The University of Newcastle

Kerrie Mengersen

Introduction to
Bayesian Methods for
QTL Analysis - 4

Armidale 2004
Case Studies

This discussion is based on:


• Baker and Mengersen (2002) Central censoring and selective genotyping, submitted to JABES.

• Various papers as referenced
Leading up to Bayesian methods

- Lander and Botstein 89; Knapp et al 89: Early methods based on assumption of at most one QTL contributing to a trait. Obvious deficiencies with many contributing QTLs.
- Knapp 91; Haley and Knott 92; Martinez and Curnow 92: approximate methods to map several QTLs.
- Jansen 92: general mixture model for multiple QTL.
- Cowen 89; Stam 1991: Multiple regression methods.
- Jansen 93; Zheng 93, 94; Jansen and Stam 94; Jansen 94: Hybrid methods combining one-QTL mixture model (L&B 1989) with multiple regression, using covariates to correct for neighbouring QTLs and reduce error variation. Output point estimates for number, location and effects of QTLs, approximate critical values for significance tests, inference on parameters via bootstrapping.
- Wu and Li 94, 96: Regression approach to joint mapping of QTLs, allowing comparison of different models for different nos’ of QTLs.
Localisation of a QTL

• Specify a genetic model, a statistical model and an analytic technique.
• Focus on half-sib design: large number of half-sib families, missing genotype data from the dams.
• Biallelic QTL
Source of information

• Pedigree structure: K unrelated half-sib families, each family has a single male parent randomly mated to a large number of females, each of which produces a single offspring.

• Marker information: N linked informative (ie heterozygous) markers genotyped on the sires and all progeny but with no information on the dams.

• Phenotypic information: Quantitative trait measured for each offspring and influenced by both genetic and environmental factors.
Models

- **Genetic model**: Assume the quantitative trait is influenced by a biallelic QTL with alleles Q and q and associated allele frequencies $p_Q$ and $(1-p_Q)$ respectively. Both additive and dominance effects are modelled.
Models

• Statistical model:

\[ Y_{ji} = \mu_{Gji} + \varepsilon_{ji} \]

- \( Y_{ji} \) is the trait value for the ith offspring in the jth family
- \( G_{ji} \in \{QQ,Qq,qq\} = \{1,2,3\} \) is the ith offspring’s QTL genotype
- \( \mu_{Gji} \) is the expected trait value given the QTL genotype \( G_{ji} \)
- \( \varepsilon_{ji} \) is the offspring’s random deviation from \( \mu_{Gji} \). Includes environmental effects and has variance \( \sigma_R^2 \) (residual variance).

- Trait’s phenotypic variance (\( \sigma_P^2 \)) is the sum of the between-family (\( \sigma_{BF}^2 \)) and the within-family (\( \sigma_{WF}^2 \)) variances. For a half-sib design, \( \sigma_{WF}^2 = (3/4)\sigma_A^2 + \sigma_D^2 + \sigma_R^2; \sigma_{BF}^2 = (1/4)\sigma_A^2 \).
**Notation**

- $y_{ji}$: phenotypic value for the $i$th offspring in the $j$th family
- $f_j$: set of marker data for the $j$th family so that $f_{ji}$ is the genotype for the $N$ linked marker loci of the $i$th individual in the $j$th family
- $s$: vector of marker genotypes and phase configurations for the $K$ sires
- $M_{ij}$: $j$th allele of the $i$th gene marker $M_i$
- $\pi_j$: $j$th locus on the chromosome, may be a marker locus or a QTL, depending on the locus order
- $Q$: set of recombination rates between locus $\pi_j$ and $\pi_{j+1}$
- $\mu=(\mu_1,\mu_2,\mu_3)$, with expected values $m+a$, $m+d$, $m-a$ ($m$ is mid homozygote value, $a$ is additive, $d$ is dominance)
- $z_s$: sire’s unknown QTL genotype: $\{QQ,Qq,qQ,qq\} = \{1,2,3,4\}$
- $z_f$: $i$th offspring’s QTL genotype in the $j$th family, $\{QQ,Qq,qq\} = \{1,2,3\}$
Likelihood

