7 multiple QTL implementation

- maximum likelihood using EM
  - composite interval mapping (Zeng)
  - multiple interval mapping (Kao Zeng)
  - sequential testing to search model space
- multiple imputation (Sen Churchill)
  - Bayesian log posterior odds
  - sequential testing and pairwise plots to search
- MCMC (Satagopan Yandell Newton)
  - Bayes factors and marginal posteriors
  - Markov chain sampling to search model space

multiple QTL likelihood

- likelihood is mixture over unknown QTL
  - likelihood = product of sum of products
  - now have multiple QTL
    - \( Q = (Q_1, Q_2, \ldots, Q_m) \)
    - \( \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m) \)
    - \( \theta = (\mu, \theta_1, \theta_2, \ldots, \theta_m, \sigma^2) \) plus interactions…

\[
L(\theta, \lambda \mid Y, X) = \text{pr}(Y \mid X, \theta, \lambda) \\
= \prod_i \text{pr}(Y_i \mid X_i, \theta, \lambda) \\
= \prod_i \sum_Q \text{pr}(Q \mid X_i, \lambda) \text{pr}(Y_i \mid Q, \theta)
\]
maximum likelihood (ML) idea

• pick QTL loci $\lambda$
  – scan whole genome for single QTL with others fixed
  – jointly scan small part of genome for two QTL
• find ML estimates of gene action $\theta$ given $\lambda$
  – maximum likelihood at peak of likelihood
  – slope is weighted average using posteriors for $Q$
  – estimate depends on $\theta$ -- need to iterate (again!)

\[
\frac{dL(\theta, \lambda|Y, X)}{d\theta} = \sum_{i,Q} \text{pr}(Q|Y_i, X_i, \theta, \lambda) \frac{d\log(\text{pr}(Y_i|Q, \theta))}{d\theta}
\]

\[
\text{pr}(Q|Y_i, X_i, \theta, \lambda) = \frac{\text{pr}(Q|X_i, \lambda)\text{pr}(Y_i|Q, \theta)}{\text{pr}(Y_i|X_i, \theta, \lambda)}
\]

EM method for interval mapping

• E-step: estimate posterior recombination
  – $P_{Qi} = \text{pr}(Q | Y_p, X_p, \theta, \lambda)$
  – individual $i$, genotype $Q = (Q_1, Q_2, \ldots, Q_m)$
  – depends on multiple QTL loci $\lambda$ and effects $\theta$
• M-steps: maximize likelihood for $\theta$
  – technical point: caution on parallel updates
    • ECM: conditional maximization depending
    • update in series, based on what has been done already
  – depends on $P_{Qi}$
  – see Kao Zeng papers for details (hard problem!)

\[
0 = \sum_{i,Q} P_{Qi} \frac{d\log(\text{pr}(Y_i|Q, \theta))}{d\theta}
\]
MIM search for QTL

- initial model: use CIM or marker regression
- refine model: simultaneous search
  - iteratively scan QTL to improve each locus
  - fix all other loci temporarily
- add QTL or epistatic interactions
  - forward selection: scan genome for new QTL
  - check all pairs of significant QTL for epistasis
  - scan genome for interactions with significant QTL
- drop QTL that may now be non-significant
  - backward elimination

CIM model

- CIM (or MQM) approximates MIM
  - scan for one QTL while fixing all others
  - use markers as "cofactors" for other QTL
- art to selection of cofactors
  - pick by stepwise marker regression
    - with model selection criterion
  - eliminate markers "too close" to current interval
    - wide window screens out loosely linked QTL
    - narrow window washes out effect of desired QTL
    - no firm criteria--depends on effects, marker spacing
**CIM idea**

![Diagram showing QTL intervals and windows.]

**examples of MIM**

- Liu et al. (1996); Zeng et al. (2000)
  - gonad shape on *Drosophila* spp.
  - reciprocal backcross, 500 individuals per cross
  - 19 QTL
- Weber et al. (1999)
  - *Drosophila* wing shape
  - 700 RI individuals
  - 19 QTL plus 10 interactions
phenotype: outline of posterior lobe
(Liu et al. 1996)

Figure 1—Posterior lobe outlines from a sample of five isofemale lines from each of four Drosophila species. Additional information about the lines is provided in Table 1.

principal components

Figure 2—A plot of the first two principal components of the Fourier coefficients from posterior lobe outlines. Many individuals from each of the five geographic clusters are represented. Each point represents an average of scores from the left and right sides of an individual except for 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
sample shapes for PC1
in parents, F1 and backcrosses

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"
pairs plot: CIM, MIM and pairscan

