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 = MSE = (bias)^2 + 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
- neither $E_1$ nor $E_2$ is rate limiting
- loss of function alleles are segregating from parent B at $E_1$ and 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
    - $|A| = 3^n$ 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

|      | $bb$ | $bB$ | $BB$ | 2 linked QTL | empty cell | with $n = 100$
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$aa$</td>
<td>6</td>
<td>15</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$aA$</td>
<td>15</td>
<td>25</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$AA$</td>
<td>3</td>
<td>15</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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^n 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
- $p_r(q|m,\lambda,\gamma)$ genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for $q$ given $m$
- $p_r(y|q,\mu,\gamma)$ 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”

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

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

likelihood mixes over missing QTL genotypes:

\[
p_r(y|m,\mu,\lambda) = \sum_q p_r(y|q,\mu) p_r(q|m,\lambda)
\]
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

\[
\text{LOD}(\lambda \ | \ \gamma_2) - \text{LOD}(\lambda \ | \ \gamma_1) \\
\text{LPD}(\lambda \ | \ \gamma_2) - \text{LPD}(\lambda \ | \ \gamma_1)
\]
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></th>
<th>smaller model</th>
<th>bigger model</th>
</tr>
</thead>
<tbody>
<tr>
<td>fit model</td>
<td>miss key features</td>
<td>fits better</td>
</tr>
<tr>
<td>estimate phenotype</td>
<td>may be biased</td>
<td>no bias</td>
</tr>
<tr>
<td>predict new data</td>
<td>may be biased</td>
<td>no bias</td>
</tr>
<tr>
<td>interpret model</td>
<td>easier</td>
<td>more complicated</td>
</tr>
<tr>
<td>estimate effects</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
QTL 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.mri.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

USDA Hatch, NIH/NIDDK (Attie), NIH/R01s (Yi, Broman)

UW-Madison Stats
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 Kendziorski
Daniel Gianola
Liang Li
Daniel Sorensen
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.rqtl.org/tutorials/rqtltour.R
- R/qtlbim tutorial
  - R/qtlbim web site: www.qtlbim.org
  - Tutorial: www.stat.wisc.edu/~yandell/qtlbim/rqtltour.pdf
  - R code: www.stat.wisc.edu/~yandell/qtlbim/rqtltour.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 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

```r
> 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>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>215</td>
<td>D1Mit267</td>
<td>5.822192</td>
</tr>
<tr>
<td>17</td>
<td>108</td>
<td>D1Mit267</td>
<td>5.822192</td>
</tr>
<tr>
<td>18</td>
<td>138</td>
<td>D1Mit267</td>
<td>5.822192</td>
</tr>
<tr>
<td>19</td>
<td>226</td>
<td>D1Mit267</td>
<td>5.822192</td>
</tr>
<tr>
<td>20</td>
<td>199</td>
<td>D1Mit267</td>
<td>5.819250</td>
</tr>
<tr>
<td>21</td>
<td>84</td>
<td>D1Mit267</td>
<td>5.808400</td>
</tr>
</tbody>
</table>

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

QTL 2: Tutorial Seattle SISG: Yandell © 2008
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
  c1.loc45  1 48.3 3.52
  D4Mit164  4 29.5 8.02

> summary(out.hk, threshold=3)
  chr pos  lod
  c1.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)
```