- Likelihood for K half-sib families is equal to a mixture of four within-family likelihoods $p_{QQ}$, $p_{Qq}$, $p_{qQ}$, $p_{qq}$ weighted by the probability of the sire having QTL genotype QQ, Qq, qQ, qq, respectively.

\[
\prod_{j=1}^{K} \left\{ p_Q^2 p_{QQ} (Y_j \mid p_Q, \mu_1, \mu_2, \sigma_R^2) \right\} \\
+ p_Q (1 - p_Q) p_{Qq} (Y_j, f_j, s_j \mid \theta, p, p_Q, \mu, \sigma_R^2, \omega_l) \\
+ (1 - p_Q) p_{qQ} (Y_j, f_j, s_j \mid \theta, p, p_Q, \mu, \sigma_R^2, \omega_l) \\
+ (1 - p_Q)^2 p_{qq} (Y_j \mid p_Q, \mu_2, \mu_3, \sigma_R^2)
\]

This likelihood is conditioned on a specific locus order $w_l$:
$w_j=QM_1M_2...M_N$; $\omega_l=...M_{l-1}QM_{l+1}$ for $l=2,...,N$;
$w_{N+1}=M_1M_2...M_NQ$. 
Single-family likelihood

- For a sire with QTL genotype QQ or qq:
  - Mixture of two Gaussian densities (mean $\mu_j$, variance $\sigma_R^2$) weighted by the allele frequencies for Q and q from the dams’ contributions.
  - Multiply over the $n_j$ offspring in family j.
  - Genetic information does not feature in $p_{QQ}$ or $p_{qq}$ since the sire’s QTL is noninformative.

- For a sire with QTL genotype Qq or qQ:
  - Mixture of three Gaussian densities (offspring’s three QTL genotype classes), weighted by the conditional probability of an offspring having genotype $f_{ji}$ and QTL genotype $Q_{ji}$ given the sire’s marker genotype and phase $s_j$, locus order $w_1$ and the sire’s QTL genotype $Q_j$.
  - Multiply over the $n_j$ offspring in family j.
Conditional Probabilities

- Consider $j$th family with assumed paternal QTL $qQ$ and corresponding within-family likelihood $p_{qQ}(.)$
- The conditional probability for an offspring with marker $f_{ji}$ and QTL $Q_{ji}$ has the functional form

\[ P(f_{ji}, Q_{ji}|...Q_j=qQ) = \frac{f_E}{\sum f_y} \]

where:
- $f_{ji}$: expected frequency of A half-sib with marker genotype $f_{ji}$ and QTL genotype $Q_{ji}=y$
- Derived from a table of expected and observed genotype numbers

**Missing data:** dams’ genotypes – need to estimate for affected offspring
Priors and posterior

- Model the hidden data structure (individuals’ unknown QTL genotypes)
- Can write down:
  - the joint posterior for parameters \( \beta=\{\theta,p,p_Q,\mu,\sigma_R,\omega_1\} \) and the missing data \( z_s \) and \( z_f \) given the observed data and \( s \) (marker genotypes and phase configurations for the K sires)
  - The conditional distribution of the offspring’s QTL genotypes given the sire’s QTL genotypes and \( \beta \)
  - The conditional distribution of the sire’s QTL genotype given \( \beta \)

All mixture distributions!
But if we know \( z \), they simplify to a product of terms
Prior for recombination rates $\theta$

- Genotyping errors, recombination hot spots, nonuniformity of recombination events along the chromosomes contribute to errors in the linkage map from which the marker positions are obtained.
- To allow for this uncertainty, estimate the recombination rates in the analysis. Incorporate published marker information through strong priors.
- Assuming independence of recombination rates along the chromosome, the prior for $\theta$ is the product of the priors for the between-marker recombination rates. Set these as Gaussian centred around the relevant published value.
- No information on the QTL’s position, so place a uniform density on the recombination rate between the QTL and the marker.
Other priors

• Prior for $p$: set of allele probabilities. Allele probabilities at a given locus do not influence allele probabilities at an alternative locus because of linkage equilibrium, so the prior for $p$ is the product of priors for all individuals. Set a Dirichlet distribution on $(p_{i1}, p_{i2}, p_{i3})$.

