Bayesian Model Selection for Multiple QTL

Brian S. Yandell
University of Wisconsin-Madison
www.stat.wisc.edu/~yandell/statgen
Jackson Laboratory, October 2007

outline

1. What is the goal of QTL study?
2. Bayesian vs. classical QTL study
3. Bayesian strategy for QTLs
4. model search using MCMC
   - Gibbs sampler and Metropolis-Hastings
5. model assessment
   - Bayes factors & model averaging
6. analysis of hyper data
7. software for Bayesian QTLs

1. what is the goal of QTL study?

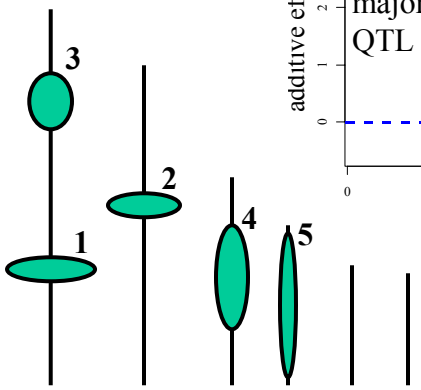
- uncover underlying biochemistry
  - identify how networks function, break down
  - find useful candidates for (medical) intervention
  - epistasis may play key role
  - statistical goal: maximize number of correctly identified QTL
- basic science/evolution
  - how is the genome organized?
  - identify 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: mimimize prediction error

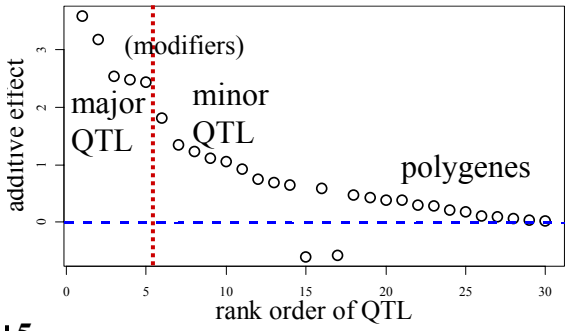
advantages of multiple QTL approach

- 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 = $(\text{bias})^2$ + variance

Pareto diagram of QTL effects

major QTL on linkage map





check QTL in context of genetic architecture

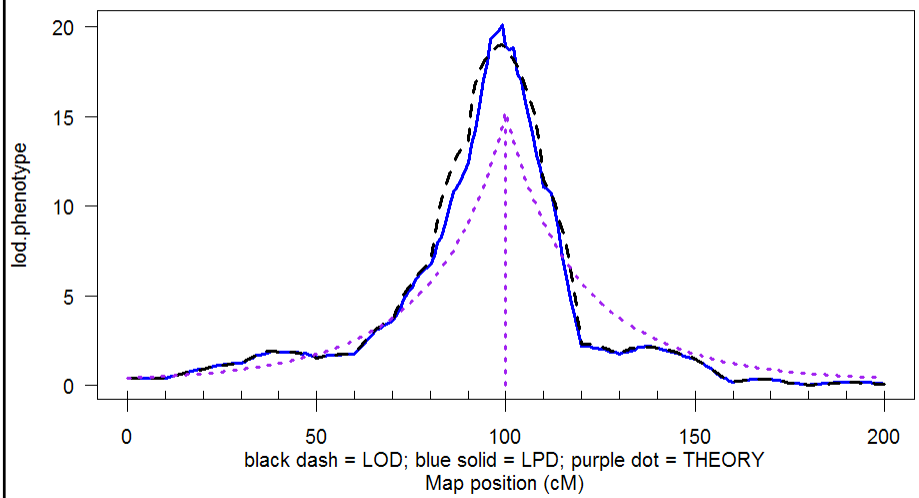
- scan for each QTL adjusting for all others
  - adjust for linked and unlinked QTL
    - adjust for linked QTL: reduce bias
    - adjust for unlinked QTL: reduce variance
  - adjust for environment/covariates
- examine entire genetic architecture
  - number and location of QTL, epistasis, GxE
  - model selection for best genetic architecture

