

# Bayesian Model Selection for Multiple QTL

Brian S. Yandell

University of Wisconsin-Madison

[www.stat.wisc.edu/~yandell/statgen](http://www.stat.wisc.edu/~yandell/statgen)↑

Jackson Laboratory, September 2006

## outline

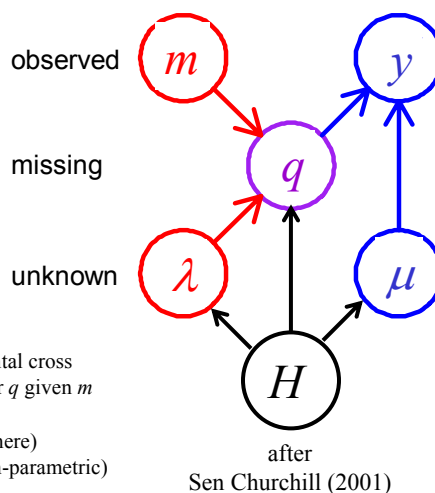
1. Bayesian vs. classical QTL study
2. Bayesian priors & posteriors
3. model search using MCMC
  - Gibbs sampler and Metropolis-Hastings
4. model assessment
  - Bayes factors & model averaging
5. data examples in detail
  - simulation & hyper data

# 1. Bayesian vs. classical QTL study

- classical study
  - maximize likelihood over unknowns
  - test for presence/absence of QTL at loci
  - model selection in stepwise fashion
- Bayesian study
  - sample unknowns from posterior
  - estimate QTL loci directly
  - sample simultaneously across models

## Bayesian QTL: key players

- observed measurements
  - $y$  = phenotypic trait
  - $m$  = markers & linkage map
  - $i$  = individual index ( $1, \dots, n$ )
- 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
  - $H$  = QTL model/genetic architecture
- $\text{pr}(q|m, \lambda, H)$  genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for  $q$  given  $m$
- $\text{pr}(y|q, \mu, H)$  phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters  $\mu$  (could be non-parametric)



## Bayes posterior vs. maximum likelihood

- *LOD*: classical *Log Odds*
  - maximizes likelihood
    - mixture over missing QTL genotypes  $q$
    - maximize phenotype model parameters  $\mu$
    - scan over possible loci  $\lambda$
  - R/qtl scanone/scantwo: method = "em"
- *LPD*: Bayesian *Log Posterior Density*
  - averages over unknowns
    - average over missing QTL genotypes  $q$
    - average phenotype model parameters  $\mu$
    - scan over possible loci  $\lambda$
  - R/qtl scanone/scantwo: method = "imp"

## Bayes posterior vs. maximum likelihood

- *LOD*: classical *Log Odds*
  - maximizes likelihood
  - R/qtl scanone/scantwo: method = "em"
- *LPD*: Bayesian *Log Posterior Density*
  - averages over unknowns
  - R/qtl scanone/scantwo: method = "imp"
- suppose genetic architecture is known
  - $H = 1$  QTL or 2 QTL model
  - available in R/qtl via scanone and scantwo routines

$$\text{LOD}(\lambda) = \log_{10} \{ \max_{\mu} \text{pr}(y | m, \mu, \lambda) \} + c$$

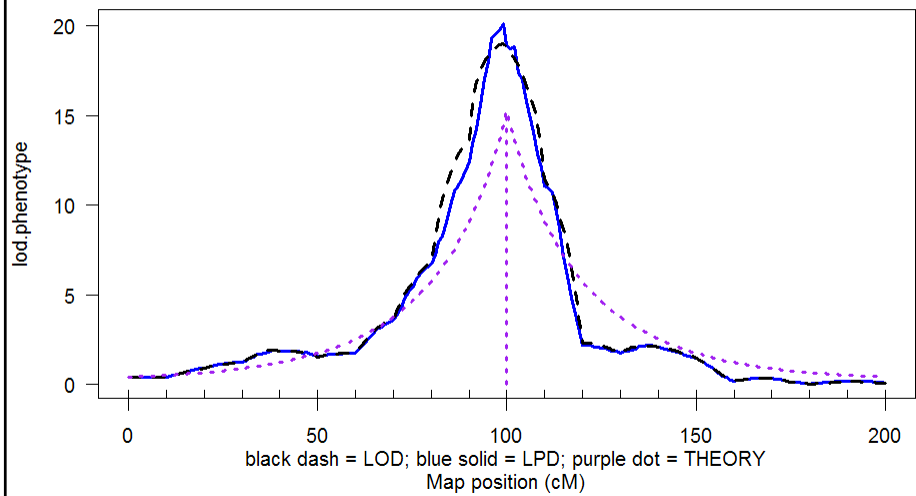
$$\text{LPD}(\lambda) = \log_{10} \{ \text{pr}(\lambda | m) \int \text{pr}(y | m, \mu, \lambda) \text{pr}(\mu) d\mu \} + C$$

with mixture over missing QTL genotypes:

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

# LOD & LPD: 1 QTL

n.ind = 100, 10 cM marker spacing



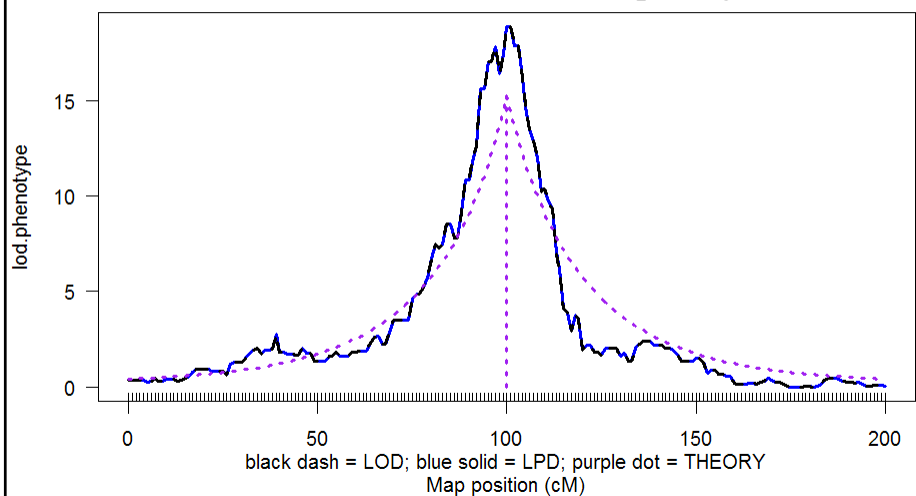
September 2006

Jax Workshop © Brian S. Yandell

7

# LOD & LPD: 1 QTL

n.ind = 100, 1 cM marker spacing



September 2006

Jax Workshop © Brian S. Yandell

8

# Bayesian strategy for QTL study

- augment data  $(y, m)$  with missing genotypes  $q$
- study unknowns  $(\mu, \lambda, H)$  given augmented data  $(y, m, q)$ 
  - find better genetic architectures  $H$
  - 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) (Yi et al. 2005)

$$\text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{constant}}$$

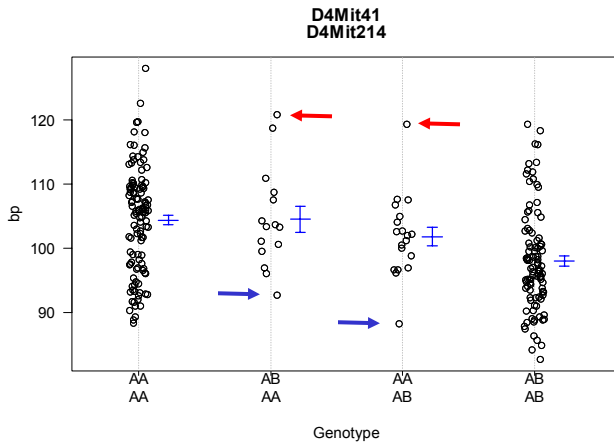
$$\text{posterior for } q, \mu, \lambda, H = \frac{\text{phenotype likelihood} * [\text{prior for } q, \mu, \lambda, H]}{\text{constant}}$$

$$\text{pr}(q, \mu, \lambda, H | y, m) = \frac{\text{pr}(y | q, \mu, H) * [\text{pr}(q | m, \lambda, H) \text{pr}(\mu | H) \text{pr}(\lambda | m, H) \text{pr}(H)]}{\text{pr}(y | m)}$$

## 2. 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 model  $H$ 
  - 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

what are likely QTL genotypes  $q$ ?  
 how does phenotype  $y$  improve guess?



what are probabilities  
 for genotype  $q$   
 between markers?

recombinants AA:AB

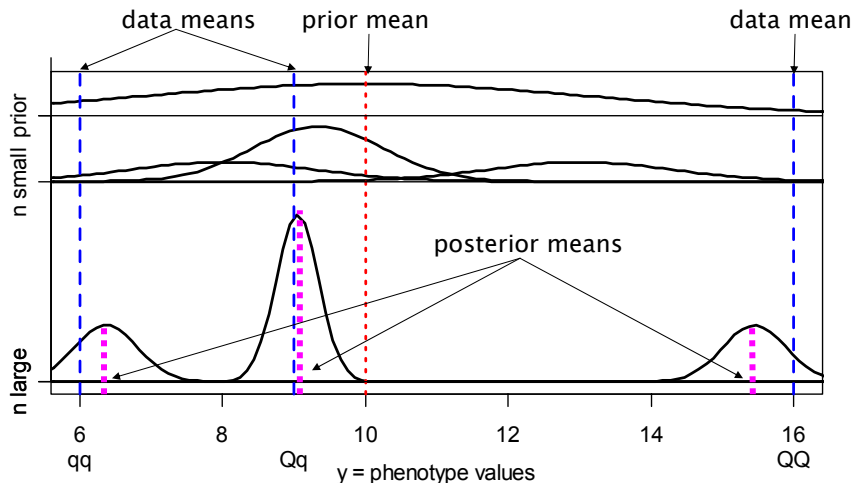
all 1:1 if ignore  $y$   
 and if we use  $y$ ?

## posterior on QTL genotypes $q$

- full conditional of  $q$  given data, parameters
  - proportional to prior  $\text{pr}(q | m, \lambda)$ 
    - weight toward  $q$  that agrees with flanking markers
  - proportional to likelihood  $\text{pr}(y|q, \mu)$ 
    - weight toward  $q$  with similar phenotype values
  - posterior recombination model balances these two
- this *is* the E-step of EM computations

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

## what values are the genotypic means? (phenotype mean for genotype $q$ is $\mu_q$ )



September 2006

Jax Workshop © Brian S. Yandell

13

## prior & posteriors: genotypic means $\mu_q$

- 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

$$\text{prior:} \quad E(\mu_q) = \bar{y}. \quad V(\mu_q) = V(y)h_q^2$$

$$\text{posterior:} \quad E(\mu_q | y) = \bar{y} \cdot (1 - b_q) + \bar{y}_q b_q \quad V(\mu_q | y) = V(\bar{y}_q) b_q$$

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

September 2006

Jax Workshop © Brian S. Yandell

14

## multiple QTL phenotype model

- phenotype affected by genotype & environment

$$E(y|q) = \mu_q = \beta_0 + \sum_{j \text{ in } H} \beta_j(q)$$

number of terms in QTL model  $H \leq 2^{n_{qtl}}$  ( $3^{n_{qtl}}$  for  $F_2$ )

- partition genotypic mean into QTL effects

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

$\mu_q$  = mean + main effects + epistatic interactions

- partition prior and posterior (details omitted)

## 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  $q$

$$\text{pr}(\lambda | m, q) = \text{pr}(\lambda) \text{pr}(q | m, \lambda) / \text{constant}$$

– 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 $H$ ?