• Prior for $p_Q$: Strong prior (normal or t-distribution) is possible using information from the design. Here we used a uniform distribution.

• Prior for $\sigma_R^2$ (residual variance): must be no larger than the observed phenotypic variance of the quantitative trait $\sigma_P^2$. So use uniform $U(0, \sigma_P^2)$.

• Prior for $\omega_l$ (locus order): equal weightings to different orders.
Update parameters, given the locus ordering

1. Allocate the $z_s$: sires are allocated QTL genotypes based on the family’s phenotypic and genotypic information and the current set of parameter values.

2. Given the sire’s QTL genotype, $z_f$ is allocated: the offspring are allocated QTL genotypes based on the offspring’s phenotypic and genotypic information and the current set of parameter values.

3. New parameter values $Q,p,p_Q,\mu,\sigma_R^2$ are sampled via the M-H algorithm for a specific locus ordering by treating $z_s$ and $z_f$ as known information.
Update locus ordering, given the parameters

Example:

- old values of locus order: $w_2^1 = AQBCD$ and recombination rates: $\theta^1 = (0.2, 0.4, 0.1, 0.2)$,
  Q is QTL; A,B,C,D are gene markers
- Convert recombination rates to map distances via the Haldane mapping function ($\delta = 0.5 \ln(1 - 2\theta)$), so that $\delta^1 = (\delta_{AQ}^1, \delta_{QB}^1, \delta_{BC}^1, \delta_{CD}^1) = (0.26, 0.81, 0.11, 0.26)$. (Map distances are easier to work with than recombination rates due to their additive nature.)
- Randomly select a marker to act as a pivot. For a forward or backward move, shift the QTL to the right or left of the pivot, respectively. The distance between the QTL and the pivot remains the same.
• Suppose a forward move is chosen and the pivot marker is B. The new locus order is dependent on the size of \( \delta_{QB}^1 \) in relation to \( \delta_{BC}^1 \) and \( \delta_{CD}^1 \). For example, \( \delta_{QB}^1 > \delta_{BC}^1 + \delta_{CD}^1 \), so \( \delta_{5}^2 = ABCDQ \).

• The vector of map distances associated with the new locus order is \( \delta_{AB}^2 = \delta_{AQ}^1 - \delta_{QB}^1 \); \( \delta_{BC}^2 = \delta_{BC}^1 \); \( \delta_{CD}^2 = \delta_{CD}^1 \); \( \delta_{DQ}^2 = \delta_{QB}^1 - (\delta_{BC}^1 + \delta_{CD}^1) \).

• Obtain \( \theta^2 \) by applying the inverse of the Haldane mapping function.

• Do a similar trick for a backward move. We can show the reversible nature of these moves.

• These relationships will change depending on the move type, the pivot selected, and the position of the markers on the chromosome, but the principle of shifting the QTL about a randomly selected pivot and adjusting the map distances accordingly is preserved.
Accepting proposed moves

• Proposed move is from $\text{old}=(\omega_2^1, \theta^1)$ to $\text{new}=(\omega_5^2, \theta^2)$

• Accept proposed move with probability

\[
\alpha = \min\left(1, \frac{p(\text{new} | \ldots)}{p(\text{old} | \ldots)} \frac{p_B}{p_F} \frac{|J|}{|J|}\right)
\]

$|J|$ is the determinant of the Jacobian (required because of the way we’ve moved)
pF: probability of making a forward move
pB: probability of making a backward move.
Complete algorithm

