Bayesian Model Selection for Quantitative Trait Loci with Markov chain Monte Carlo in Experimental Crosses

Brian S. Yandell

University of Wisconsin-Madison

www.stat.wisc.edu/~yandell/statgen↥

Jackson Laboratory, September 2004

outline

1. What is the goal of QTL study?
2. Bayesian priors & posteriors
3. Model search using MCMC
    - Gibbs sampler and Metropolis-Hastings
    - Reversible jump MCMC
    - Fully MCMC approach (loci indicators)
4. Model assessment
    - Bayes factors
    - model selection diagnostics
    - simulation and *Brassica napus* example

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

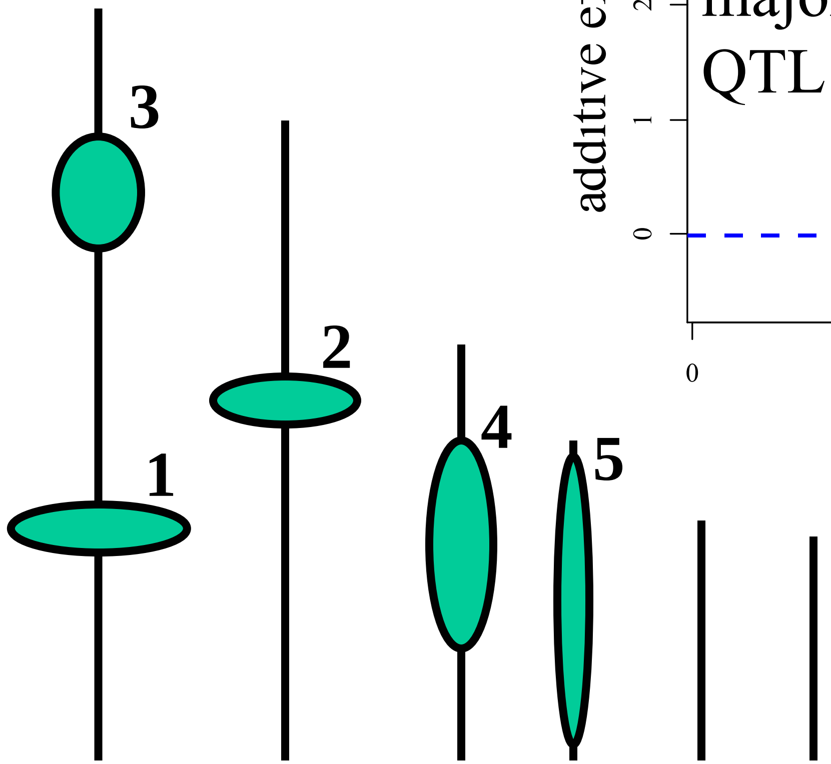
- improve statistical power, precision
  - increase number of QTL that can be detected
  - better estimates of loci and effects: less bias, smaller intervals
- improve inference of complex genetic architecture
  - infer number of QTL and their pattern across chromosomes
  - construct “good” estimates of effects
    - gene action (additive, dominance) and epistatic interactions
  - assess relative contributions of different QTL
- improve estimates of genotypic values
  - want less bias (more accurate) and smaller variance (more precise)
  - balance in mean squared error = MSE = $(\text{bias})^2$ + variance
    - always a compromise…

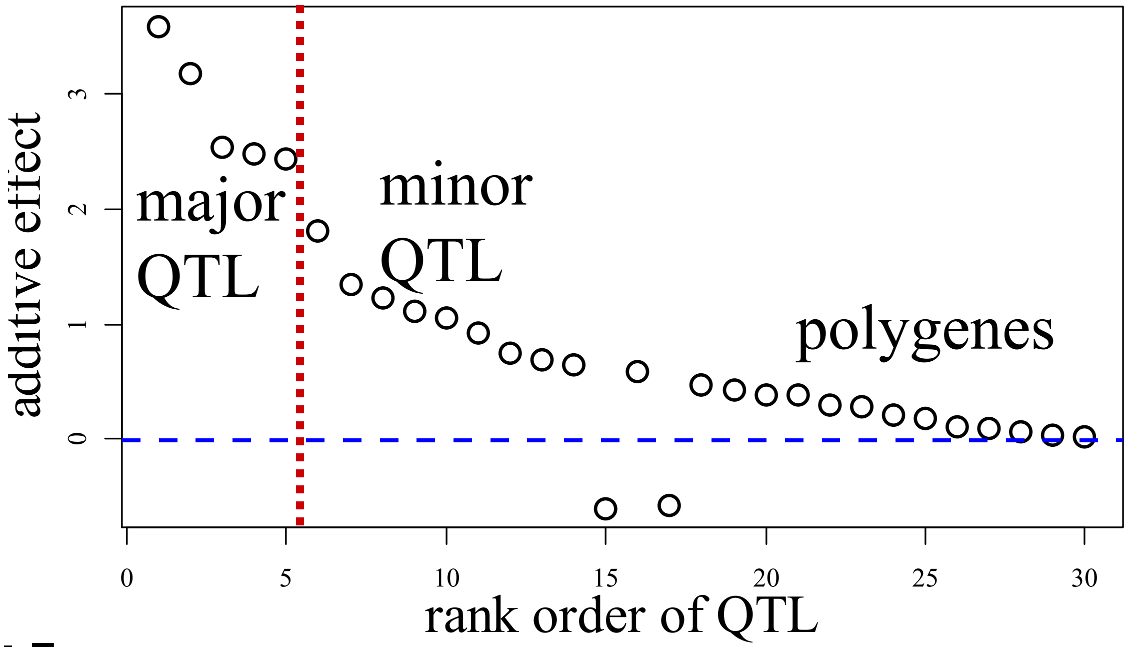
why worry about multiple QTL?