2. 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

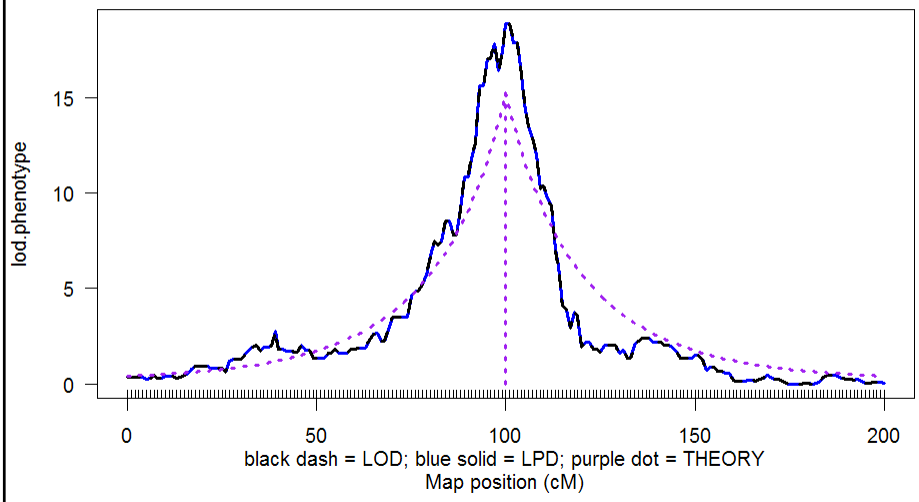
LOD & LPD: 1 QTL

n.ind = 100, 10 cM marker spacing



LOD & LPD: 1 QTL

n.ind = 100, 1 cM marker spacing

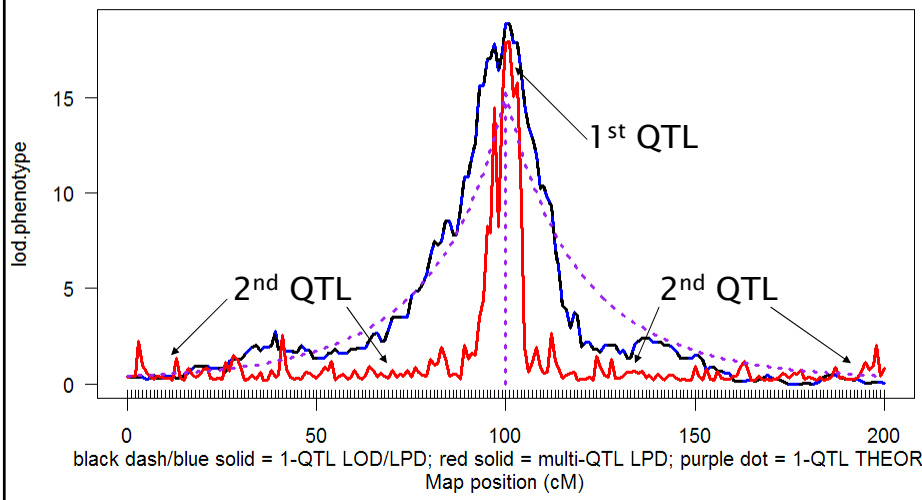


marginal LOD or LPD

- What is contribution of a QTL adjusting for all others?
  - improvement in LPD due to QTL at locus λ
  - contribution due to main effects, epistasis, GxE?
- How does adjusted LPD *differ* from unadjusted LPD?
  - raised by removing variance due to unlinked QTL
  - raised or lowered due to bias of linked QTL
  - analogous to Type III adjusted ANOVA tests
- can ask these same questions using classical LOD
  - see Broman's newer tools for multiple QTL inference

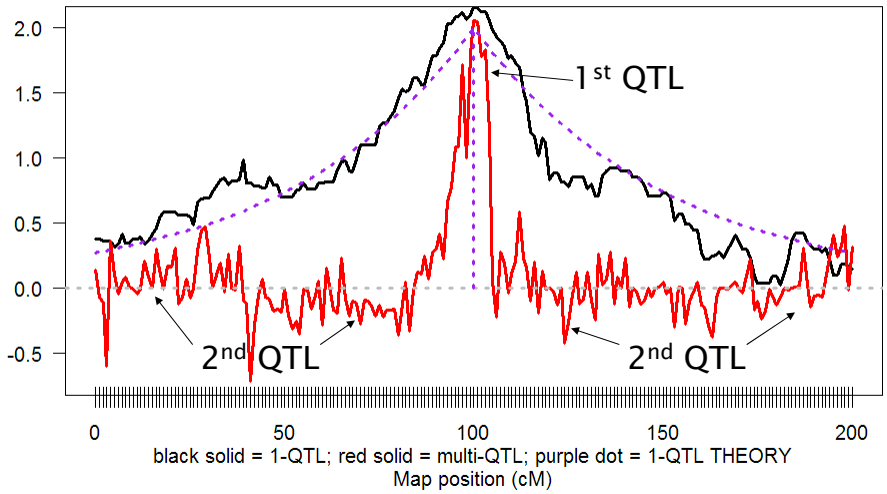
LPD: 1 QTL vs. multi-QTL

marginal contribution to LPD from QTL at λ



substitution effect: 1 QTL vs. multi-QTL

single QTL effect vs. marginal effect from QTL at λ

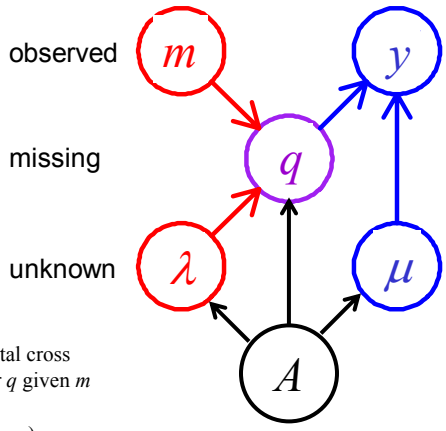


3. Bayesian strategy for QTLs

- augment data (y,m) with missing genotypes q
- build model for augmented data
  - genotypes (q) evaluated at loci (λ)
    - depends on flanking markers (m)
  - phenotypes (y) centered about effects (μ)
    - depends on missing genotypes (q)
  - λ and μ depend on genetic architecture (A)
    - How complicated is model? number of QTL, epistasis, etc.
- sample from model in some clever way
- infer most probable genetic architecture
  - estimate loci, their main effects and epistasis
  - study properties of estimates

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
  - λ = QT locus (or loci)
  - μ = phenotype model parameters
  - A = QTL model/genetic architecture
- pr($q|m,\lambda,A$) genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for q given m
- pr($y|q,\mu,A$) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters μ (could be non-parametric)



after
Sen Churchill (2001)

likelihood and posterior

- likelihood relates "known" data (y,m,q) to unknown values of interest (μ,λ,A)
  - $\text{pr}(y,q|m,\mu,\lambda,A) = \text{pr}(y|q,\mu,A)\ \text{pr}(q|m,\lambda,A)$
  - mix over unknown genotypes (q)
- posterior turns likelihood into a distribution
  - weight likelihood by priors
  - rescale to sum to 1.0
  - posterior = likelihood * prior / constant

likelihood and posterior

$$\text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{constant}} \ : \ \text{Bayes' rule}$$

$$\text{pr}(\mu,\lambda,A \mid y,m) = \frac{\text{pr}(y \mid m,\mu,\lambda,A) * \text{pr}(\mu \mid A)\text{pr}(\lambda \mid m,A)\text{pr}(A)}{\text{pr}(y \mid m)}$$

likelihood mixes over missing QTL genotypes:

$$\text{pr}(y \mid m,\mu,\lambda) = \sum_q \text{pr}(y \mid q,\mu)\text{pr}(q \mid m,\lambda)$$