1. Given the locus ordering:
   1. Allocate the sires’ QTL genotypes
   2. Allocate the offspring’s QTL genotypes
   3. Update the parameter values

2. Update the locus ordering.
Example

- 6 designs:
  - sires = 20, 5
  - offspring = 200, 50
  - $\sigma_R = 0.5, 2.0$
  - $a/\sigma_p = 1.15a, 0.47a,$
  - $\sigma_Q^2/\sigma_P^2 = 0.67, 0.11$

- Design 1: $q_{M2Q}, p_Q, \mu_1, \sigma_R$
  
<table>
<thead>
<tr>
<th>True</th>
<th>Mean</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>.10</td>
<td>.10</td>
<td>(.08,.12)</td>
</tr>
<tr>
<td>.50</td>
<td>.48</td>
<td>(.45,.51)</td>
</tr>
<tr>
<td>71.00</td>
<td>71.02</td>
<td>(70.96,71.07)</td>
</tr>
<tr>
<td>.50</td>
<td>.50</td>
<td>(.48,.52)</td>
</tr>
</tbody>
</table>

- For each design, the true locus order had the highest posterior probability.
- Use Bayes factors to confirm strength of evidence. Some designs suggested spurious multiple QTL, due to half-sib data.
• Posterior probabilities for the locus order. True locus order is $M_1M_2QM_3M_4$.

<table>
<thead>
<tr>
<th>Locus Order</th>
<th>Design 1</th>
<th>Design 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q M_1M_2M_3M_4$</td>
<td>0.000</td>
<td>0.020</td>
</tr>
<tr>
<td>$M_1QM_2M_3M_4$</td>
<td>0.002</td>
<td>0.346</td>
</tr>
<tr>
<td>$M_1M_2QM_3M_4$</td>
<td>0.998</td>
<td>0.365</td>
</tr>
<tr>
<td>$M_1M_2M_3QM_4$</td>
<td>0.000</td>
<td>0.233</td>
</tr>
<tr>
<td>$M_1M_2QM_3M_4Q$</td>
<td>0.000</td>
<td>0.036</td>
</tr>
</tbody>
</table>

• Average percent of correctly allocated sires and offspring?

Sire allocation:  
Design 1: 1.00  
Design 2: 0.99

Offspring allocation:  
0.77  
0.51
Conclusions

• Weak priors result in poor estimation of some parameters.
• Stable estimation of locus ordering takes longer than estimation of parameters within a locus ordering.
• Poor mapping of QTL under some designs, perhaps due to half-sib design.
• Correct allocation of sires is influenced by the size of the QTL and the number of offspring in the half-sib families.
• Correct allocation of the offspring depends only on the strength of the individual’s data.
• The size of the QTL has a profound effect on the accuracy of the parameter estimates.
• Family size influences the performance of the sampler: easier to achieve convergence with a small number of large half-sib families compared to a large number of small half-sib families.
• ANIMAP overestimated the additive effect and failed to detect small QTL.
Example: Selective genotyping

- Introduced by Lander & Botstein (1989): increase power to detect QTL of a smaller effect
- Phenotype a population but genotype, via molecular markers, those with extreme phenotypes
Why selective genotyping?

- Lander & Botstein (1989): increase power to detect QTL of a smaller effect
- Darvasi & Soller (1992): for a single marker linked to a QTL, genotyping the upper & lower 25% is nearly as efficient as genotyping the whole population.
- Lin & Ritland (1996): for large linked QTL, may need tails of larger size or genotype all individuals
- Muranty & Goffinet (1997): for a trait other than that used for SG there is no loss of accuracy of QTL detection compared to genotyping a random sample of the same size.
Methods of analysis

• Lander & Botstein (1989): EM approach; need bias adjustment (Darvasi & Soller 1992)

• Baker et al (2000): Bayesian mixture model
  - General Bayesian package BUGS
Cattle mapping study (Davis et al, 1998)