- 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

## 3. 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  $\lambda$  given  $q, H$  (using Metropolis-Hastings step)
  - sample genotypes  $q$  given  $\lambda, \mu, y, H$  (using Gibbs sampler)
  - sample effects  $\mu$  given  $q, y, H$  (using Gibbs sampler)
  - sample QTL model  $H$  given  $\lambda, \mu, y, q$  (using Gibbs or M-H)

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

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

# 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

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

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

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

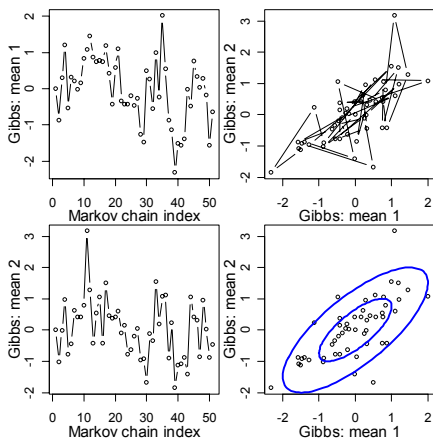
September 2006

Jax Workshop © Brian S. Yandell

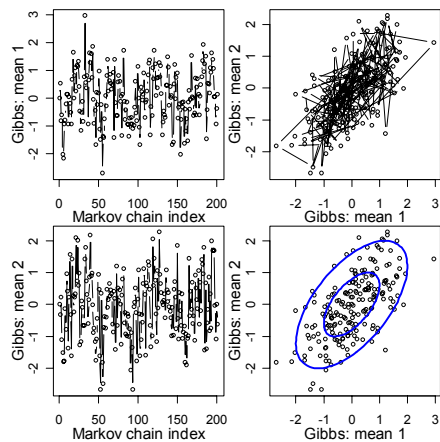
19

## Gibbs sampler samples: $\rho = 0.6$

$N = 50$  samples



$N = 200$  samples



September 2006

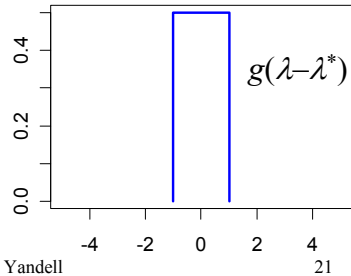
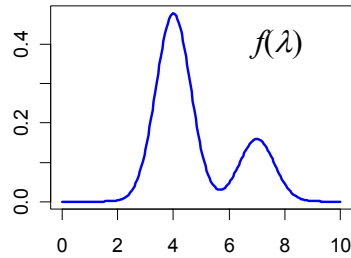
Jax Workshop © Brian S. Yandell

20

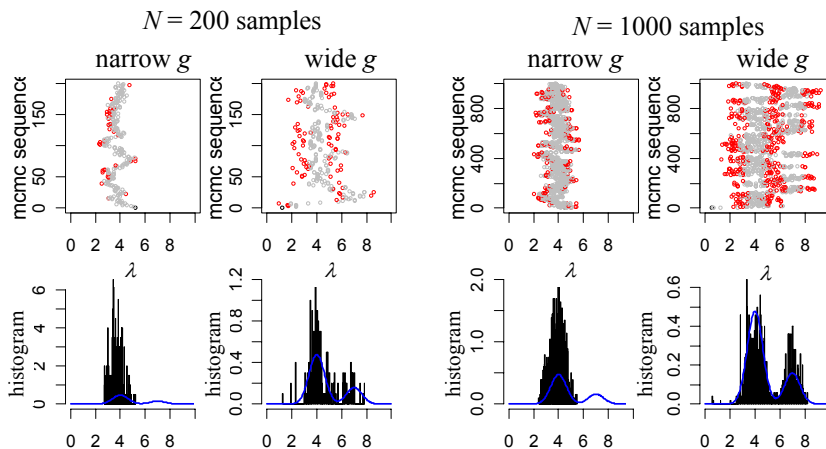
# Metropolis-Hastings idea

- 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

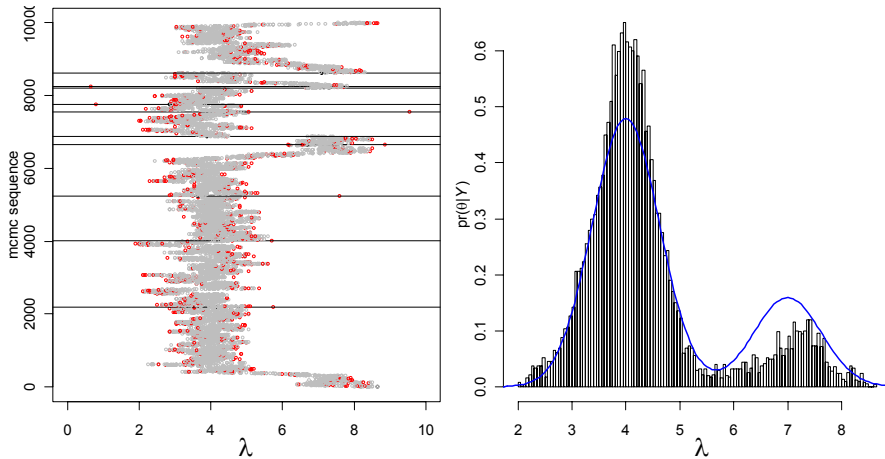
$$a = \min\left(1, \frac{f(\lambda^*)g(\lambda - \lambda^*)}{f(\lambda)g(\lambda^* - \lambda)}\right)$$



# Metropolis-Hastings samples



# MCMC realization



added twist: occasionally propose from whole domain

# sampling across QTL models $H$

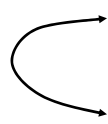


action steps: draw one of three choices

- 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  $\lambda$ 
  - 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  $H$


$$\begin{aligned}nqtl = 1 : Y &= \beta_0 + \beta_1(q_1) + e \\nqtl = 2 : Y &= \beta_0 + \beta_1(q_1) + \beta_2(q_2) + e\end{aligned}$$

## 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  $\lambda_j$  depends on variance of  $\beta_j$
  - 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

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

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

# 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 Xu (2000) *Genetics*; Yi, Xu, Allison (2003) *Genetics*; Yi (2004)