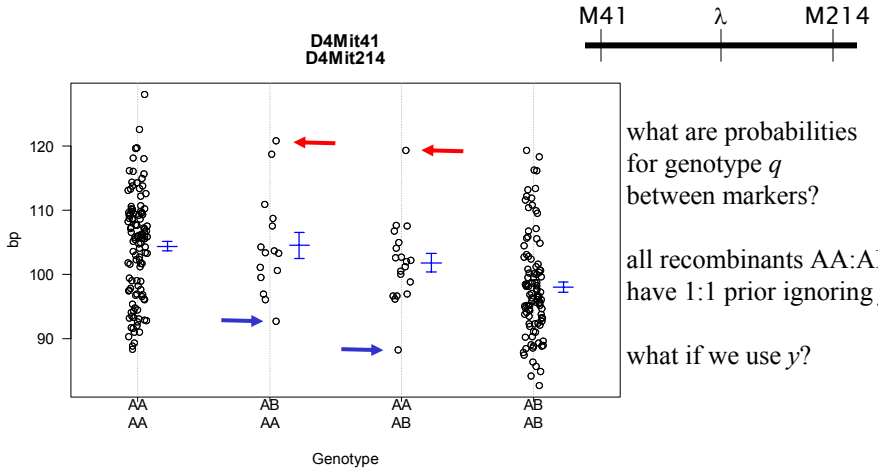
Bayes posterior vs. maximum likelihood
(genetic architecture A = single QTL at λ)

- LOD: classical Log ODds
  - maximize likelihood over effects μ
  - `R/qtl scanone/scantwo: method = "em"`
- *LPD*: Bayesian *L*og *P*osterior *D*ensity
  - average posterior over effects μ
  - `R/qtl scanone/scantwo: method = "imp"`

$$\mathrm{LOD}(\lambda) = \log_{10}\left(\max_{\mu} \mathrm{pr}(y \mid m, \mu, \lambda)\right)$$

$$\mathrm{LPD}(\lambda) = \log_{10}\left(\mathrm{pr}(\lambda \mid m)\sum_{\mu} \mathrm{pr}(y \mid m, \mu, \lambda)\mathrm{pr}(\mu)\right)$$

do phenotypes help to guess genotypes?
posterior on QTL genotypes q



what are probabilities for genotype q between markers?

all recombinants AA:AB have 1:1 prior ignoring y

what if we use y?

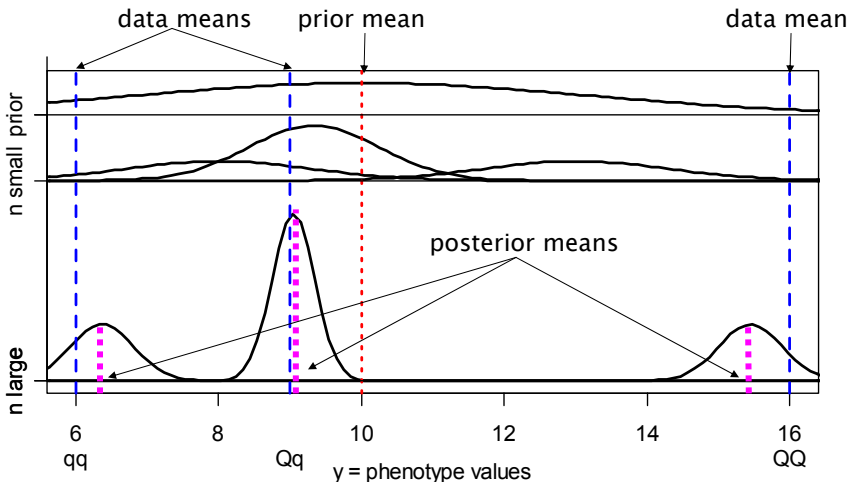
posterior on QTL genotypes q

- 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($y|q,\mu$)
    - weight toward q with similar phenotype values
  - posterior balances these two
- this *is* the E-step of EM computations

$$\text{pr}(q \mid y, m, \mu, \lambda) = \frac{\text{pr}(y \mid q, \mu) * \text{pr}(q \mid m, \lambda)}{\text{pr}(y \mid m, \mu, \lambda)}$$

where are the genotypic means?

(phenotype mean for genotype q is μ_q)



prior & posteriors: genotypic means μ_q

- prior for genotypic means
  - centered at grand mean
  - variance related to heritability of effect
    - hyper-prior on variance (details omitted)
- posterior
  - shrink genotypic means toward grand mean
  - shrink variance of genotypic mean

prior: $E(\mu_q) = \bar{y}_\bullet$ $V(\mu_q) = V(y)h_q^2$

posterior: $E(\mu_q \mid y) = \bar{y}_\bullet(1-b_q) + \bar{y}_q b_q$ $V(\mu_q \mid y) = V(\bar{y}_q)b_q$

shrinkage: $b_q = 1 - \dfrac{V(\bar{y}_q)}{V(\bar{y}_q) + V(y)h_q^2} \approx 1$

multiple QTL phenotype model

- phenotype affected by genotype & environment

  $\mathrm{E}(y|q) = \mu_q = \beta_0 + \mathrm{sum}_{\{j \text{ in } H\}}\, \beta_j(q)$

  number of terms in QTL model $H \le 2^{nqtl}$ (3^{nqtl} for F_2)
- partition genotypic mean into QTL effects

  $\mu_q = \beta_0 + \beta_1(q_1) + \beta_2(q_2) + \beta_{12}(q_1,q_2)$

  μ_q = mean + main effects + epistatic interactions
- partition prior and posterior (details omitted)

Where are the loci λ 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

  $\Pr(\lambda \mid m,q) = \Pr(\lambda) \Pr(q \mid m,\lambda) / \text{constant}$
  - constant determined by averaging
    - over all possible genotypes q
    - over all possible loci λ on entire map
- no easy way to write down posterior

What is the genetic architecture A?