Sire

Progeny: n=250 n=250 n=250

Phenotype data (adj. for fixed effects)

4 chromosomes A-D: 5, 5, 4, 4 markers
13 markers heterozygous for Sire 1;
16 for Sire 2
2 linked QTL on A, single QTL on C

Dams: up to 5 markers
Cattle Mapping study (Davis et al, 1998)

- 3 sires 2507, 2508, 2509 each with 250 progeny
- Phenotype data adjusted for fixed effects via linear models
- Four chromosomes A, B, C, E had 5, 5, 4, 4 markers respectively. Markers were between 20 & 35 cM apart.
- Two linked QTL on chromosome A at approx. 15 & 60 cM from 1st marker and a single QTL on chromosome C at approx 75 cM from 1st marker C1.
- 18 markers: 13 heterozygous for sires 2507 and 2509; 16 heterozygous for sire 2508.
Notation

• Offspring markers labelled as $M_1, M_2$ if they may have come from the sire or $M_3$ otherwise

• If a sire has marker alleles $M_1$ & $M_2$ and the QTL has 2 alleles $Q$ (positive for the trait) and $q$, then we assume the sire has marker-QTL genotype $M_1Q/M_2q$.

• The dams have 3 marker alleles $M_1, M_2, M_3$ with unknown proportions $t_1, t_2, (1- t_1- t_2)$. 
Genetic effects

Genetic effect
Of QTL alleles
\{\mu_{qq}, \mu_{Qq}, \mu_{QQ}\}

Additive (>0) & dominance components

\[
\begin{pmatrix}
G_0 \\
G_1 \\
G_2
\end{pmatrix} =
\begin{pmatrix}
1 \\
1 \\
1
\end{pmatrix} \mu +
\begin{pmatrix}
-1 & -1/2 \\
0 & 1/2 \\
1 & -1/2
\end{pmatrix}
\begin{pmatrix}
a \\
d
\end{pmatrix} =
1 \mu + D \begin{pmatrix}
a \\
d
\end{pmatrix}
\]

Overall phenotypic mean
Genetic effects

- Define genetic effect of QTL alleles \{\mu_{qq}, \mu_{Qq}, \mu_{QQ}\} as G_0, G_1, G_2.
- overall phenotypic mean \mu
- additive & dominance components a>0, d
- Genetic effects written as:

\[
\begin{pmatrix}
G_0 \\
G_1 \\
G_2
\end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} -1 & -1/2 \\ 0 & 1/2 \\ 1 & -1/2 \end{pmatrix} \begin{pmatrix} a \\ d \end{pmatrix} = \mu + D \begin{pmatrix} a \\ d \end{pmatrix}
\]
Without marker information

X is design matrix for fixed effects

\[ y \sim \sum_{k=1}^{3} p_k N \left( X\beta + \Delta_k D \begin{pmatrix} a \\ d \end{pmatrix}, \sigma^2 I \right) \]

\( \Delta \) is nx3 matrix, 1 = QTL genotype, 0 otherwise

Biallelic QTL: Q,q
p is prop’n Q alleles in dam population
Without marker information

• Define $\Delta$ as nx3 matrix with elements 1 to indicate QTL genotype, 0 otherwise
• Define $X$ as design matrix for fixed effects $\beta$.
• For observed phenotype data without marker information:

\[
y \sim \sum_{k=1}^{3} p_k N \left( X\beta + \Delta_k D \begin{pmatrix} a \\ d \end{pmatrix}, \sigma^2 I \right)
\]

• (could replace by skewed $t$ or multivariate $t$ within clusters)
With marker information

