. number of estimated parameters (k)

\[ \text{AIC} = \text{Akaike's information criterion} \]
\[ \text{AIC} = 2k - 2 \ln(L) \]

\[ \text{BIC} = \text{Bayesian information criterion (Schwarz criterion)} \]
\[ \text{BIC} = k \ln(n)/n - 2 \ln(L)/n \]
BIC penalizes free parameters more strongly than AIC

Other measures. Smaller is better
Model averaging

Model averaging: Generate a composite model by weighting (averaging) the various models, using AIC, BIC, or other

Idea: Perhaps no “best” model, but several models all extremely close. Better to report this “distribution” rather than the best one

One approach is to average the coefficients on the “best-fitting” models using some scheme to return a composite model
Stochastic search variable selection (SSVS)

- A Bayesian approach approach to search through a space of possible models
  - Fit the model $y_i = \mu + x_i^T \beta + e$, including ALL possible covariates of the full model, $X = (x_i^T \cdots x_n^T)$
  - Idea: Assume model parameters fall into two classes: those with values very near zero and those with larger values
  - We use the latent (unobservable) variable $\gamma_i$, which determines into which class the covariate falls
  - $\beta_i \sim (1- \gamma_i)N(0,\tau_i^2) + \gamma_i N(0, c_i^2 \tau_i^2)$
  - $\tau_i^2$ small, hence values near zero
  - $c_i^2 \tau_i^2$ large, hence values can deviate substantially from zero
  - Posterior probability of $(\gamma_i)$ is the probability that the parameter is included in the final model
$\beta_i \sim (1- \gamma_i)N(0, \tau_i^2) + \gamma_i N(0, c_i^2 \tau_i^2)$

While $\tau$ and $c$ can vary over covariates (model variables), typically they are assigned the same values over all $i$ (e.g., $\tau_i^2 = 0.001$, $c_i^2 \tau_i^2 = 10$)

Let $\gamma$ be the vector of indicator random variables, with a value of one if in the model, zero otherwise

Given a current $\gamma$ vector, the conditional prior of the $\beta$ values is MVN with mean zero and covariance matrix $D_{\gamma} R D_{\gamma}$

- Where $R$ is the prior correlation matrix, either taken as $R = I$ or $R = (X^TX)^{-1}$
- And $D$ is a diagonal matrix whose ith element is 1 if $\gamma_i = 0$ and $c_i$ if $\gamma_i = 1$. 
• The idea is that with a current estimate of $\mu$, $\sigma_e^2$, and $\gamma$ in hand, we can update $\beta$, which is drawn from a MVN distribution
  - Mean = $(X^TX + \sigma_e^2(D_{\gamma}RD_{\gamma})^{-1})^{-1}X^T(y - \mu I)$
  - Variance = $(X^TX + \sigma_e^2(D_{\gamma}RD_{\gamma})^{-1})^{-1}$

• With an updated $\beta$ vector in hand, we update the $\gamma$ vector

$$
Pr(\gamma_i | \gamma(-i), \mu, \beta, \sigma_e^2) = \frac{p(\beta | \gamma(-i), \gamma_i = 1) Pr(\gamma_i = 1)}{p(\beta | \gamma(-i), \gamma_i = 1) Pr(\gamma_i = 1) + p(\beta | \gamma(-i), \gamma_i = 0) Pr(\gamma_i = 0)}
$$

Usually, the prior on the $\gamma_i$ is independent of $i$ and taken to be $p_i = 1/2$
Supersaturated Models

A problem with many QTL approaches is that there are far more parameters (p) to estimate than there are independent samples (n).

Case in point: epistasis

Such supersaturated models arise commonly in Genomics. How do we deal with them?

Again, an approach like SSVS, where all parameters are included, but some are shrunk back (regressed) towards zero by assigning them a very small posterior variance.
Shrinkage estimators

Shrinkage estimates: Rather than adding interaction terms one at a time, a shrinkage method starts with all interactions included, and then shrinks most back to zero.

Under a Bayesian analysis, any effect is random. One can assume the effect for (say) interaction \( ij \) is drawn from a normal with mean zero and variance \( \sigma^2_{ij} \)

Further, the interaction-specific variances are themselves random variables drawn from a hyperparameter distribution, such as an inverse chi-square.

One then estimates the hyperparameters and uses these to predict the variances, with effects with small variances shrinking back to zero, and effects with large variances remaining in the model.
Key ideas in QTL mapping

Look for marker-trait associations

• Many difference crossing designs (F2, BC, RIL)
• Many difference methods of analysis: Linear models, MLE, Bayesian
• Most studies UNDERPOWERED, esp. for “fine” mapping
What is a “QTL”

• A detected “QTL” in a mapping experiment is a region of a chromosome detected by linkage.
• Usually large (typically 10-40 cM)
• When further examined, most “large” QTLs turn out to be a linked collection of locations with increasingly smaller effects
• The more one localizes, the more subregions that are found, and the smaller their effect