- components of genetic architecture
  - how many QTL?
  - where are loci (λ)? how large are effects (μ)?
  - which pairs of QTL are epistatic?
- use priors to weight posterior
  - toward guess from previous analysis
  - improve efficiency of sampling from posterior
    - increase samples from architectures of interest

4. QTL Model Search using MCMC

- 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 λ given q,H (using Metropolis-Hastings step)
  - sample genotypes q given λ,μ,y,H (using Gibbs sampler)
  - sample effects μ given q,y,H (using Gibbs sampler)
  - sample QTL model H given λ,μ,y,q (using Gibbs or M-H)

$$(\lambda,q,\mu,H) \sim \text{pr}(\lambda,q,\mu,H \mid y,m)$$

$$(\lambda,q,\mu,H)_1 \rightarrow (\lambda,q,\mu,H)_2 \rightarrow \cdots \rightarrow (\lambda,q,\mu,H)_N$$

Gibbs sampler idea

- toy problem
  - want to study two correlated effects
  - could sample directly from their bivariate distribution
- instead use Gibbs sampler:
  - sample each effect from its full conditional given the other
  - pick order of sampling at random
  - repeat many times

$$\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} \sim N\left(\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$

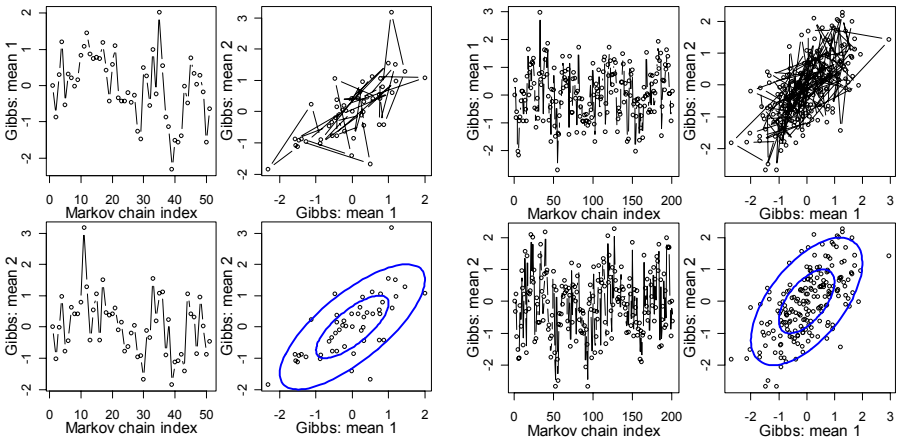
$$\mu_1 \sim N\left(\rho\mu_2, 1-\rho^2\right)$$

$$\mu_2 \sim N\left(\rho\mu_1, 1-\rho^2\right)$$

Gibbs sampler samples: $\rho = 0.6$

$N = 50$ samples

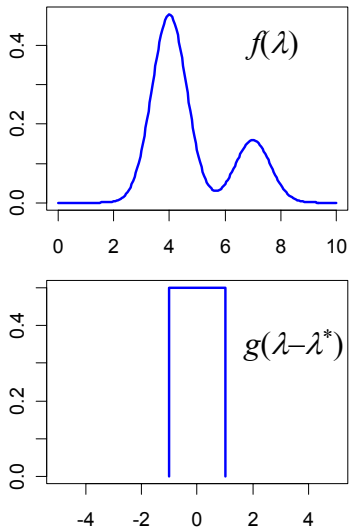
$N = 200$ samples



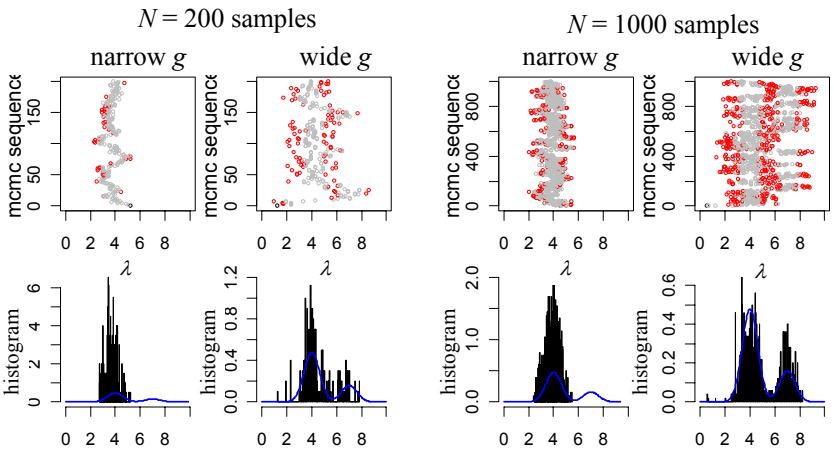
Metropolis-Hastings idea

- 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 λ^*
    - near (?) current value λ
    - from some distribution g
  - accept new value with prob a
    - Gibbs sampler: $a = 1$ always