- many, many QTL may affect most any trait
  - how many QTL are detectable with these data?
    - limits to useful detection (Bernardo 2000)
    - depends on sample size, heritability, environmental variation
  - consider probability that a QTL is in the model
    - avoid sharp in/out dichotomy
    - major QTL usually selected, minor QTL sampled infrequently
- build M = model = genetic architecture into model
  - M = {loci 1,2,…,*m*, plus interactions 12,13,…}
  - directly allow uncertainty in genetic architecture
  - model selection over number of QTL, genetic architecture
  - use Bayes factors and model averaging
    - to identify “better” models

Pareto diagram of QTL effects

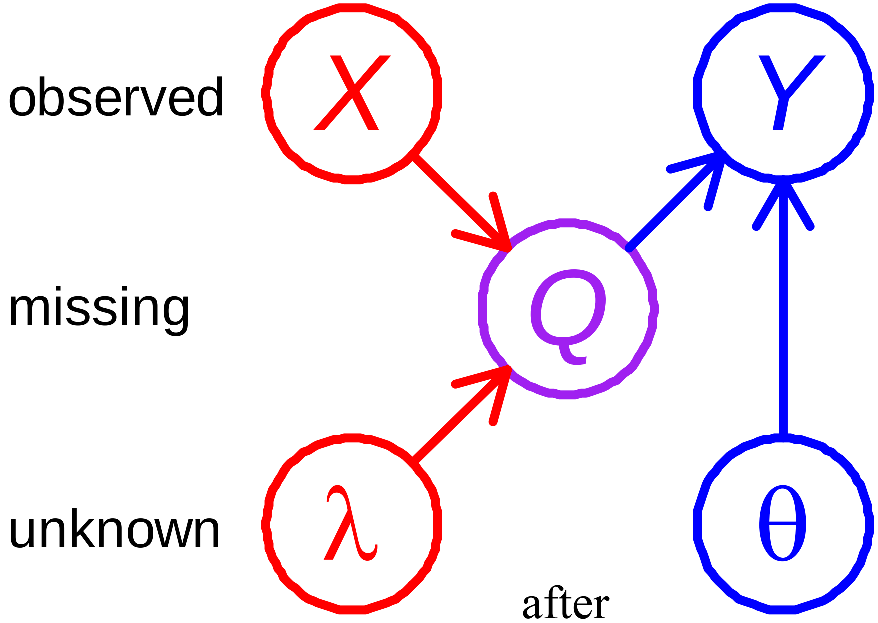
major QTL on linkage map





interval mapping basics

- observed measurements
  - Y = phenotypic trait
  - X = 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)
  - θ = phenotype model parameters
  - m = number of QTL
- pr($Q|X,\lambda,m$) genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for Q given X
- pr($Y|Q,\theta,m$) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters θ (could be non-parametric)



after
Sen Churchill (2001)

2. Bayesian priors for QTL

- genomic region = locus λ
  - may be uniform over genome
  - pr(λ | X) = 1 / length of genome
  - or may be restricted based on prior studies
- missing genotypes Q
  - depends on marker map and locus for QTL
  - pr(Q | X, λ)
  - genotype (recombination) model is formally a prior
- genotypic means and variance $\theta = (G_q, \sigma^2)$
  - pr(θ) = pr(G_q | σ^2) pr(σ^2)
  - use conjugate priors for normal phenotype
    - pr(G_q | σ^2) = normal
    - pr(σ^2) = inverse chi-square

Bayesian model posterior

- augment data (*Y*,*X*) with unknowns *Q*
- study unknowns (θ,λ,*Q*) given data (*Y*,*X*)
  - properties of posterior pr(θ,λ,*Q* | *Y*, *X*)
- sample from posterior in some clever way
  - multiple imputation or MCMC