QTL 2: Tutorial Seattle SISG: Yandell © 2008

black = EM
blue = HK

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

```r
> 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))

<table>
<thead>
<tr>
<th>position</th>
<th>position</th>
<th>lod.full</th>
<th>lod.fv1</th>
<th>lod.int</th>
<th>position</th>
<th>position</th>
<th>lod.add</th>
<th>lod.av1</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1 : c4</td>
<td>68.3</td>
<td>30.0</td>
<td>14.13</td>
<td>6.51</td>
<td>0.225</td>
<td>68.3</td>
<td>30.0</td>
<td>13.90</td>
</tr>
<tr>
<td>c2 : c19</td>
<td>47.7</td>
<td>0.0</td>
<td>6.71</td>
<td>5.01</td>
<td>3.458</td>
<td>52.7</td>
<td>0.0</td>
<td>3.25</td>
</tr>
<tr>
<td>c3 : c3</td>
<td>37.2</td>
<td>42.2</td>
<td>6.10</td>
<td>5.08</td>
<td>0.226</td>
<td>37.2</td>
<td>42.2</td>
<td>5.87</td>
</tr>
<tr>
<td>c6 : c15</td>
<td>60.0</td>
<td>20.5</td>
<td>7.17</td>
<td>5.22</td>
<td>3.237</td>
<td>25.0</td>
<td>20.5</td>
<td>3.93</td>
</tr>
<tr>
<td>c9 : c18</td>
<td>67.0</td>
<td>37.2</td>
<td>6.31</td>
<td>4.79</td>
<td>4.083</td>
<td>67.0</td>
<td>12.2</td>
<td>2.23</td>
</tr>
<tr>
<td>c12 : c19</td>
<td>1.1</td>
<td>40.0</td>
<td>6.48</td>
<td>4.79</td>
<td>4.090</td>
<td>1.1</td>
<td>0.0</td>
<td>2.39</td>
</tr>
</tbody>
</table>

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

upper triangle/left scale: epistasis LOD
lower triangle/right scale: 2-QTL LOD
R/qtl: ANOVA imputation at QTL