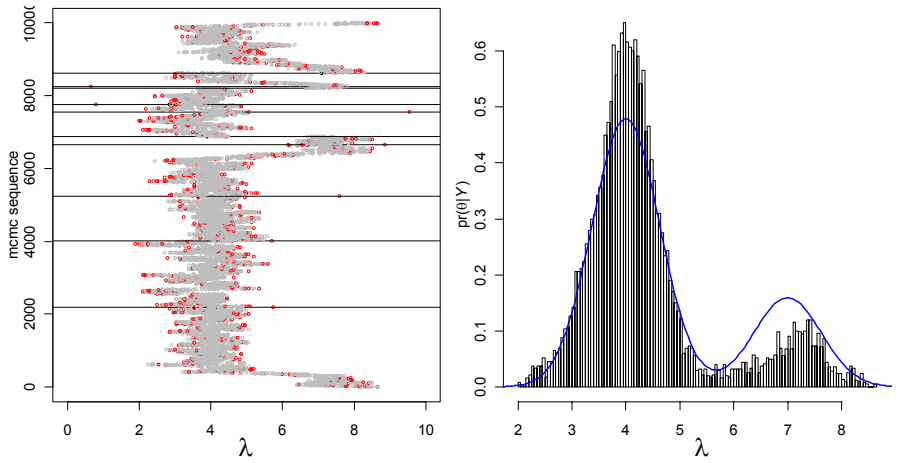
$$a = \min\left(1, \frac{f(\lambda^*)g(\lambda^* - \lambda)}{f(\lambda)g(\lambda - \lambda^*)}\right)$$



Metropolis-Hastings samples



MCMC realization



added twist: occasionally propose from whole domain

sampling across QTL models *H*

action steps: draw one of three choices

- update QTL model *H* with probability 1-*b*(*H*)-*d*(*H*)
  - update current model using full conditionals
  - sample QTL loci, effects, and genotypes
- add a locus with probability *b*(*H*)
  - 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*(*H*)
  - propose dropping one of existing loci
  - decide whether to accept the "death" of locus

reversible jump MCMC

- consider known genotypes *q* at 2 known loci λ
  - 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*

$$nqtl = 1 : Y = \beta_0 + \beta_1(q_1) + e$$

$$nqtl = 2 : Y = \beta_0 + \beta_1(q_1) + \beta_2(q_2) + e$$

Gibbs sampler with loci indicators

- partition genome into intervals
  - at most one QTL per interval
  - interval = 1 cM in length
  - assume QTL in middle of interval
- use loci to indicate presence/absence of QTL in each interval
  - $\gamma = 1$ if QTL in interval
  - $\gamma = 0$ if no QTL
- Gibbs sampler on loci indicators
  - see work of Nengjun Yi (and earlier work of Ina Hoeschele)

$$Y = \beta_0 + \gamma_1 \beta_1(q_1) + \gamma_2 \beta_2(q_1) + e$$

Bayesian shrinkage estimation

- soft loci indicators
  - strength of evidence for λ_j depends on variance of β_j
  - similar to $\gamma > 0$ on grey scale
- include all possible loci in model
  - pseudo-markers at 1cM intervals
- Wang et al. (2005 *Genetics*)
  - Shizhong Xu group at U CA Riverside

$$Y = \beta_0 + \beta_1(q_1) + \beta_2(q_1) + ... + e$$

$$\beta_j(q_j) \sim N(0, \sigma_j^2), \sigma_j^2 \sim \text{inverse - chisquare}$$

epistatic interactions

- model space issues
  - 2-QTL interactions only?
  - Fisher-Cockerham partition vs. tree-structured?
  - general interactions among multiple QTL
- 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?
- Yi et al. (2005, 2007)

5. Model Assessment

- balance model fit against model complexity

	smaller model	bigger model
model fit	miss key features	fits better
prediction	may be biased	no bias
interpretation	easier	more complicated
parameters	low variance	high variance

- information criteria: penalize likelihood by model size
  - compare IC = – 2 log L(model | data) + penalty(model size)
- Bayes factors: balance posterior by prior choice
  - compare pr(data | model)

Bayes factors

- ratio of model likelihoods
  - ratio of posterior to prior odds for architectures
  - average over unknown effects (μ) and loci (λ)

$$BF = \frac{\text{pr}(\text{data} \mid \text{model } A_1)}{\text{pr}(\text{data} \mid \text{model } A_2)}$$

- roughly equivalent to BIC
  - BIC maximizes over unknowns
  - BF averages over unknowns

$$2\log_{10}(BF) = 2LOD + (\text{change in model size})\log_{10}(n)$$

issues in computing Bayes factors

- *BF* insensitive to shape of prior on A
  - geometric, Poisson, uniform
  - precision improves when prior mimics posterior
- *BF* sensitivity to prior variance on effects θ
  - 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
  - apply Bayes' rule and solve for pr($y \mid m, A$)
    - pr($A \mid y, m$) = pr($y \mid m, A$) pr($A \mid m$) / constant
    - pr(data|model) = constant * pr(model|data) / pr(model)
  - posterior pr($A \mid y, m$) is marginal histogram

marginal BF scan by QTL

- compare models with and without QTL at λ
  - average over all possible models
  - estimate as ratio of samples with/without QTL
- scan over genome for peaks
  - 2log(BF) seems to have similar properties to LPD

$$BF_\lambda = \frac{\text{pr}(y \mid m, \text{model with } \lambda)}{\text{pr}(y \mid m, \text{model without } \lambda)}$$

Bayesian model averaging

- average summaries over multiple architectures
- avoid selection of "best" model
- focus on "better" models
- examples in data talk later