## 4. 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  $L$  by model size  $|H|$ 
  - compare  $IC = -2 \log L(H | y) + \text{penalty}(H)$
- Bayes factors: balance posterior by prior choice
  - compare  $\text{pr}(\text{data } y | \text{model } H)$

## Bayes factors and BIC

- Bayesian interpretation
  - $\text{pr}(\text{data} | \text{model}) = \text{pr}(\text{model} | \text{data}) / \text{pr}(\text{model})$
  - $\text{pr}(\text{data} | \text{model}) = \text{model posterior} / \text{model prior}$
  - marginal model averaged over all parameters
- Bayes Information Criteria
  - $BIC = 2\log(\text{likelihood}) + \text{d.f.} * \log(n.\text{ind})$
  - downweight data likelihood by complexity
  - complexity penalty matches Bayesian idea

## Bayes factors and BIC

- Bayes factor ( $BF$ ) for model comparison
  - ratio of  $\text{pr}(\text{data} \mid \text{model})$  for 2 models
  - often reported as  $2\log(BF)$
  - weak/moderate/strong evidence: 3/10/30
- $BIC$  comparison
  - difference of two  $BIC$  values
  - same as  $LR$  statistic with penalty when
    - comparing two nested models
    - simple hypotheses (e.g. 1 vs 2 QTL)
- $BF = BIC$  comparison for nested models

## QTL Bayes factors

- use to compare genetic architectures
  - $n.qtl$  = number of QTL
  - pattern of QTL across chromosomes
  - epistatic pairs
- can compare nested or non-nested architectures
  - 1 vs. 2 QTL
  - QTL on chr 1,2,5 vs. QTL on chr 2,3,4

$$BF = \frac{\text{pr}(\text{data} \mid n.qtl)}{\text{pr}(\text{data} \mid n.qtl + 1)} = \frac{\text{pr}(n.qtl \mid \text{data})/\text{pr}(n.qtl)}{\text{pr}(n.qtl + 1 \mid \text{data})/\text{pr}(n.qtl + 1)}$$



## marginal LOD or LPD

- compare two architectures at locus
  - with ( $H_2$ ) or without ( $H_1$ ) QTL at  $\lambda_2$ 
    - preserve model hierarchy (e.g. drop any epistasis with QTL at  $\lambda_2$ )
  - with ( $H_2$ ) or without ( $H_1$ ) epistasis at  $\lambda_2$
  - allow for QTL at all other loci  $\lambda_1$  in architecture  $H_1$
- use marginal LPD or other diagnostic
  - posterior, Bayes factor, heritability

$$\text{LOD}(\lambda_1, \lambda_2 | H_2) - \text{LOD}(\lambda_1 | H_1)$$

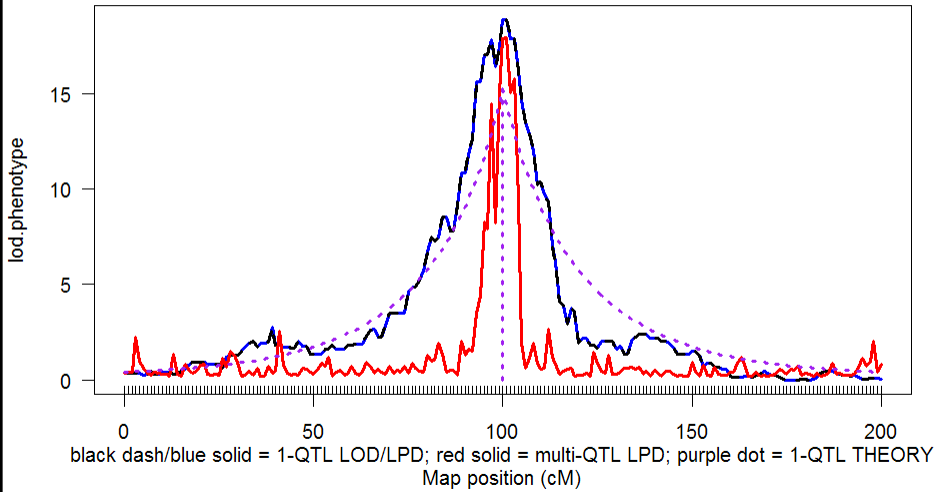
$$\text{LPD}(\lambda_1, \lambda_2 | H_2) - \text{LPD}(\lambda_1 | H_1)$$

## 5. simulations and data analyses

- revisit 1 QTL simulation
  - refining position by marginal scan
    - single QTL vs. marginal on multi-QTL
    - $2\log(\text{BF})$
  - substitution effect: 1-QTL vs. multi-QTL
- R/ctl hyper dataset (Sugiyama *et al.* 2001)
  - higher LPD with multi-QTL
  - detecting epistasis and linked QTL