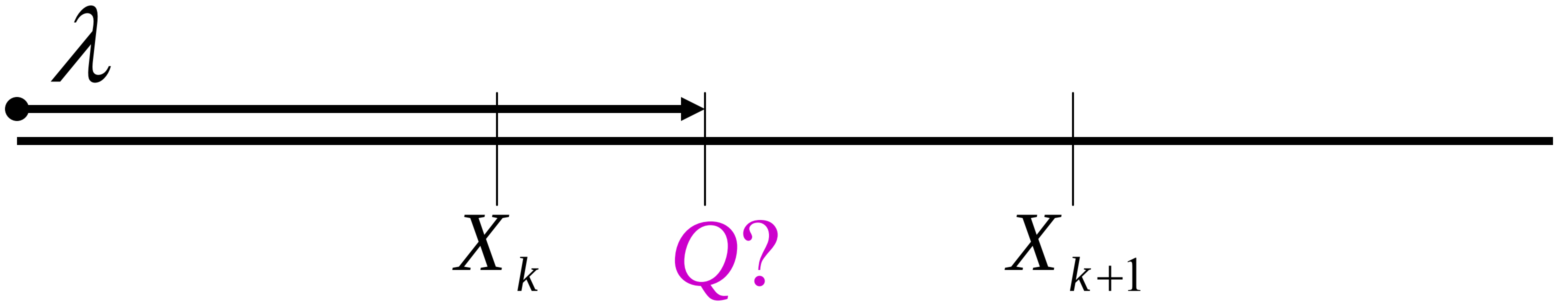
$$\text{pr}(\theta, \lambda, Q \,|\, Y, X) = \frac{\text{pr}(Y \,|\, Q, \theta)\text{pr}(Q \,|\, X, \lambda)\text{pr}(\theta)\text{pr}(\lambda \,|\, X)}{\text{pr}(Y \,|\, X)}$$

$$\text{pr}(\theta, \lambda \,|\, Y, X) = \text{sum}_Q \; \text{pr}(\theta, \lambda, Q \,|\, Y, X)$$

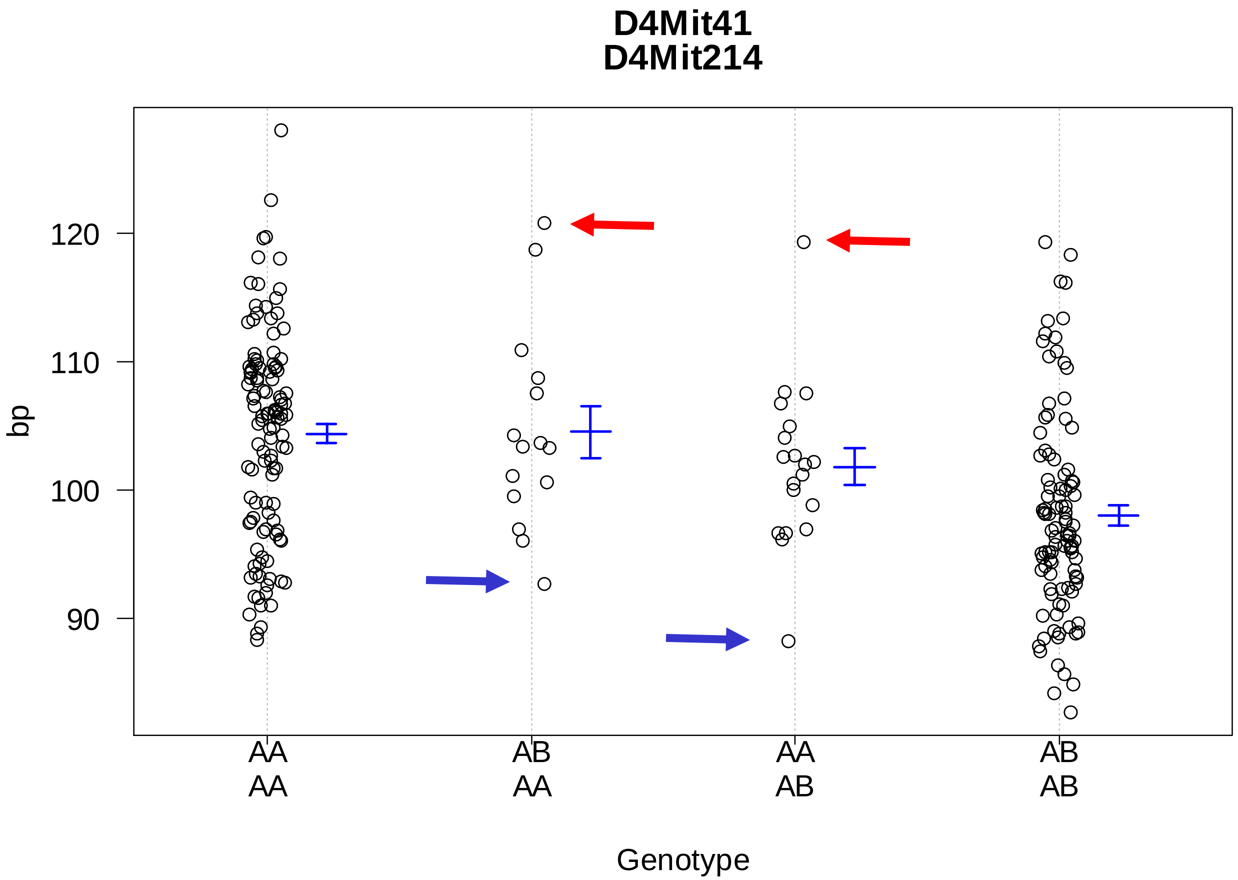
genotype prior model: $\mathrm{pr}(Q|X,\lambda)$

- locus λ is distance along linkage map
  - map assumed known from earlier study
  - λ identifies flanking marker interval
- use flanking markers to approximate prior on Q
  - slight inaccuracy by ignoring multipoint map function
  - use next flanking markers if missing data

$$\mathrm{pr}(Q|X,\lambda) = \mathrm{pr}(\text{geno} \mid \text{map, locus}) \approx$$
$$\mathrm{pr}(\text{geno} \mid \text{flanking markers, locus})$$



how does phenotype *Y* improve posterior for genotype *Q*?