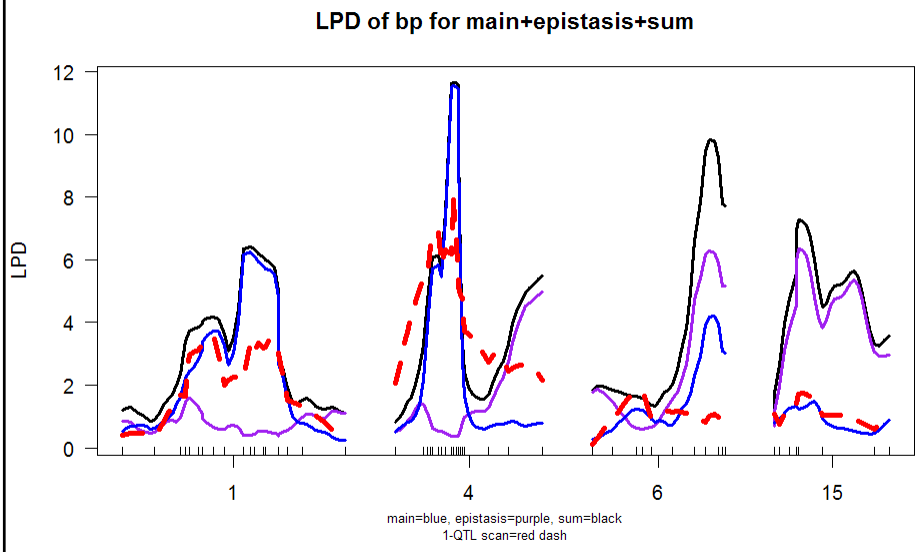
6. analysis of hyper data

- marginal scans of genome
  - detect significant loci
  - infer main and epistatic QTL, GxE
- infer most probable genetic architecture
  - number of QTL
  - chromosome pattern of QTL with epistasis
- diagnostic summaries
  - heritability, unexplained variation

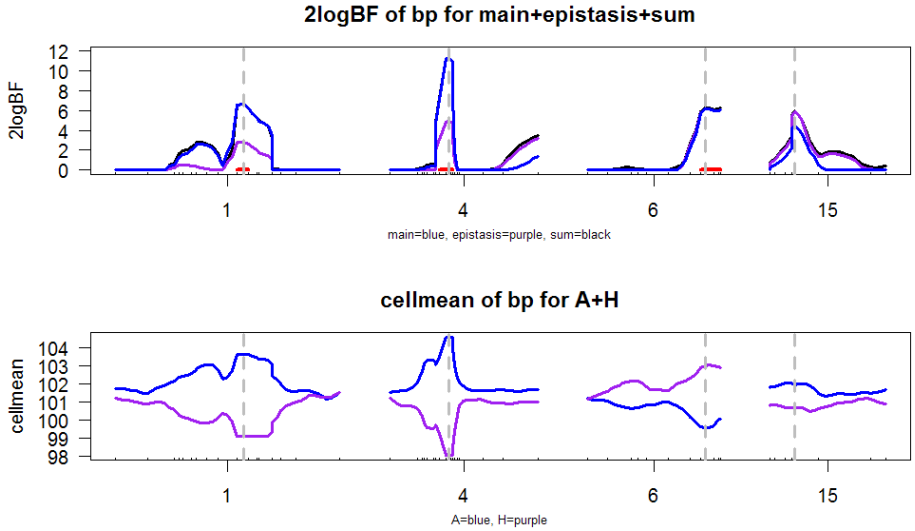
marginal scans of genome

- LPD and 2log(BF) “tests” for each locus
- estimates of QTL effects at each locus
- separately infer main effects and epistasis
  - main effect for each locus (blue)
  - epistasis for loci paired with another (purple)
    - identify epistatic QTL in 1-D scan
    - infer pairing in 2-D scan