```r
> hyper <- sim.geno(hyper, step=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>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>0</td>
</tr>
<tr>
<td>Error</td>
<td>243</td>
<td>11879.847</td>
<td>48.88826</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>249</td>
<td>17668.936</td>
<td>48.88826</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Drop one QTL at a time ANOVA table:

```
------------------
Model formula is:  y ~ Q1 + Q2 + Q3 + Q4 + Q5 + Q4:Q5
------------------

<table>
<thead>
<tr>
<th>Chr</th>
<th>df</th>
<th>Type III SS</th>
<th>LOD</th>
<th>%var</th>
<th>F</th>
<th>Pvalue(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr1@50</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>Chr1@76</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>Chr4@30</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>Chr6@70</td>
<td>2</td>
<td>1435.721</td>
<td>6.194</td>
<td>8.126</td>
<td>14.684</td>
<td>9.55e-07 ***</td>
</tr>
<tr>
<td>Chr15@20</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>Chr6@70:Chr15@20</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 '.' 0.1 ' ' 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
Bayesian Interval Mapping

1. Bayesian strategy 3-19
2. Markov chain sampling 20-27
3. sampling genetic architectures 28-35
4. criteria for model selection 36-44

QTL model selection: key players

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

after Sen Churchill (2001)
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)

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

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

\[
\text{pr}(q, \mu, \lambda, \gamma \mid y, m) = \frac{\text{pr}(y \mid q, \mu, \lambda, \gamma) \times \text{[pr}(q \mid m, \lambda, \gamma) \times \text{pr}(\mu \mid \gamma) \times \text{pr}(\lambda \mid m, \gamma) \times \text{pr}(\gamma))]}{\text{pr}(y \mid m)}
\]

Bayes posterior for normal data

- For small prior variance, the posterior mean is closer to the actual mean.
- For large prior variance, the posterior mean is closer to the prior mean.

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

posterior: mean tends to sample mean

single individual \( \mu \sim N(\mu_0 + b_1(y_1 - \mu_0), b_1 \sigma^2) \)

sample of \( n \) individuals \( \mu \sim N\left(b_n \bar{y} + (1 - b_n)\mu_0, b_n \sigma^2 / n\right) \)

with \( \bar{y} = \text{sum } y_i / n \)

shrinkage factor (shrinks to 1) \( b_n = \frac{\kappa \mu}{\kappa \mu + 1} \rightarrow 1 \)

what values are the genotypic means?

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

![Diagram showing the relationship between data means, prior mean, and posterior means with QTL 2: Bayes Seattle SISG: Yandell © 2008 on the bottom right corner.]
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:
\[ E(\mu_q) = \overline{y} \quad V(\mu_q) = \kappa \sigma^2 \]

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

\[ n_q = \text{count}\{q_i = q\} \quad \overline{y}_q = \frac{\sum y_i / n_q}{\left\{q_i = q\right\}} \]

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

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 \( \hat{\beta}_q = \text{sum}_i [\text{sum}_j \hat{\beta}(q_{ij})] = \text{sum}_i w_{qi} y_i \)

variance

\[ V(\hat{\beta}_q) = \text{sum}_i w_{qi}^2 \sigma^2 = C_q \sigma^2 \]

shrinkage

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

\[
\begin{align*}
&m_1 \quad m_2 \quad q? \quad m_3 \quad m_4 \quad \text{markers} \quad m_5 \quad m_6 \\
&\lambda \quad \text{distance along chromosome}
\end{align*}
\]
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 $\Pr(q \mid m, \lambda)$
    - weight toward $q$ that agrees with flanking markers
  - proportional to likelihood $\Pr(y \mid q, \mu)$
    - weight toward $q$ with similar phenotype values
  - posterior recombination model balances these two

- this is the E-step of EM computations

$$
\Pr(q \mid y, m, \mu, \lambda) = \frac{\Pr(y \mid q, \mu) \cdot \Pr(q \mid m, \lambda)}{\Pr(y \mid m, \mu, \lambda)}
$$
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 $q$
  \[
  \text{pr}(\lambda \mid m,q) = \text{pr}(\lambda) \text{pr}(q \mid m,\lambda) / \text{constant}
  \]
  - 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
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 $q, y$ (using Metropolis-Hastings step)
  - sample genotypes $q$ given $\lambda, \mu, y$ (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)

\[
(\lambda, q, \mu, y) \sim p(\lambda, q, \mu, y \mid y, m)
\]

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

\[
\begin{align*}
\left( \mu_1 \right) & \sim N \left( \begin{pmatrix} 0 \\ \rho \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \\
\mu_1 & \sim N \left( \rho \mu_2, 1 - \rho^2 \right) \\
\mu_2 & \sim N \left( \rho \mu_1, 1 - \rho^2 \right)
\end{align*}
\]

Gibbs sampler samples: $\rho = 0.6$

$N = 50$ samples

$N = 200$ samples
full conditional for locus

- cannot easily sample from locus full conditional
  \[ \text{pr}(\lambda | y, m, \mu, q) = \text{pr}(\lambda | m, q) = \text{pr}(q | m, \lambda) \text{pr}(\lambda) / \text{constant} \]
- 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 © 2008
3. sampling genetic architectures

- search across genetic architectures $A$ 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

reversible jump MCMC

- consider known genotypes $q$ at 2 known loci $\lambda$
  - models with 1 or 2 QTL
- M-H step between 1-QTL and 2-QTL models
  - model changes dimension (via careful bookkeeping)
  - consider mixture over QTL models $H$

\[ \gamma = 1 \text{QTL} : Y = \beta_0 + \beta(q_1) + e \]

\[ \gamma = 2 \text{QTL} : Y = \beta_0 + \beta(q_1) + \beta(q_2) + e \]
geometry of reversible jump

Move Between Models

Reversible Jump Sequence

c21 = 0.7

geometry allowing $q$ and $\lambda$ to change

a short sequence

first 1000 with $m<3$
collinear QTL = correlated effects

4-week

8-week

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

sampling across QTL models $\gamma$

action steps: draw one of three choices
• update QTL model $\gamma$ with probability $1 - b(\gamma) \cdot d(\gamma)$
  – 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
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)

\[
\mu_q = \mu + \gamma_1 \beta_1(q_1) + \gamma_2 \beta_2(q_2), \quad \gamma_k = 0,1
\]

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

\[
\mu_q = \beta_0 + \gamma_1 \beta_1(q_1) + \gamma_2 \beta_2(q_2), \quad 0 \leq \gamma_k \leq 1
\]
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 $\text{pr}(y \mid \gamma)$
  – average over unknowns

classical information criteria

• start with likelihood $L(\gamma \mid y, m)$
  – 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
  – $|\gamma| = \text{size of genetic architecture}$
    • BC: $|\gamma| = 1 + n.gtl + n.gtl(n.gtl - 1) = 1 + 4 + 12 = 17$
    • F2: $|\gamma| = 1 + 2n.gtl +4n.gtl(n.gtl - 1) = 1 + 8 + 48 = 57$
classical information criteria

• construct information criteria
  – balance fit to complexity
  – Akaike AIC = \(-2 \log(L) + 2 |\gamma|\)
  – Bayes/Schwartz BIC = \(-2 \log(L) + |\gamma| \log(n)\)
  – Broman BIC\(\delta\) = \(-2 \log(L) + \delta |\gamma| \log(n)\)
  – general form: 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
    • 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

\[
B_{12} = \frac{\text{pr}(\gamma_1 | y,m) / \text{pr}(\gamma_2 | y,m)}{\text{pr}(\gamma_1) / \text{pr}(\gamma_2)} = \frac{\text{pr}(y | m,\gamma_1)}{\text{pr}(y | m,\gamma_2)}
\]

- 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 $\gamma$

• $|\gamma| = \text{number of QTL}$
  – 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**

\[ \beta_{ij} \sim N\left(0, \sigma_G^2/m\right), \sigma_G^2 = h^2 \sigma_{\text{total}}^2, h^2 \text{ fixed} \]

**BF insensitivity to random effects prior**

\[ \beta_{ij} \sim N\left(0, \sigma_G^2/m\right), \sigma_G^2 = h^2 \sigma_{\text{total}}^2, \frac{1}{2} h^2 \sim \text{Beta}(a,b) \]
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

R/qtlbim: tutorial
(www.stat.wisc.edu/~yandell/qtlbim)