## LPD: 1 QTL vs. multi-QTL

marginal contribution to LPD from QTL at  $\lambda$



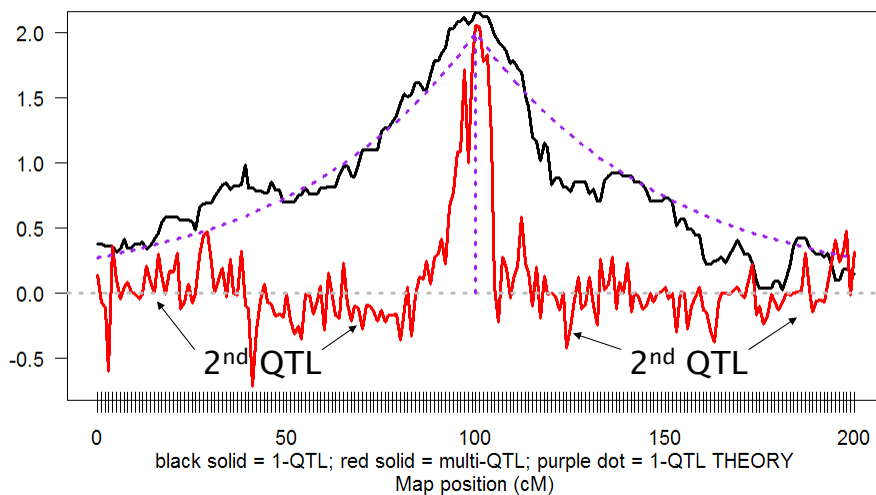
September 2006

Jax Workshop © Brian S. Yandell

35

## substitution effect: 1 QTL vs. multi-QTL

single QTL effect vs. marginal effect from QTL at  $\lambda$



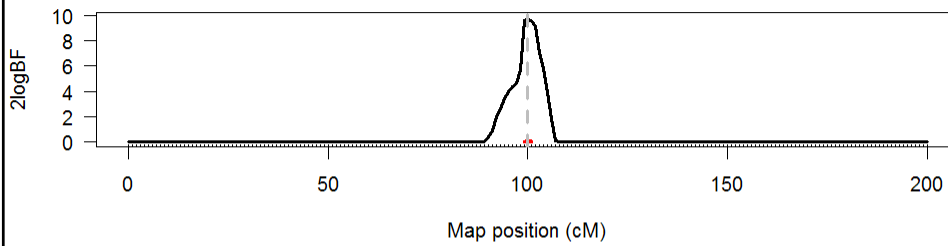
September 2006

Jax Workshop © Brian S. Yandell

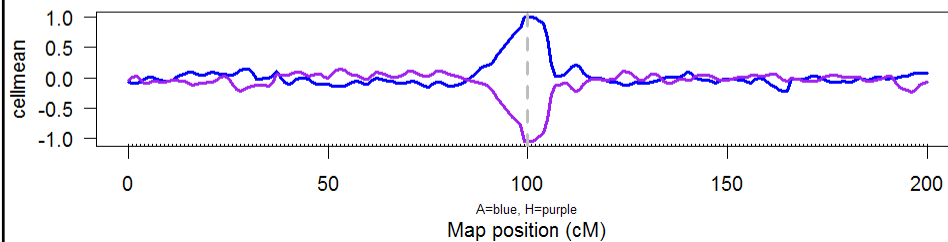
36

# scan of marginal Bayes factor

2logBF of phenotype for main



cellmean of phenotype for A+H



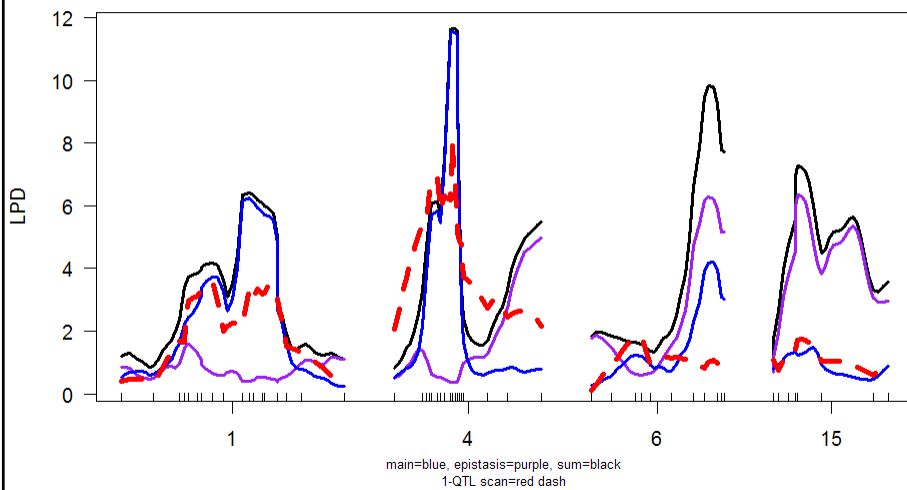
September 2006

Jax Workshop © Brian S. Yandell

37

# hyper data: scanone

LPD of bp for main+epistasis+sum



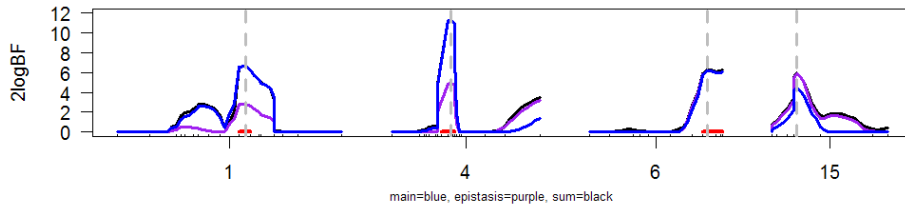
September 2006

Jax Workshop © Brian S. Yandell

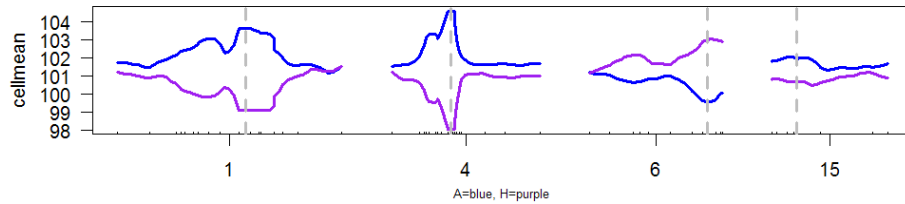
