Taking the Broad View of Model Selection for QTL in Experimental Crosses

Brian S. Yandell
University of Wisconsin-Madison
www.stat.wisc.edu/~yandell/statgen
with Chunfang “Amy” Jin, UW-Madison,
Patrick J. Gaffney, Lubrizol,
and Jaya M. Satagopan, Sloan-Kettering
Plant & Animal Genome XI, January 2003
Pareto diagram of QTL effects

major QTL on linkage map

additive effect

rank order of QTL

0 5 10 15 20 25 30

0 1 2 3

(sqrt order of QTL)

polygenes

major QTL

minor QTL

(modifiers)
how many (detectable) QTL?

• build $m =$ number of QTL detected into model
  – directly allow uncertainty in genetic architecture
  – model selection over number of QTL, architecture
  – use Bayes factors and model averaging
    • to identify “better” models

• many, many QTL may affect most any trait
  – how many QTL are detectable with these data?
    • limits to useful detection (Bernardo 2000)
    • depends on sample size, heritability, environmental variation
  – consider probability that a QTL is in the model
    • avoid sharp in/out dichotomy
    • major QTL usually selected, minor QTL sampled infrequently
interval mapping basics

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

after Sen Churchill (2001)
Classical vs. Bayesian IM

• MIM: classical LOD: mix over genotypes $Q$
  
  $L(\lambda, \theta|Y,m) = \text{pr}(Y|X,\lambda,\theta,m)$
  
  $= \text{product}_i \left[ \sum_Q \text{pr}(Q_i|X_i,\lambda,m) \text{pr}(Y_i|Q_i,\theta,m) \right]$  
  
  maximize LOD($\lambda$) = $2.3 \log(LR(\lambda)) = \max_{\theta} \log_{10} L(\lambda,\theta|Y,m)/L(\mu|Y)$
  
  threshold for testing presence of QTL
  
  Kao Zeng Teasdale 1999; Zeng et al. 2000; Broman Speed 2002

• BIM: Bayesian posterior: $Q$ as missing data
  
  sample genotypes $Q$, loci $\lambda$, effects $\theta$ and number of QTL $m$
  
  $\text{pr}(\lambda,Q,\theta,m|Y,X) = \left[ \text{product}_i \text{pr}(Q_i|X_i,\lambda,m) \text{pr}(Y_i|Q_i,\theta,m) \right] \text{pr}(\lambda,\theta|X,m)\text{pr}(m)$
  
  study marginal posteriors
  
  $\text{pr}(\lambda,\theta|Y,X,m) = \sum_Q \text{pr}(\lambda,Q,\theta|Y,X,m)$ with $m$ fixed
  
  $\text{pr}(m|Y,X) = \sum_{(\lambda,\theta)} \text{pr}(\lambda,\theta|Y,X,m)\text{pr}(m)$
  
  threshold for posterior “power” (positive false discovery rate)
  
  Satagopan et al. 1996; Gaffney 2001; Yi Xu 2002
Model Selection for QTL

• what is the genetic architecture?
  – \( M = \text{model} = (\lambda, \theta, m) \)
  – \( \lambda = \text{QT locus (or loci)} \)
  – \( \theta = \text{genetic action (additive, dominance, epistasis)} \)
  – \( m = \text{number of QTL} \)

• how to assess models?
  – MIM: various flavors of AIC, BIC
  – BIM: Bayes factors

• how to search model space?
  – MIM: sequential forward selection/backward elimination
    • scan loci systematically across genome
  – BIM: sample forward/backward: transdimensional MCMC
    • sample loci at random across genome
Bayes factors to assess models

• Bayes factor: which model best supports the data?
  – ratio of posterior odds to prior odds
  – ratio of model likelihoods

• equivalent to $LR$ statistic when
  – comparing two nested models
  – simple hypotheses (e.g. 1 vs 2 QTL)

• related to Bayes Information Criteria (BIC)
  – Schwartz introduced for model selection in general settings
  – penalty to balance model size ($p = \text{number of parameters}$)

\[
BF = \frac{\text{pr}(m | Y, X)}{\text{pr}(m+1 | Y, X)} = \frac{\text{pr}(Y | m, X)}{\text{pr}(Y | m+1, X)}
\]
\[
-2 \log(BF) = -2 \log(LR) - 2 \log(n)
\]
QTL Bayes factors & RJ-MCMC

• easy to compute Bayes factors from samples
  – posterior \( \Pr(m|Y,X) \) is marginal histogram
  – posterior affected by prior \( \Pr(m) \)

• \( BF \) insensitive to shape of prior
  – 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
multiple QTL phenotype model