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

- full conditional of Q given data, parameters
  - proportional to prior pr(Q | X_i, λ)
    - weight toward Q that agrees with flanking markers
  - proportional to likelihood $\text{pr}(Y_i|Q,\theta)$
    - weight toward Q so that group mean $G_Q \approx Y_i$
- phenotype and flanking markers may conflict
  - posterior recombination balances these two weights

$$\text{pr}(Q \mid Y_i, X_i, \theta, \lambda) = \frac{\text{pr}(Q \mid X_i, \lambda)\text{pr}(Y_i \mid Q, \theta)}{\text{pr}(Y_i \mid X_i, \theta, \lambda)}$$

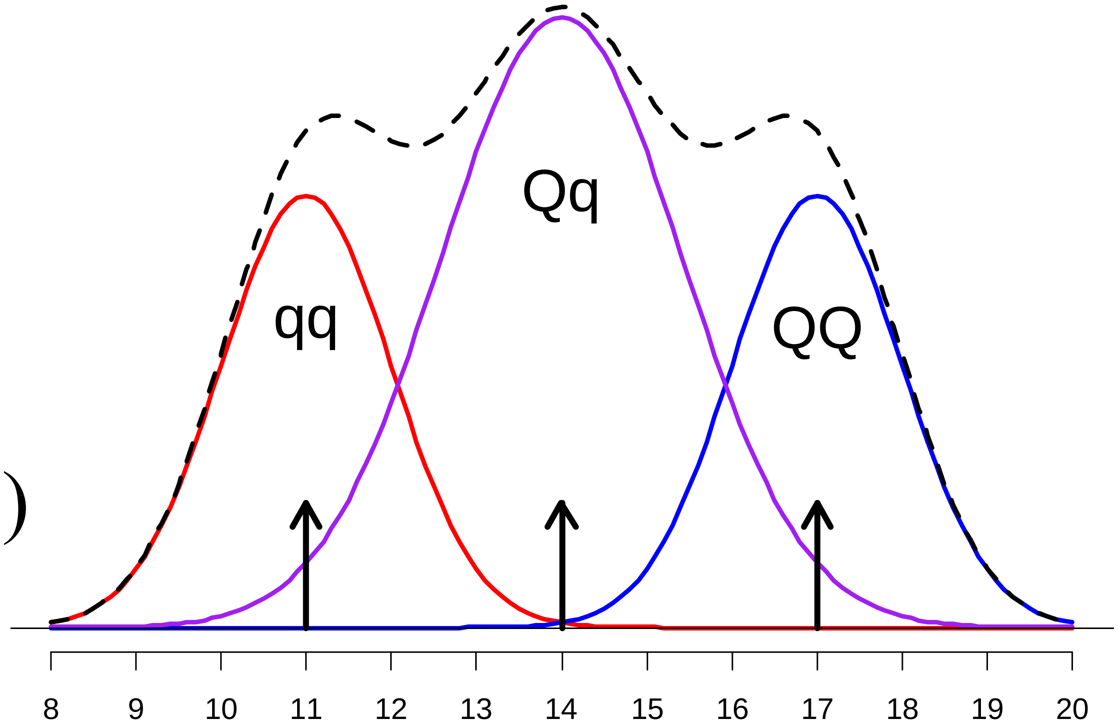
idealized phenotype model

- trait = mean + additive + error
- trait = effect_of_geno + error
- pr(trait | geno, effects)