```r
> 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.
> qb.load(hyper, 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>varadd</th>
<th>varaa</th>
<th>var</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>2.000</td>
<td>97.42</td>
<td>28.07</td>
<td>5.112</td>
<td>0.000</td>
<td>5.112</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>5.000</td>
<td>101.00</td>
<td>44.33</td>
<td>17.010</td>
<td>1.639</td>
<td>20.180</td>
</tr>
<tr>
<td>Median</td>
<td>7.000</td>
<td>101.30</td>
<td>48.57</td>
<td>20.060</td>
<td>4.580</td>
<td>25.160</td>
</tr>
<tr>
<td>Mean</td>
<td>6.543</td>
<td>101.30</td>
<td>48.80</td>
<td>20.310</td>
<td>5.321</td>
<td>25.630</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>8.000</td>
<td>103.70</td>
<td>53.11</td>
<td>23.480</td>
<td>7.862</td>
<td>30.370</td>
</tr>
<tr>
<td>Max.</td>
<td>13.000</td>
<td>103.90</td>
<td>74.03</td>
<td>51.730</td>
<td>34.940</td>
<td>65.220</td>
</tr>
</tbody>
</table>

Percentages for number of QTL detected:

<table>
<thead>
<tr>
<th></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>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>2</td>
<td>3</td>
<td>9</td>
<td>14</td>
<td>21</td>
<td>19</td>
<td>17</td>
<td>10</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Percentages for number of epistatic pairs detected:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>29</td>
<td>31</td>
<td>23</td>
<td>11</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

Percentages for common epistatic pairs:

|         | 0.15 | 0.15 | 0.6 | 1.75 | 1.4 | 1.6 | 4.9 | 1.15 | 1.17 | 1.5 | 5.11 | 1.2 | 1.6 | 4.9 | 1.15 | 1.17 | 1.5 | 5.11 | 1.2 |
|---------|------|------|-----|------|-----|-----|-----|------|------|-----|------|-----|-----|-----|------|------|-----|------|-----|-----|
| Total   | 63   | 18   | 10  | 6    | 6   | 5   | 4   | 4    | 3    | 3   | 2    | 2   | 2   | 2   | 2    | 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>67.8</td>
<td>6.10</td>
<td>0.442</td>
<td>6.57</td>
</tr>
<tr>
<td>c4</td>
<td>1.377</td>
<td>29.5</td>
<td>29.5</td>
<td>11.49</td>
<td>0.375</td>
<td>11.61</td>
</tr>
<tr>
<td>c6</td>
<td>0.838</td>
<td>59.0</td>
<td>59.0</td>
<td>3.99</td>
<td>6.265</td>
<td>9.60</td>
</tr>
<tr>
<td>c15</td>
<td>0.961</td>
<td>17.5</td>
<td>17.5</td>
<td>1.30</td>
<td>6.325</td>
<td>7.28</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</th>
<th>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</td>
<td>2.71e-05</td>
<td>24.30</td>
</tr>
<tr>
<td>1,4,6,6,15,6:15</td>
<td>6</td>
<td>0.00467</td>
<td>5.22e-06</td>
<td>17.40</td>
</tr>
<tr>
<td>1,1,4,6,15,6:15</td>
<td>6</td>
<td>0.00600</td>
<td>9.05e-06</td>
<td>12.80</td>
</tr>
<tr>
<td>1,4,5,6,15,6:15</td>
<td>7</td>
<td>0.00267</td>
<td>4.11e-06</td>
<td>12.60</td>
</tr>
<tr>
<td>1,6,15,15,6:15</td>
<td>6</td>
<td>0.00300</td>
<td>4.96e-06</td>
<td>11.70</td>
</tr>
<tr>
<td>1,4,6,6,6:15,15</td>
<td>6</td>
<td>0.00300</td>
<td>5.81e-06</td>
<td>10.00</td>
</tr>
<tr>
<td>1,2,4,6,15,6:15</td>
<td>6</td>
<td>0.00767</td>
<td>1.54e-05</td>
<td>9.66</td>
</tr>
<tr>
<td>1,4,5,6,15,6:15</td>
<td>6</td>
<td>0.00500</td>
<td>1.28e-05</td>
<td>7.56</td>
</tr>
<tr>
<td>1,2,4,5,6,15,6:15</td>
<td>7</td>
<td>0.00267</td>
<td>6.98e-06</td>
<td>7.41</td>
</tr>
<tr>
<td>1,4</td>
<td>2</td>
<td>0.01430</td>
<td>1.51e-04</td>
<td>1.84</td>
</tr>
<tr>
<td>1,1,2,4</td>
<td>4</td>
<td>0.00300</td>
<td>3.66e-05</td>
<td>1.59</td>
</tr>
<tr>
<td>1,2,4</td>
<td>3</td>
<td>0.00733</td>
<td>1.03e-04</td>
<td>1.38</td>
</tr>
<tr>
<td>1,1,4</td>
<td>3</td>
<td>0.00400</td>
<td>6.05e-05</td>
<td>1.28</td>
</tr>
<tr>
<td>1,4,19</td>
<td>3</td>
<td>0.00300</td>
<td>5.82e-05</td>
<td>1.00</td>
</tr>
</tbody>
</table>

> plot(qb.BayesFactor(qbHyper, item = "nqtl"))
```

hyper: number of QTL

posterior, prior, Bayes factors

![Graph of QTL posterior and Bayes factor ratios](image)
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

chrom locus locus.LCL locus.UCL n.qtl
247  1  69.9  24.44875   95.7985 0.8026667
245  4  29.5  14.20000   74.3000 0.8800000
248  6  59.0  13.83333   66.7000 0.7096667
246 15  19.5  13.10000   55.7000 0.8450000
```

```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
how close are other patterns?

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

Score by sample number of qtl

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2</td>
<td>1.437</td>
<td>1.735</td>
<td>1.919</td>
<td>1.919</td>
<td>2.000</td>
</tr>
<tr>
<td>3.2</td>
<td>1.351</td>
<td>1.735</td>
<td>1.916</td>
<td>1.919</td>
<td>2.016</td>
</tr>
<tr>
<td>4.0</td>
<td>1.270</td>
<td>1.916</td>
<td>2.437</td>
<td>2.648</td>
<td>3.574</td>
</tr>
<tr>
<td>5.0</td>
<td>1.295</td>
<td>1.919</td>
<td>2.835</td>
<td>2.798</td>
<td>3.611</td>
</tr>
<tr>
<td>6.0</td>
<td>1.257</td>
<td>2.254</td>
<td>3.451</td>
<td>3.029</td>
<td>3.648</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13.2</td>
<td>3.694</td>
<td>3.694</td>
<td>3.694</td>
<td>3.694</td>
<td>3.694</td>
</tr>
</tbody>
</table>

Score by sample chromosome pattern

<table>
<thead>
<tr>
<th>Percent</th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>4@1,4,6,15,6:15</td>
<td>3.4</td>
<td>2.946</td>
<td>3.500</td>
<td>3.630</td>
<td>3.613</td>
<td>3.758</td>
</tr>
<tr>
<td>2@1,4</td>
<td>1.4</td>
<td>1.437</td>
<td>1.735</td>
<td>1.919</td>
<td>1.832</td>
<td>1.919</td>
</tr>
<tr>
<td>5@1,2,4,6,15,6:15</td>
<td>0.8</td>
<td>3.137</td>
<td>3.536</td>
<td>3.622</td>
<td>3.611</td>
<td>3.777</td>
</tr>
<tr>
<td>3@1,2,4</td>
<td>0.7</td>
<td>1.351</td>
<td>1.700</td>
<td>1.821</td>
<td>1.808</td>
<td>1.919</td>
</tr>
<tr>
<td>5@1,4,6,15,6:15</td>
<td>0.6</td>
<td>3.257</td>
<td>3.484</td>
<td>3.563</td>
<td>3.575</td>
<td>3.698</td>
</tr>
<tr>
<td>5@1,4,5,6,15,6:15</td>
<td>0.5</td>
<td>3.237</td>
<td>3.515</td>
<td>3.595</td>
<td>3.622</td>
<td>3.777</td>
</tr>
<tr>
<td>5@1,4,6,15,6:15</td>
<td>0.5</td>
<td>3.203</td>
<td>3.541</td>
<td>3.646</td>
<td>3.631</td>
<td>3.757</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
> plot(close)
> plot(close, category = "nqtl")
```
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>64.5</td>
<td>72.1</td>
<td>6.692</td>
<td>103.611</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3.228</td>
<td>29.5</td>
<td>25.1</td>
<td>31.7</td>
<td>11.169</td>
<td>104.584</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>1.033</td>
<td>59.0</td>
<td>56.8</td>
<td>66.7</td>
<td>6.054</td>
<td>99.637</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>0.159</td>
<td>17.5</td>
<td>17.5</td>
<td>17.5</td>
<td>5.837</td>
<td>101.972</td>
</tr>
</tbody>
</table>