MIM effects for *Dm*
Bayesian model posterior

- augment data \((Y,X)\) with unknowns \(Q\)
  - study unknowns \((\theta, \lambda, Q)\) given data \((Y,X)\)
  - \(Q \sim p(Q \mid Y, X, \theta, \lambda)\)
- sample genotypes \(Q\) for every individual at \(m\) QTL
  - and multiple loci \(\lambda\) and multiple genetic effects and epistasis \(\theta\)
- study properties of posterior
  - use priors that are independent between QTL
  - draw samples from posterior in some clever way
    - multiple imputation or MCMC

\[
p(\theta, \lambda, Q \mid Y, X) = \frac{p(Q \mid X, \lambda) p(Y \mid Q, \theta) p(\lambda \mid X) p(\theta)}{p(Y \mid X)}
\]

2 QTL: EM versus multiple imputation for \(Dm\)
MCMC idea for QTLs

• 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
• update $m$-QTL model components from full conditionals
  – update effects $\theta$ given genotypes & traits
  – update locus $\lambda$ given genotypes & marker map
  – update genotypes $Q$ given traits, marker map, locus & effects

\[
(\lambda, Q, \theta, m) \sim \text{pr}(\lambda, Q, \theta, m | Y, X)
\]
\[
(\lambda, Q, \theta, m_1) \rightarrow (\lambda, Q, \theta, m_2) \rightarrow \cdots \rightarrow (\lambda, Q, \theta, m_N)
\]

sample from full conditionals
for model with $m$ QTL

• hard to sample from joint posterior
  – $\text{pr}(\lambda, Q, \theta | Y, X) = \text{pr}(\theta)\text{pr}(\lambda)\text{pr}(Q|X, \lambda)\text{pr}(Y|Q, \theta) / \text{constant}$
• easy to sample parameters from full conditionals
  – full conditional for genetic effects
    • $\text{pr}(\theta | Y, X, \lambda, Q) = \text{pr}(\theta | Y, Q) = \text{pr}(\theta)\text{pr}(Y|Q, \theta) / \text{constant}$
  – full conditional for QTL locus
    • $\text{pr}(\lambda | Y, X, \theta, Q) = \text{pr}(\lambda | X, Q) = \text{pr}(\lambda)\text{pr}(Q|X, \lambda) / \text{constant}$
  – full conditional for QTL genotypes
    • $\text{pr}(Q | Y, X, \lambda, \theta) = \text{pr}(Q | X, \lambda)\text{pr}(Y|Q, \theta) / \text{constant}$
Bayesian interval mapping

- sample missing genotypes $Q$
- decouple effects $\theta$ from QTL $\lambda$
- but $Q$ depends on $(\theta, \lambda)$ and vice versa
- also need to specify priors

\[
\begin{align*}
\lambda &\sim \frac{\text{pr}(Q \mid X, \lambda) \text{pr}(\lambda \mid X)}{\text{pr}(Q \mid X)} \\
Q &\sim \text{pr}(Q \mid Y_1, X, \theta, \lambda) \\
\theta &\sim \frac{\text{pr}(Y \mid Q, \theta) \text{pr}(\theta)}{\text{pr}(Y \mid Q)}
\end{align*}
\]

MCMC diagnostics for $Dm$ shape

- $m \sim \text{Poisson}(15)$ prior on number of QTL
- Bayesian LOD (log posterior density)
- Heritability
- 5% burnin
- 1,000,000 samples – every 1000\textsuperscript{th} recorded
- note stable mean
MCMC sampled loci

- markers as blue lines
  - horizontal jittering
- note denser regions
  - 10-11 broad regions
- jointly sampling
  - 15-30 QTL at once

MCMC model selection

- \( m \) = number of QTL
  - prior: Poisson(15)
    - rescaled in blue
    - posterior: mean 22.4
    - Bayes factor increases
  - pattern across genome
    - prior depends on \( m \) and length of chromosomes
    - posterior mode: \( m = 20 \)
    - Bayes factor favors
      - \( m = 24 \)
      - 3*1, 8*2, 13*3
MCMC model selection restricted to “better models”

- models with minimum
  - \( m \geq 24 \)
  - pattern \( \geq 3*1, 8*2, 13*3 \)
- note uncertainty in BF
  - estimate \( \pm 2 \) SE
- mode is chosen pattern
  - \(~14\% \) of samples
- BF similar to more complicated patterns
  - parsimony: simpler model
  - 2SE intervals overlap

MCMC loci and effects

- model averaging
  - over all models
  - 1000 samples
- histogram of loci
  - marginal posteriors
  - superimposed on genome
  - 12 peaks identified
- scatterplot: loci & effects
  - smoothed mean \( \pm 2 \) SE
Brassica napus data

- 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

Brassica 4- & 8-week data

summaries of raw data
joint scatter plots
(identity line)
separate histograms
Brassica 8-week data locus MCMC with $m=2$

4-week vs 8-week vernalization

- **4-week vernalization**
  - longer time to flower
  - larger LOD at 40cM
  - modest LOD at 80cM
  - loci well determined

- **8-week vernalization**
  - shorter time to flower
  - larger LOD at 80cM
  - modest LOD at 40cM
  - loci poorly determined