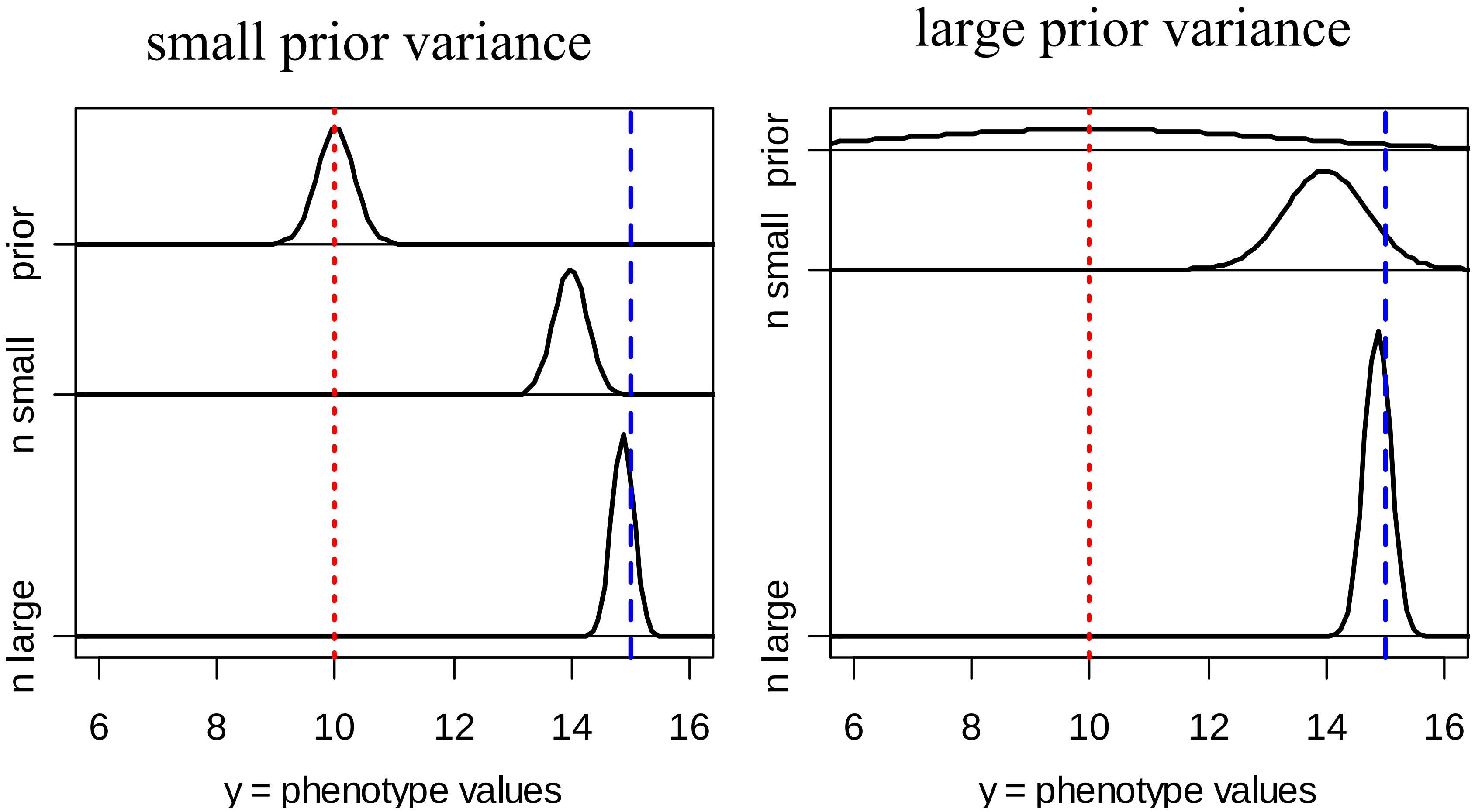
$$Y = G_Q + E$$

$$\mathrm{pr}(Y \mid Q, \theta) =$$

$$\mathrm{normal}(G_Q, \sigma^2)$$



priors & posteriors: normal data



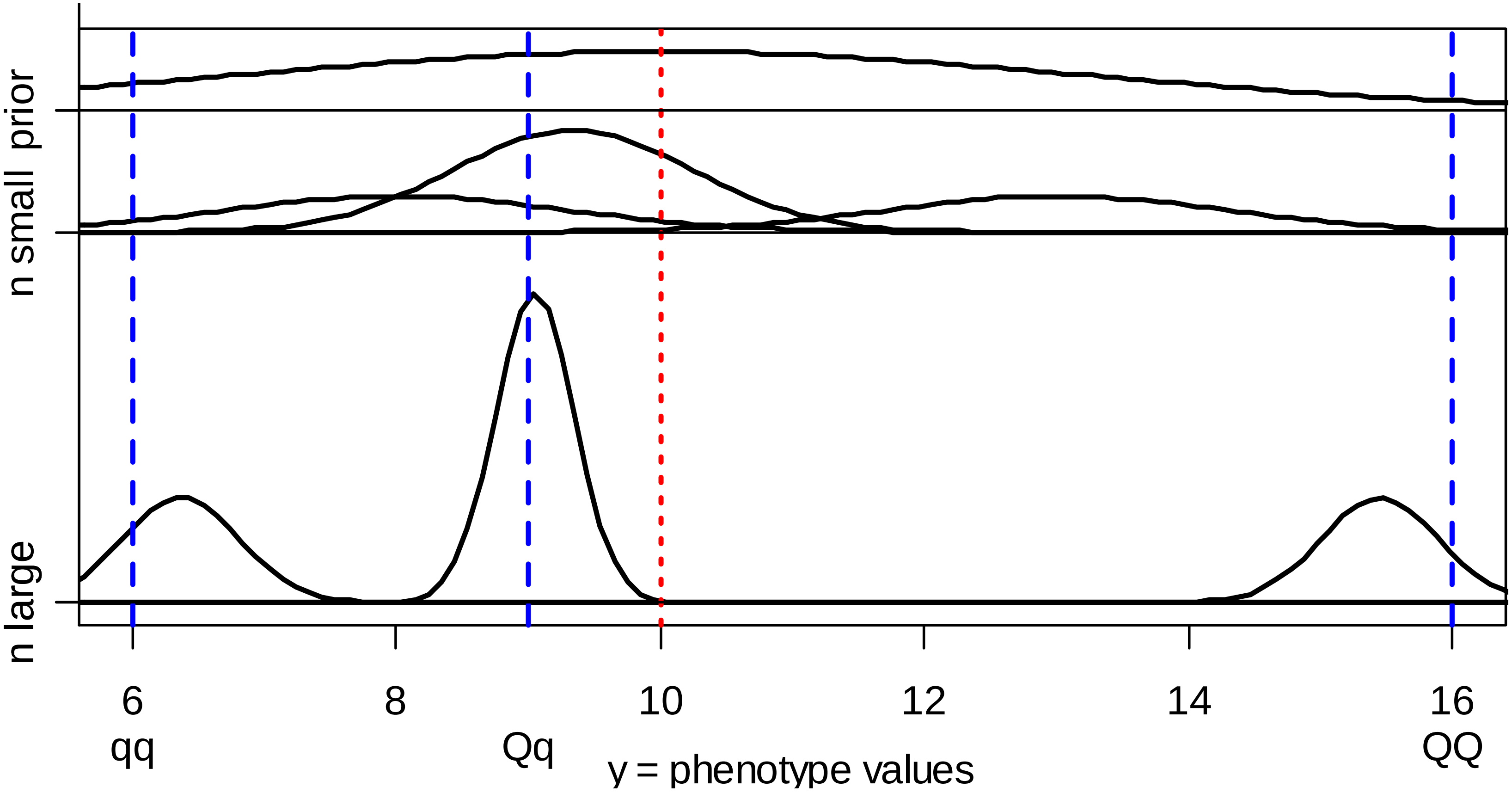
small prior variance

large prior variance

priors & posteriors: normal data

model	$Y_i = \mu + E_i$
environment	$E \sim N(0, \sigma^2)$, σ^2 known
likelihood	$Y \sim N(\mu, \sigma^2)$
prior	$\mu \sim N(\mu_0, \kappa\sigma^2)$, κ 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	$\mu \sim N\left(B_n\bar{Y}_\bullet + (1 - B_n)\mu_0, B_n \frac{\sigma^2}{n} \right)$
	with $\bar{Y}_\bullet = \text{sum}\frac{Y_i}{n}$
fudge factor (shrinks to 1)	$B_n = \frac{\kappa n}{\kappa n + 1} \to 1$

prior & posteriors: genotypic means G_Q



prior & posteriors: genotypic means G_Q

posterior centered on sample genotypic mean
but shrunken slightly toward overall mean
κ is related to heritability

prior:

$$G_Q \sim N\left(\overline{Y}_{\bullet}, \kappa\sigma^2\right)$$

posterior:

$$G_Q \sim N\left(B_Q \overline{Y}_Q + (1 - B_Q)\overline{Y}_{\bullet}, B_Q \frac{\sigma^2}{n_Q} \right)$$

$$n_Q = \text{count}\{Q_i = Q\}, \overline{Y}_Q = \underset{\{i:Q_i = Q\}}{\text{sum}} \frac{Y_i}{n_Q}$$

fudge factor:

$$B_Q = \frac{\kappa n_Q}{\kappa n_Q + 1} \to 1$$

What if variance σ^2 is unknown?

- sample variance is proportional to chi-square
  - $ns^2 / \sigma^2 \sim \chi^2 (n)$
  - likelihood of sample variance s^2 given n, σ^2
- conjugate prior is inverse chi-square
  - $\nu\tau^2 / \sigma^2 \sim \chi^2 (\nu)$
  - prior of population variance σ^2 given ν, τ^2
- posterior is weighted average of likelihood and prior
  - $(\nu\tau^2+ns^2) / \sigma^2 \sim \chi^2 (\nu+n)$
  - posterior of population variance σ^2 given n, s^2, ν, τ^2
- empirical choice of hyper-parameters
  - $\tau^2= s^2/3$, $\nu=6$
  - $E(\sigma^2 \mid \nu,\tau^2) = s^2/2$, $\mathrm{Var}(\sigma^2 \mid \nu,\tau^2) = s^4/4$

multiple QTL phenotype model

- phenotype affected by genotype & environment

  $\mathrm{pr}(Y|Q=q,\theta) \sim N(G_q, \sigma^2)$

  $Y = G_Q + \text{environment}$

- partition genotypic mean into QTL effects

  $G_q = \mu \quad + \beta_{1q} + \ldots + \beta_{mq} \quad + \beta_{12q} + \ldots$

  $G_q = \text{mean} \quad + \text{main effects} \quad + \text{epistatic interactions}$

- general form of QTL effects for model M

  $G_q = \mu \quad + \mathrm{sum}_{j \text{ in } M}\, \beta_{jq}$

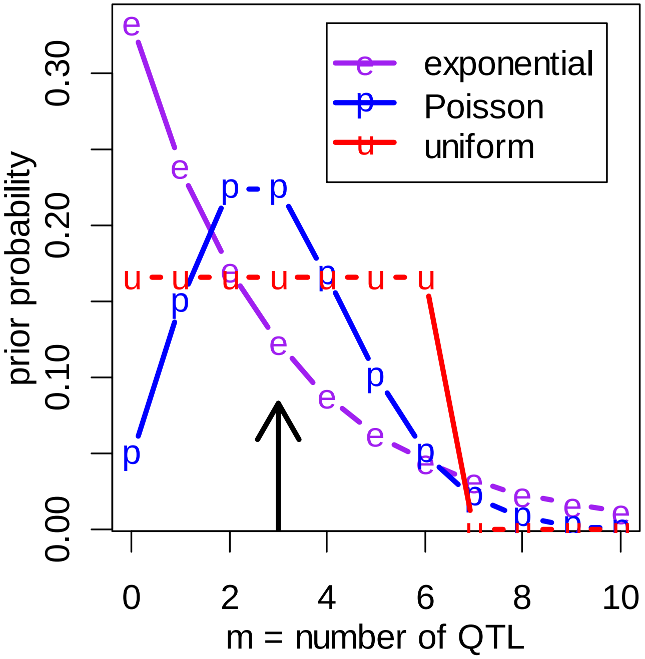
  $|M|$ = number of terms in model $M < 2^m$