- $p_k$ are conditional on observed marker types
- Let $r$ be recombination rate between the QTL and marker.
- Probabilities of a QTL given a linked marker:

\[
R = \begin{pmatrix}
M_1M_1 & QQ & Qq & qq \\
M_1M_2 & p(1-r) & (1-p)(1-r)+rp & (1-p)r \\
M_1M_3 & t_1p+r_2p(1-r) & t_1(p(1-r)+r(1-p))+t_2(pr+(1-p)(1-r)) & t_1(1-p)(1-r)+t_2(1-p)r \\
M_2M_2 & t_1+t_2 & (1-p)(1-r)+rp & (1-p)r \\
M_2M_3 & p(1-r) & p(1-r)+r(1-p) & (1-p)(1-r) \\
M_2M_3 & p(1-r) & p(1-r)+r(1-p) & (1-p)(1-r)
\end{pmatrix}
\]
Likelihood: without SG

- \( j = 1 \) for \( M_1M_2 \), 2 for \( M_1M_2 \), 3 for \( M_2M_2 \) etc
- \( \delta_{M(j)} = 1 \) if offspring has marker type \( j \); 0 otherwise
- \( \delta_l = \) nx1 vector with \( l^{th} \) element 1; 0 otherwise

\[
f(y | \mu, a, d, r, p, \beta, \sigma) = \prod_{l=1}^{n} \left\{ \sum_{j=1}^{5} \delta_{M(j)} \left( \sum_{k=1}^{3} R_{(jk)} N(y_l, \delta_l^T \left(X\beta + \Delta_k D \begin{bmatrix} a \\ d \end{bmatrix}, \sigma^2) \right) \right) \right\}
\]
Likelihood: with SG

• $\delta_{M(j)}$ is not 0 or 1 if the marker type is not recorded for that individual

• Instead, $\delta_{M(j)}$ is $\Pr(\text{marker type} = j), j=1,..,5$) and must be estimated.

• This results in a finite mixture with 15 components for those individuals not genotyped.
Bayesian approach

• Posterior $\propto$ Likelihood x prior

\[
f(\mu, a, d, r, p, t_1, t_2, \beta_1, \ldots, \beta_k, \sigma \mid y) \\
\propto f(y \mid \ldots) f(\mu) f(a) \ldots f(\sigma)
\]

• Marginal posterior distribution of any parameter is found by integrating the joint posterior over all other parameters.

• Use MCMC for computation.
Priors

- No informative prior information, so disperse proper priors were adopted:

\[
\begin{align*}
\mu & \sim N(0,10^{12}), \ a \sim N(0,10^{6})I(a > 0), \\
r & \sim U(0,0.5), \ \tau = \sigma^{-2} \sim Ga(.001,.001) \\
G & \sim Dirichlet(1,2,1)
\end{align*}
\]

- For selective genotyping, prior on marker types:

\[
\delta_{M(j)} \sim Dirichlet(1,...,1)
\]
Mixtures as latent variables

- Estimate $\delta_{M(j)}$.
- Then allocate ungenotyped individuals to the 5 marker types.
- For each marker type, estimate $R_{(jk)}$.
- Then allocate individuals to the 3 mixture components.
- Estimate component-specific parameters using those individuals allocated to that component.
- Estimate other parameters.
Results

- Four chromosomes A, B, C, E had 5, 5, 4, 4 markers respectively.
- Markers were between 20 & 35 cM apart.
- Two linked QTL on chromosome A at approx. 15 & 60 cM from 1st marker and a single QTL on chromosome C at approx 75 cM from 1st marker C1.

<table>
<thead>
<tr>
<th>Sire</th>
<th>Marker</th>
<th>Full Data</th>
<th>S.G. Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dist (cM)</td>
<td>a</td>
</tr>
<tr>
<td>2509</td>
<td>A3</td>
<td>29 27 33 17.8</td>
<td>29 28 35 15.4</td>
</tr>
<tr>
<td>2508</td>
<td>C1</td>
<td>32 24 24 5.1</td>
<td>31 24 22 4.1</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>32 20 17 3.9</td>
<td>28 19 15 4.0</td>
</tr>
<tr>
<td></td>
<td>C3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BF: Effect size
Full vs SG datasets: Bayes Factors