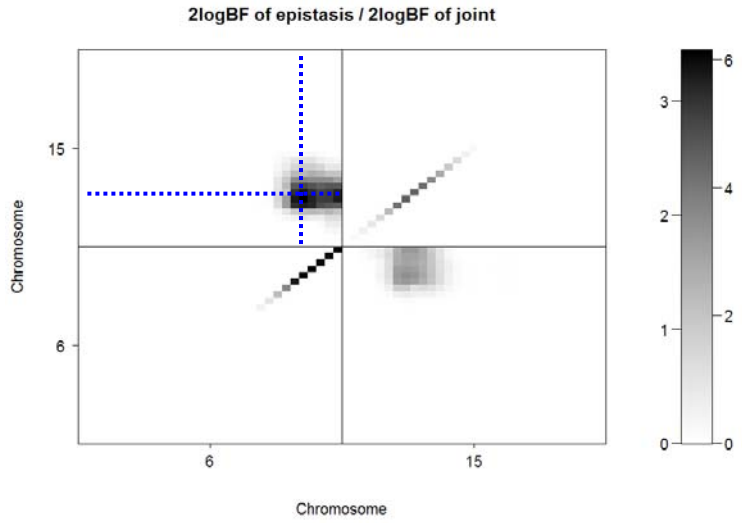
hyper data: scanone



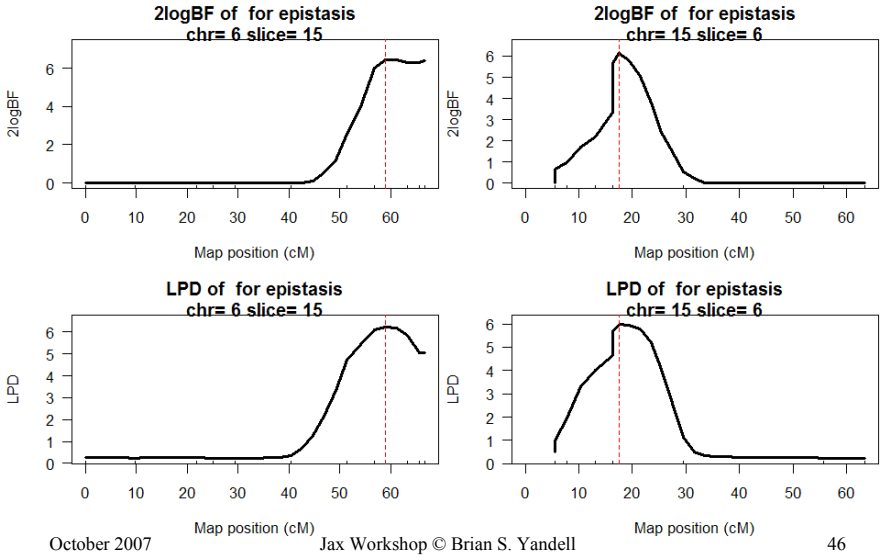
2log(BF) scan with 50% HPD region



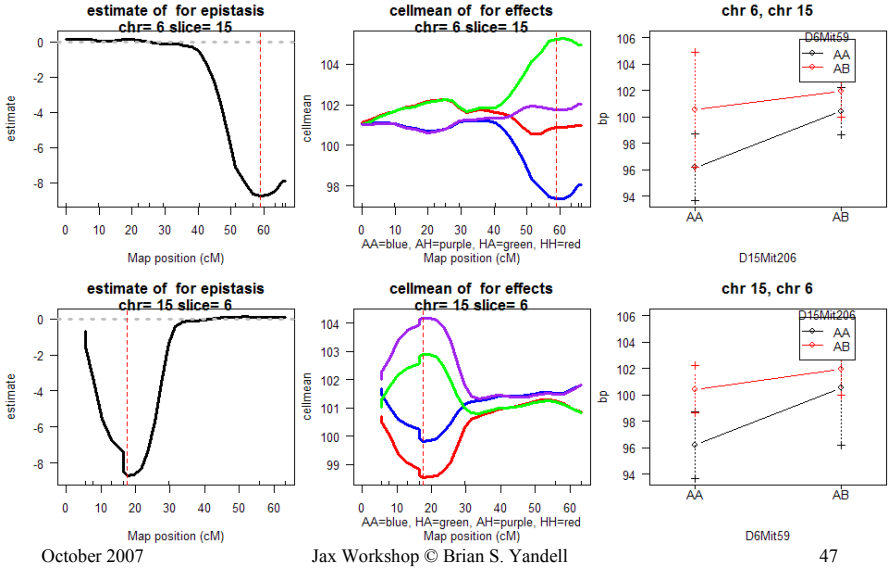
2-D plot of 2logBF: chr 6 & 15



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



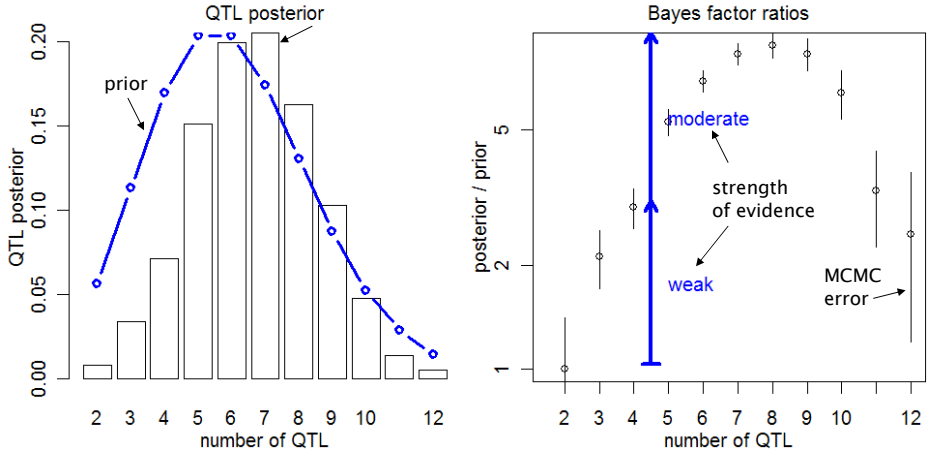
1-D Slices of 2-D scans: chr 6 & 15



What is best genetic architecture?

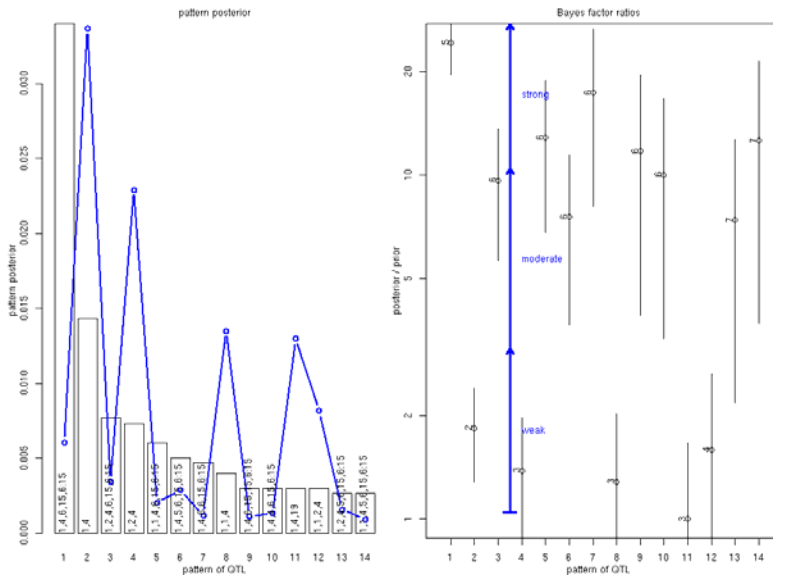
- How many QTL?
- What is pattern across chromosomes?
- examine posterior relative to prior
  - prior determined ahead of time
  - posterior estimated by histogram/bar chart
  - Bayes factor ratio = pr(model|data) / pr(model)