- can partition prior and posterior into effects β_{jq}

  (details omitted)

prior & posterior on number of QTL

- what prior on number of QTL?
  - uniform over some range
  - Poisson with prior mean
  - geometric with prior mean
- prior influences posterior
  - good: reflects prior belief
    - push data in discovery process
  - bad: skeptic revolts!
    - “answer” depends on “guess”



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
- update *m*-QTL model components from full conditionals
  - update locus λ given Q,X (using Metropolis-Hastings step)
  - update genotypes Q given λ,θ,Y,X (using Gibbs sampler)
  - update effects θ given Q,Y (using Gibbs sampler)

$$(\lambda,Q,\theta,m) \sim \mathrm{pr}(\lambda,Q,\theta,m \mid Y,X)$$

$$(\lambda,Q,\theta,m)_1 \rightarrow (\lambda,Q,\theta,m)_2 \rightarrow \cdots \rightarrow (\lambda,Q,\theta,m)_N$$

Gibbs sampler idea

- two correlated normals (genotypic means in BC)
  - could draw samples from both together
  - but easier to sample one at a time
- Gibbs sampler:
  - sample each from its full conditional
  - pick order of sampling at random
  - repeat *N* times

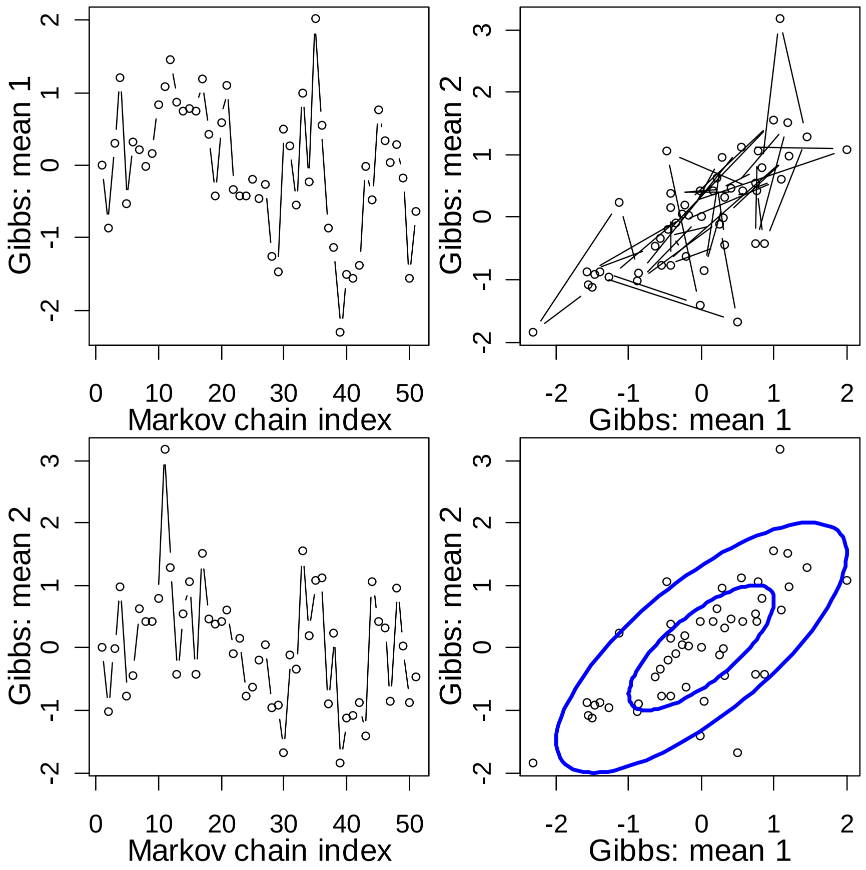
$$G_{\mathrm{QQ}} \sim N(0,1); G_{\mathrm{Qq}} \sim N(0,1) \text{ but } cor(G_{\mathrm{QQ}}, G_{\mathrm{Qq}}) = \rho$$

$$G_{\mathrm{QQ}} \text{ given } G_{\mathrm{Qq}} \sim N\left(\rho G_{\mathrm{Qq}}, 1-\rho^2\right)$$