<table>
<thead>
<tr>
<th>cM</th>
<th>add 4-week</th>
<th>add 8-week</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>0.30</td>
<td>0.06</td>
</tr>
<tr>
<td>80</td>
<td>0.16</td>
<td>0.13</td>
</tr>
</tbody>
</table>
Brassica credible regions

4-week

8-week

reversible jump MCMC

action steps: draw one of three choices

- update \( m \)-QTL model with probability \( 1 - b(m+1) - d(m) \)
  - update current model using full conditionals
  - sample \( m \) QTL loci, effects, and genotypes
- add a locus with probability \( b(m+1) \)
  - 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 \( d(m) \)
  - propose dropping one of existing loci
  - decide whether to accept the “death” of locus
sampling the number of QTL

• use reversible jump MCMC to change $m$
  – bookkeeping helps in comparing models
  – adjust to change of variables between models
  – Green (1995); Richardson Green (1997)
  – other approaches out there these days…

• think model selection in multiple regression
  – but regressors (QT genotypes) are unknown
  – linked loci = collinear regressors = correlated effects
  – consider additive effects with coding $Q_{ij} = -1, 0, 1$

$$\theta_{ijQ} = \alpha_j (Q_{ij} - \bar{Q}_j)$$

Model Selection in Regression

• consider known genotypes ($Q$)
  – models with 1 or 2 QTL at known loci

• jump between 1-QTL and 2-QTL models
  – adjust posteriors when model changes
  – due to collinearity of QTL genotypes

$m = 1 : Y_i = \mu + \alpha (Q_{i1} - \bar{Q}_1) + e_i$

$m = 2 : Y_i = \mu + \alpha_1 (Q_{i1} - \bar{Q}_1) + \alpha_2 (Q_{i1} - \bar{Q}_1) + e_i$
collinear QTL = correlated effects

4-week

8-week

cor = -0.81

cor = -0.7

• linked QTL: collinear genotypes & correlated effect estimates
  – sum of linked effects usually well determined
• which QTL to go after in breeding, genome walking?

Geometry of Reversible Jump

Move Between Models

Reversible Jump Sequence

c21 = 0.7

m=2

m=1
QT additive Reversible Jump

a short sequence

first 1000 with m<3

a complicated simulation

• 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%

<table>
<thead>
<tr>
<th>QTL</th>
<th>chr</th>
<th>loci</th>
<th>effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>11</td>
<td>-3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>50</td>
<td>-5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>62</td>
<td>+2</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>107</td>
<td>-3</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>152</td>
<td>+3</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>32</td>
<td>-4</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>54</td>
<td>+1</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>195</td>
<td>+2</td>
</tr>
</tbody>
</table>
loci pattern across genome

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

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>m1</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>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>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>377</td>
<td></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>1</td>
<td>0</td>
<td>218</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>218</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>198</td>
<td></td>
</tr>
</tbody>
</table>

Bmapqtl: our RJ-MCMC software

- www.stat.wisc.edu/~yandell/qtl/software/Bmapqtl
  - module using QtlCart format
  - compiled in C for Windows/NT
  - extensions in progress
  - R post-processing graphics
    - library(bim) is cross-compatible with library(qtl)
- Bayes factor and reversible jump MCMC computation
- enhances MCMCQTL and revjump software
  - initially designed by JM Satagopan (1996)
  - major revision and extension by PJ Gaffney (2001)
    - whole genome
    - multivariate update of effects; long range position updates
    - substantial improvements in speed, efficiency
    - pre-burnin: initial prior number of QTL very large
**B. napus** 8-week vernalization whole genome study

- 108 plants from double haploid
  - similar genetics to backcross: follow 1 gamete
  - parents are Major (biennial) and Stellar (annual)
- 300 markers across genome
  - 19 chromosomes
  - average 6cM between markers
    - median 3.8cM, max 34cM
  - 83% markers genotyped
- phenotype is days to flowering
  - after 8 weeks of vernalization (cooling)
  - Stellar parent requires vernalization to flower

---

Markov chain Monte Carlo sequence

burnin (sets up chain)
mcmc sequence

number of QTL
environmental variance
$h^2 = \text{heritability}$
(genetic/total variance)
LOD = likelihood
MCMC sampled loci

subset of chromosomes N2, N3, N16

points jittered for view
blue lines at markers

note concentration on chromosome N2

Bayesian model assessment

row 1: # QTL
row 2: pattern

col 1: posterior
col 2: Bayes factor
note error bars on bf

evidence suggests 4-5 QTL
N2(2-3), N3, N16
Bayesian estimates of loci & effects

- histogram of loci
- blue line is density
- red lines at estimates

- estimate additive effects (red circles)
- grey points sampled from posterior
- blue line is cubic spline
- dashed line for 2 SD

Bayesian model diagnostics

- pattern: N2(2), N3, N16
- col 1: density
- col 2: boxplots by m

- environmental variance $\sigma^2 = .008$, $\sigma = .09$
- heritability $h^2 = 52\%$
- LOD = 16 (highly significant)
- but note change with m