38

# 2log(BF) scan with 50% HPD region

2logBF of bp for main+epistasis+sum



cellmean of bp for A+H



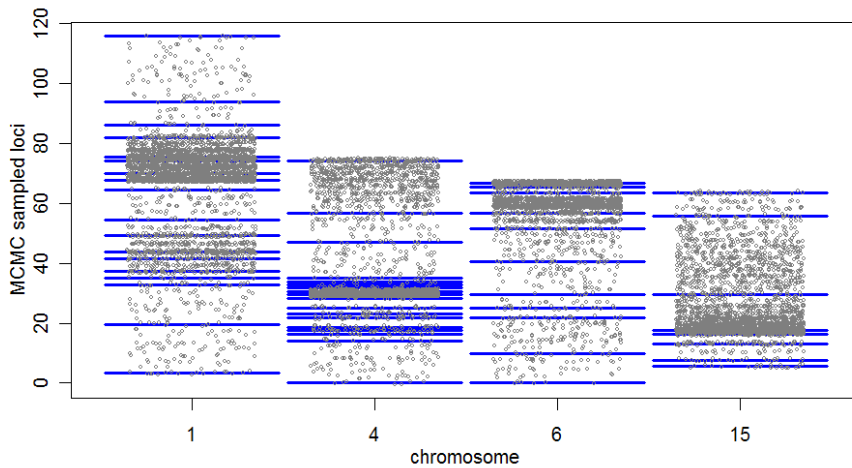
September 2006

Jax Workshop © Brian S. Yandell

39

# sampled QTL by chromosome

blue lines = markers

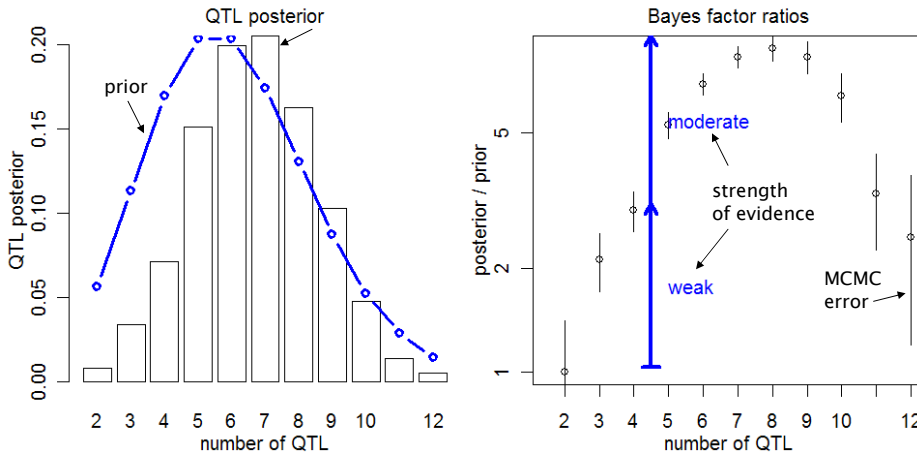


September 2006

Jax Workshop © Brian S. Yandell

40

# hyper: number of QTL posterior, prior, Bayes factors

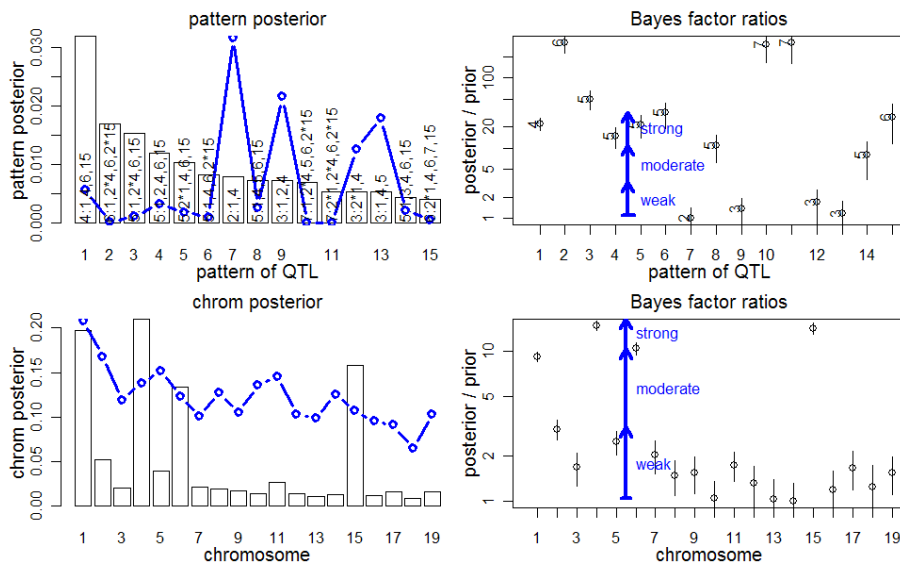


September 2006

Jax Workshop © Brian S. Yandell

41

# pattern of QTL on chromosomes

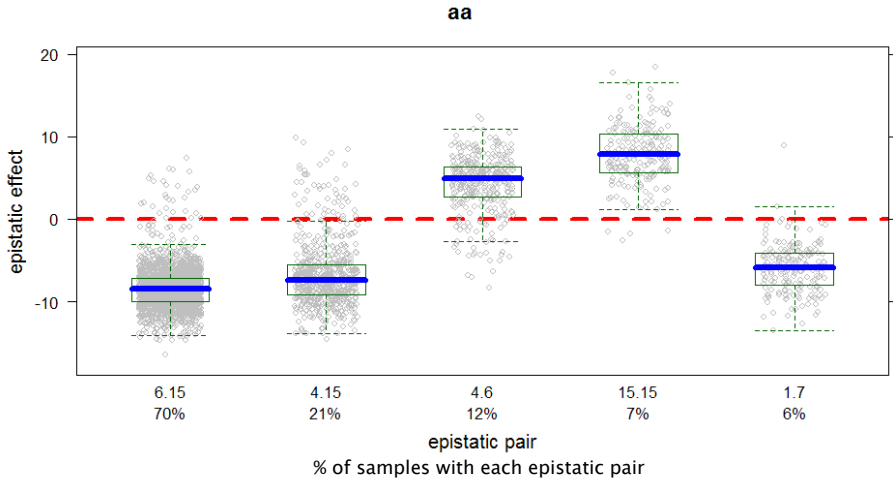


September 2006

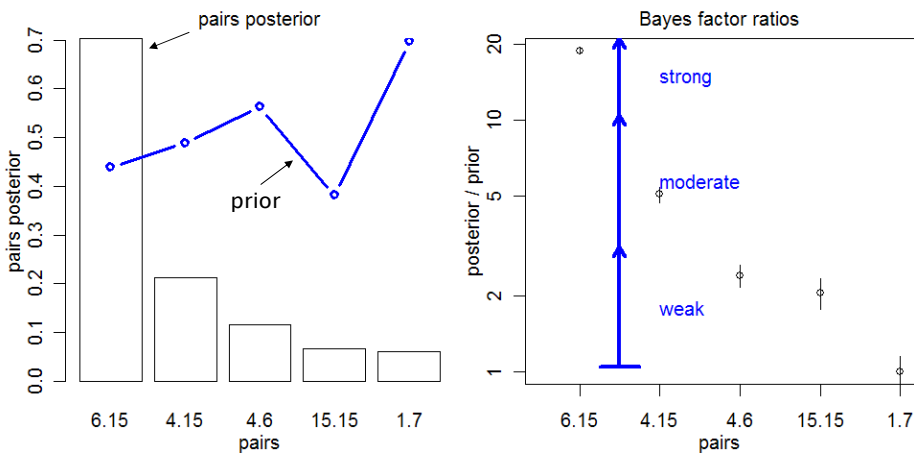
Jax Workshop © Brian S. Yandell

42

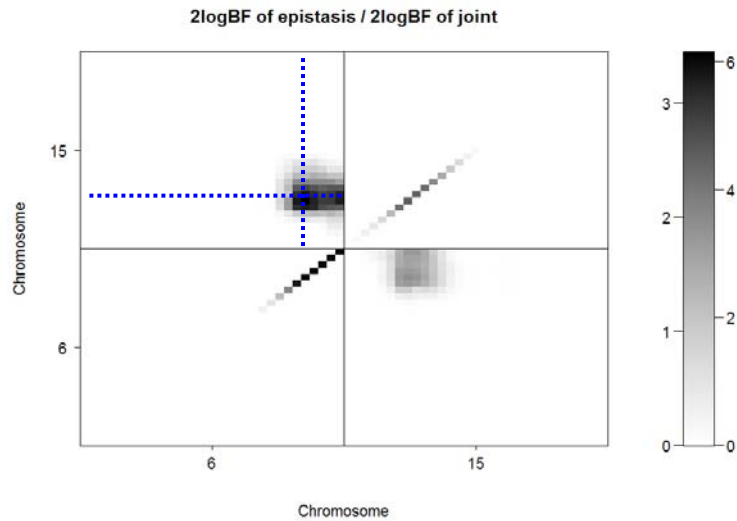
# Cockerham epistatic effects



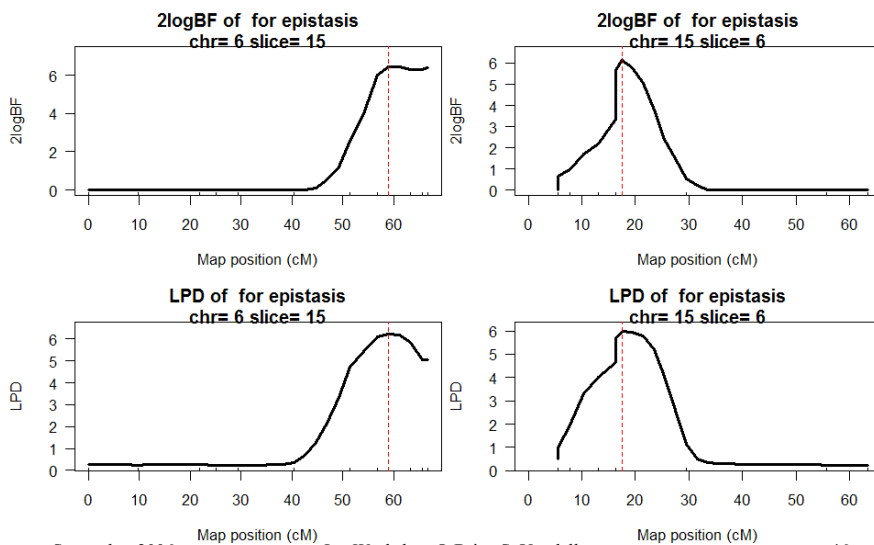