> plot(hpd)
```

2log(BF) scan with 50% HPD region
R/qtlbim: 2-D *(not 2-QTL)* scans

```r
> 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)
```

2-D plot of 2logBF: chr 6 & 15
1-D Slices of 2-D scans: chr 6 & 15

R/qtlbim: slice of epistasis

```r
> 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>18.1</td>
</tr>
<tr>
<td>c15</td>
<td>0.961</td>
<td>17.5</td>
<td>17.5</td>
<td>17.5</td>
<td>60.6</td>
</tr>
</tbody>
</table>

Cellmean of bp for AA, HA, AH, HH

<table>
<thead>
<tr>
<th>n.qtl</th>
<th>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>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>60.6</td>
</tr>
</tbody>
</table>

Estimate of bp for epistasis

<table>
<thead>
<tr>
<th>n.qtl</th>
<th>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>18.1</td>
</tr>
<tr>
<td>c15</td>
<td>0.961</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
examples in detail

- simulation study (after Stephens & Fisch (1998)) 2-3
- obesity in mice \((n = 421)\) 4-12
  - epistatic QTLs with no main effects
- expression phenotype (SCD1) in mice \((n = 108)\) 13-22
  - multiple QTL and epistasis
- mapping two correlated phenotypes 23-35
  - Jiang & Zeng 1995 paper
  - *Brassica napus* vernalization
- gonad shape in *Drosophila* spp. (insect) \((n = 1000)\) 36-42
  - multiple traits reduced by PC
  - many QTL and epistasis

**QTL 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%

<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>8</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>3371</td>
</tr>
<tr>
<td>9</td>
<td>3</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>751</td>
<td></td>
</tr>
<tr>
<td>7</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>377</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>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>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>198</td>
<td></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

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
posterior probability of effects

model selection for pairs
scatterplot estimates of epistatic loci

stronger epistatic effects

QTL 2: Data
Seattle SISG: Yandell © 2008
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, G6Pase, PPARgamma, PPARalpha, G6Pase, PDL...

Multiple Interval Mapping (QTLCart)
SCD1: multiple QTL plus epistasis!
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!

sub-peaks can be easily overlooked!
epistatic model fit

Cockerham epistatic effects

QTL 2: Data  Seattle SISG: Yandell © 2008

QTL 2: Data  Seattle SISG: Yandell © 2008
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  correlation only  both
Korol et al. (2001)
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

\[
\rho_P = 0.06, \rho_G = 0.68, \rho_E = -0.2
\]

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
  - \( \log(\text{days to flower}) \)
- 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.genoprobs(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)
```
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")
```
scatterplot adjusted for covariate

```r
## 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")
```
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 3—A plot of the first two principal components of the Fourier coefficients from geometric lobe outlines. Many individuals from each of five genotypic classes are represented. Each point represents an average of scores from the left and right sides of an individual except for those 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