How many QTL?
posterior, prior, Bayes factor ratios

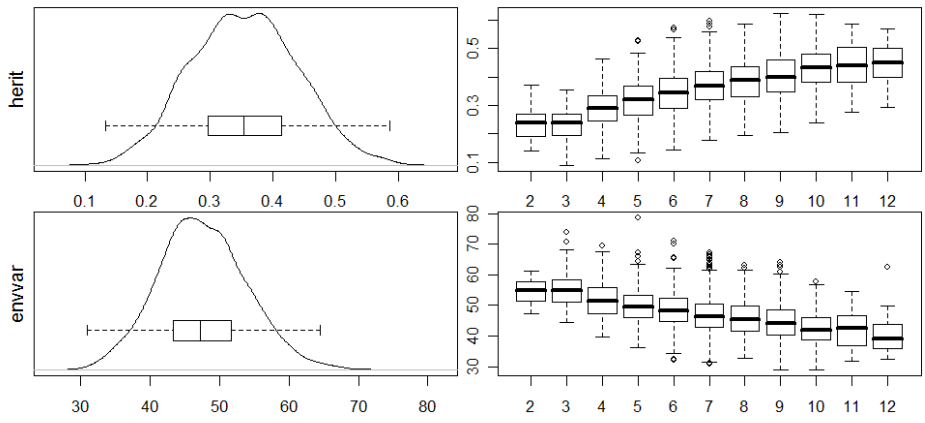


most probable patterns

	nqtl	posterior	prior	bf	bfse
1,4,6,15,6:15	5	0.03400	2.71e-05	24.30	2.360
1,4,6,6,15,6:15	6	0.00467	5.22e-06	17.40	4.630
1,1,4,6,15,6:15	6	0.00600	9.05e-06	12.80	3.020
1,1,4,5,6,15,6:15	7	0.00267	4.11e-06	12.60	4.450
1,4,6,15,15,6:15	6	0.00300	4.96e-06	11.70	3.910
1,4,4,6,15,6:15	6	0.00300	5.81e-06	10.00	3.330
1,2,4,6,15,6:15	6	0.00767	1.54e-05	9.66	2.010
1,4,5,6,15,6:15	6	0.00500	1.28e-05	7.56	1.950
1,2,4,5,6,15,6:15	7	0.00267	6.98e-06	7.41	2.620
1,4	2	0.01430	1.51e-04	1.84	0.279
1,1,2,4	4	0.00300	3.66e-05	1.59	0.529
1,2,4	3	0.00733	1.03e-04	1.38	0.294
1,1,4	3	0.00400	6.05e-05	1.28	0.370
1,4,19	3	0.00300	5.82e-05	1.00	0.333



diagnostic summaries



7. Software for Bayesian QTLs R/qtlbim

- publication
  - CRAN release Fall 2006
  - Yandell et al. (2007 *Bioinformatics*)
- properties
  - cross-compatible with R/qtl
  - epistasis, fixed & random covariates, GxE
  - extensive graphics

R/qtlbim: software history

- Bayesian module within WinQTLCart
  - WinQTLCart output can be processed using R/bim
- Software history
  - initially designed (Satagopan Yandell 1996)
  - major revision and extension (Gaffney 2001)
  - R/bim to CRAN (Wu, Gaffney, Jin, Yandell 2003)
  - R/qtlbim total rewrite (Yandell et al. 2007)

other Bayesian software for QTLs

- R/bim*: Bayesian Interval Mapping
  - Satagopan Yandell (1996; Gaffney 2001) CRAN
  - no epistasis; reversible jump MCMC algorithm
  - version available within WinQTLCart (statgen.ncsu.edu/qtlcart)
- R/qtl*
  - Broman et al. (2003 Bioinformatics) CRAN
  - multiple imputation algorithm for 1, 2 QTL scans & limited mult-QTL fits
- Bayesian QTL / Multimapper
  - Sillanpää Arjas (1998 Genetics) www.rni.helsinki.fi/~mjs
  - no epistasis; introduced posterior intensity for QTLs
- (no released code)
  - Stephens & Fisch (1998 Biometrics)
  - no epistasis
- R/bqtl
  - C Berry (1998 TR) CRAN
  - no epistasis, Haley Knott approximation

* Jackson Labs (Hao Wu, Randy von Smith) provided crucial technical support

many thanks

Karl Broman
Jackson Labs
Gary Churchill
Hao Wu
Randy von Smith
U AL Birmingham
David Allison
Nengjun Yi
Tapan Mehta
Samprit Banerjee
Ram Venkataraman
Daniel Shriner
Tom Osborn
David Butruille
Marcio Ferrera
Josh Udahl
Pablo Quijada
Alan Attie
Jonathan Stoehr
Hong Lan
Susie Clee
Jessica Byers
Mark Keller
Michael Newton
Hyuna Yang
Daniel Sorensen
Daniel Gianola
Liang Li
my students
Jaya Satagopan
Fei Zou
Patrick Gaffney
Chunfang Jin
Elias Chaibub
W Whipple Neely
Jee Young Moon
USDA Hatch, NIH/NIDDK (Attie), NIH/R01 (Yi, Broman)