# relative importance of epistasis



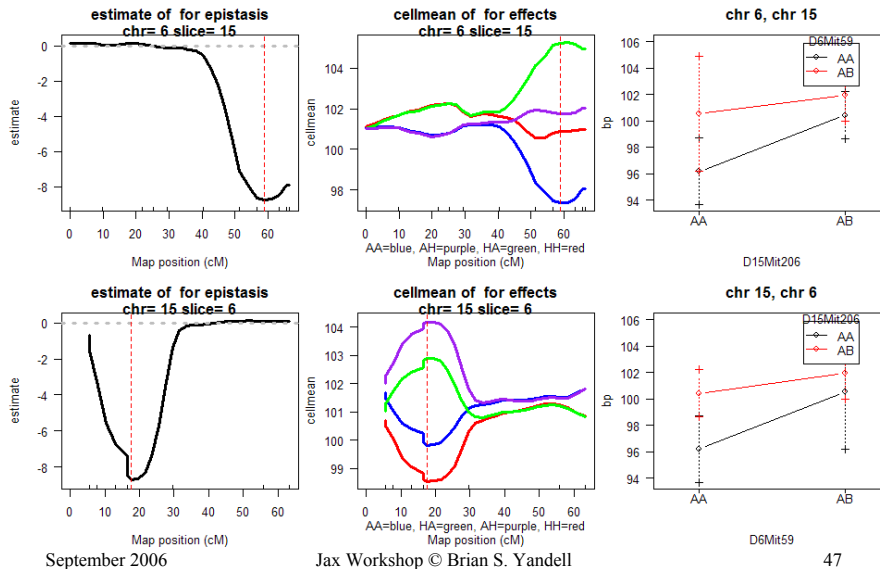
## 2-D plot of 2logBF: chr 6 & 15



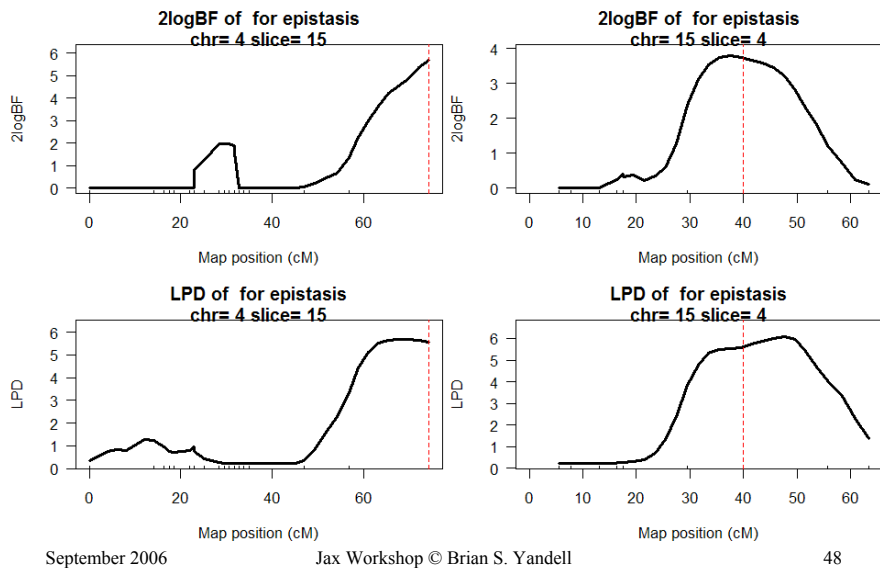
## 1-D Slices of 2-D scans: chr 6 & 15



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

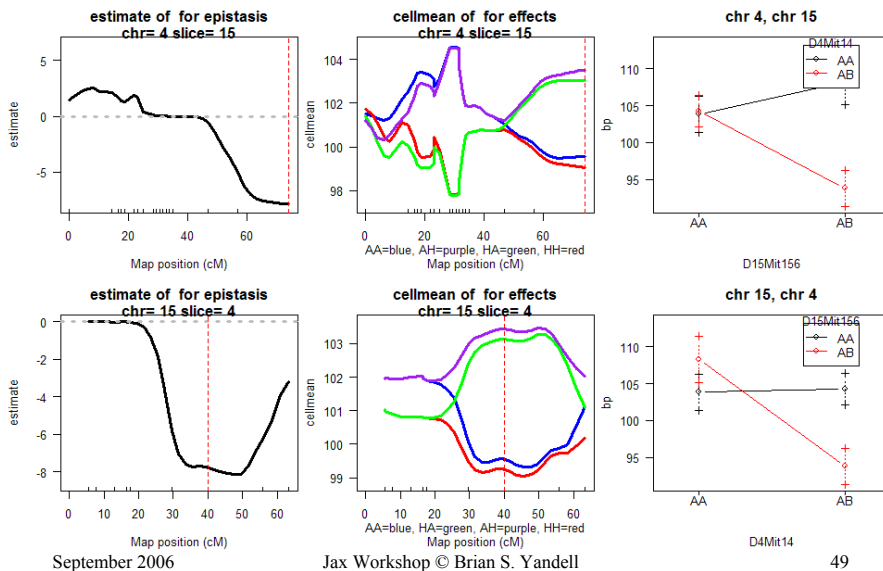


# 1-D Slices of 2-D scans: chr 4 & 15

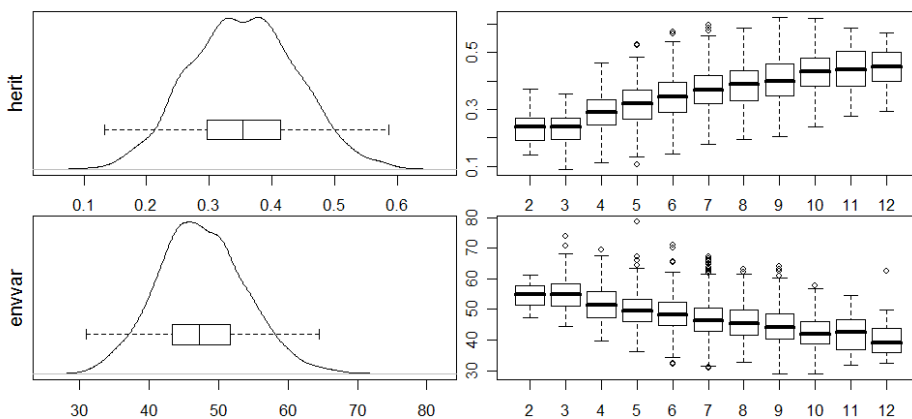




# 1-D Slices of 2-D scans: chr 4 & 15



# diagnostic summaries



# QTL for Bayesian Interval Mapping

## R/qtlbim: our software

- publication
  - Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005 *Genetics*)
  - Yi et al. Yandell (in review)
  - CRAN release Fall 2006
- properties
  - new MCMC algorithms
    - Gibbs with loci indicators; no reversible jump
  - epistasis, fixed & random covariates, GxE
  - extensive graphics

## R/qtlbim: our software

- R/qtlbim is cross-compatible with R/qtl
- 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 to CRAN (Yi, Yandell, Mehta, Banerjee, Shriner, Neely, von Smith 2006)

## 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 multi-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

Jackson Labs	Tom Osborn	Michael Newton
Gary Churchill	David Butruille	Hyuna Yang
Hao Wu	Marcio Ferrera	Daniel Sorensen
Randy von Smith	Josh Udahl	Daniel Gianola
U AL Birmingham	Pablo Quijada	Liang Li
David Allison	Alan Attie	my students
Nengjun Yi	Jonathan Stoehr	Jaya Satagopan
Tapan Mehta	Hong Lan	Fei Zou
Samprit Banerjee	Susie Clee	Patrick Gaffney
	Jessica Byers	Chunfang Jin
		Elias Chaibub
		W Whipple Neely

USDA Hatch, NIH/NIDDK (Attie), NIH/R01 (Yi)