- \( Y = \mu + G_Q + \text{environment} \)
- partition genotypic effect into separate QTL effects
  \( G_Q = \text{main QTL effects} + \text{epistatic interactions} \)
  \( G_Q = \theta_1 Q + \ldots + \theta_m Q + \theta_{12} Q + \ldots \)
- priors on mean and effects
  \( G_Q \sim N(0, h^2 s^2) \) model independent genotypic prior
  \( \theta_{jQ} \sim N(0, \kappa_1 s^2/m.) \) effects and interactions
  \( \theta_{j2Q} \sim N(0, \kappa_2 s^2/m.) \) down-weighted
- hyperparameters (to reduce sensitivity of Bayes factors to prior)
  - \( s^2 = \text{total sample variance} \)
  - \( m. = m + m_2 = \text{number of QTL effects and interactions} \)
  - \( h^2 = \kappa_1 + \kappa_2 = \text{unknown heritability, } h^2/2 \sim \text{Beta}(a,b) \)
Markov chain Monte Carlo idea

have posterior $pr(\theta|Y)$
want to draw samples

 propose $\theta \sim pr(\theta|Y)$
(ideal: Gibbs sample)

propose new $\theta$ “nearby”
accept if more probable
toss coin if less probable
based on relative heights
(Metropolis-Hastings)
MCMC realization

added twist: occasionally propose from whole domain
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>

0 50 100 150 200

ch1 ch2 ch3 ch4 ch5 ch6 ch7 ch8 ch9 ch10

Genetic map

posterior

frequency in %

number of QTL
loci pattern across genome

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

### Chromosome

<table>
<thead>
<tr>
<th>m</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>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>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>
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
$h^2 =$ 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

evidence suggests 4-5 QTL N2(2-3),N3,N16

col 1: posterior

col 2: Bayes factor

row 1: # QTL

row 2: pattern

note error bars on bf
Bayesian model diagnostics

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

environmental variance
\[ \sigma^2 = 0.008, \quad \sigma = 0.09 \]
heritability
\[ h^2 = 52\% \]
LOD = 16
(highly significant)

but note change with \( m \)

January 2003
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
loci marginal posteriors

unlinked loci

linked loci

January 2003

PAG XI © Brian S. Yandell
mapping gene expression

- 108 F2 mice
- mRNA to RT-PCR
- multivariate screen
  - clustering
  - PC analysis
- highlight SCD
- Lan et al. (2003)

- ch2 dominance
false detection rates and posteriors

• multiple comparisons: test QTL across genome
  – size = \( \Pr( \text{LOD}(\lambda) > t \mid \text{no QTL at } \lambda ) \)
  – genome-wise threshold
    • theoretical value or permutation value (Churchill Doerge 1995)
  – threshold guards against a single false detection
  – difficult to extend to multiple QTL

• positive false discovery rate (Storey 2001)
  – \( \text{pFDR} = \Pr( \text{no QTL at } \lambda \mid \text{LOD}(\lambda) > t ) \)
  – consider proportion of false detections for threshold
  – related to Bayesian posterior
  – extends naturally to multiple QTL
pFDR and QTL posterior

• single QTL case
  – pick a rejection region \( R = \{ \lambda | \text{LOD}(\lambda) > t \} \) for some \( t \)
  – \( \text{pFDR} = \frac{\text{Pr}(m=0) \times \text{size}}{\text{Pr}(m=0) \times \text{size} + \text{Pr}(m=1) \times \text{power}} \)
  – \( \text{power} = \text{Pr}(\lambda \text{ in } R \mid Y, X, m = 1) \)
  – \( \text{size} = (\text{length of } R) / (\text{length of genome}) \)

• multiple QTL case
  – \( \text{pFDR} = \frac{\text{Pr}(m=0) \times \text{size}}{\text{Pr}(m=0) \times \text{size} + \text{Pr}(m>1) \times \text{power}} \)
  – \( \text{power} = \text{Pr}(\lambda \text{ in } R \mid Y, X, m > 1) \)

• extends to other null hypotheses
  – \( \text{pFDR} = \frac{\text{Pr}(m=1) \times \text{size}}{\text{Pr}(m=1) \times \text{size} + \text{Pr}(m>2) \times \text{power}} \)
B napus with $m \sim \text{Poisson}(1)$
Summary

• Bayesian posteriors and Bayes factors
  – Bayes factors for model assessment
  – posteriors can reveal subtle hints of QTL

• graphical tools for model selection
  – Bayes factor ratios on log scale
  – model identified by $m$ or genetic architecture

• connection to false discovery rate
  – whole genome evaluation
  – calibrate posterior region with pFDR