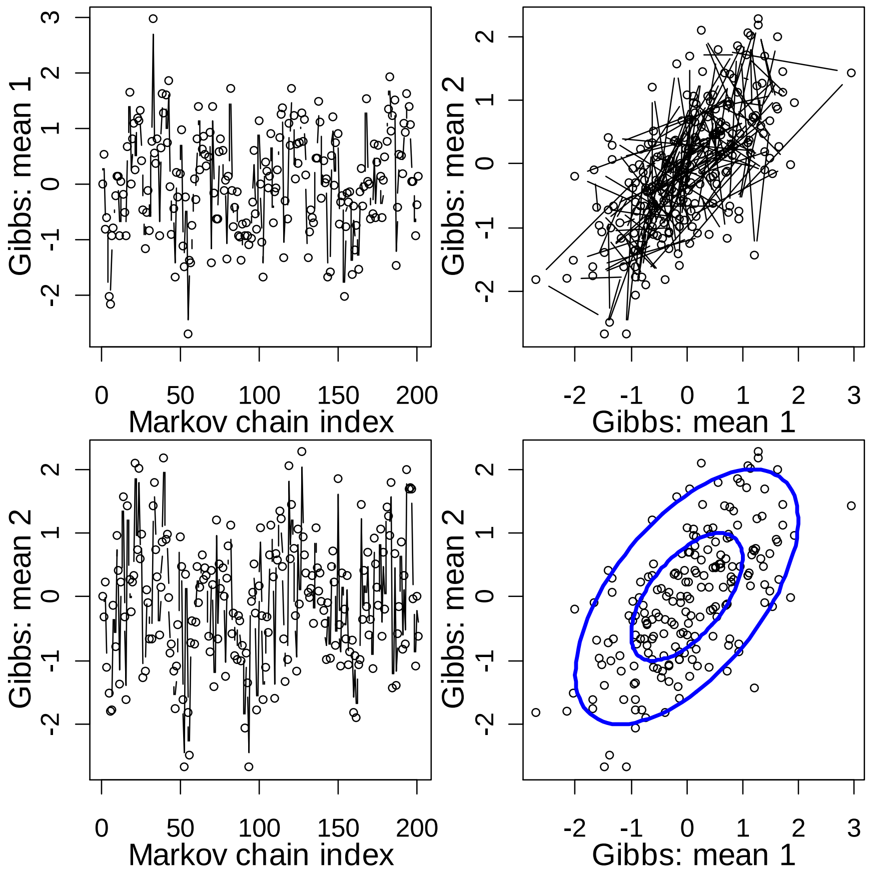
$$G_{\mathrm{Qq}} \text{ given } G_{\mathrm{QQ}} \sim N\left(\rho G_{\mathrm{QQ}}, 1-\rho^2\right)$$

Gibbs sampler samples: $\rho = 0.6$

$N = 50$ samples



$N = 200$ samples



How to sample a locus λ?

- cannot easily sample from locus full conditional

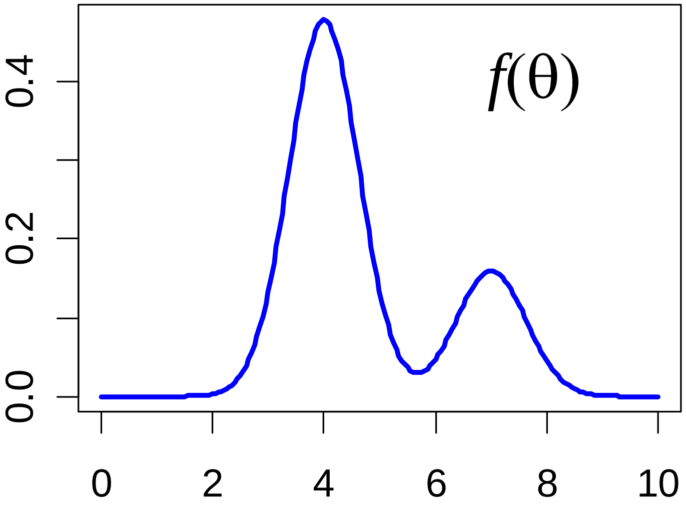
$$\begin{aligned}\text{pr}(\lambda\,|Y,X,\theta,Q) &= \text{pr}(\ \lambda \mid X,Q)\\ &= \text{pr}(\lambda\ )\ \text{pr}(\ Q \mid X,\ \lambda\ )\ /\ \text{constant}\end{aligned}$$

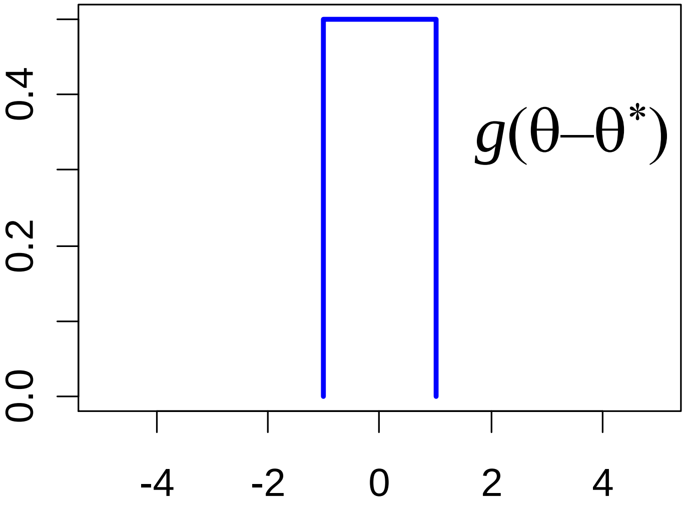
- to explicitly determine constant, must average
  - over all possible genotypes
  - over entire map
- Gibbs sampler will not work in general
  - but can use method based on ratios of probabilities
  - Metropolis-Hastings is extension of Gibbs sampler

Metropolis-Hastings idea

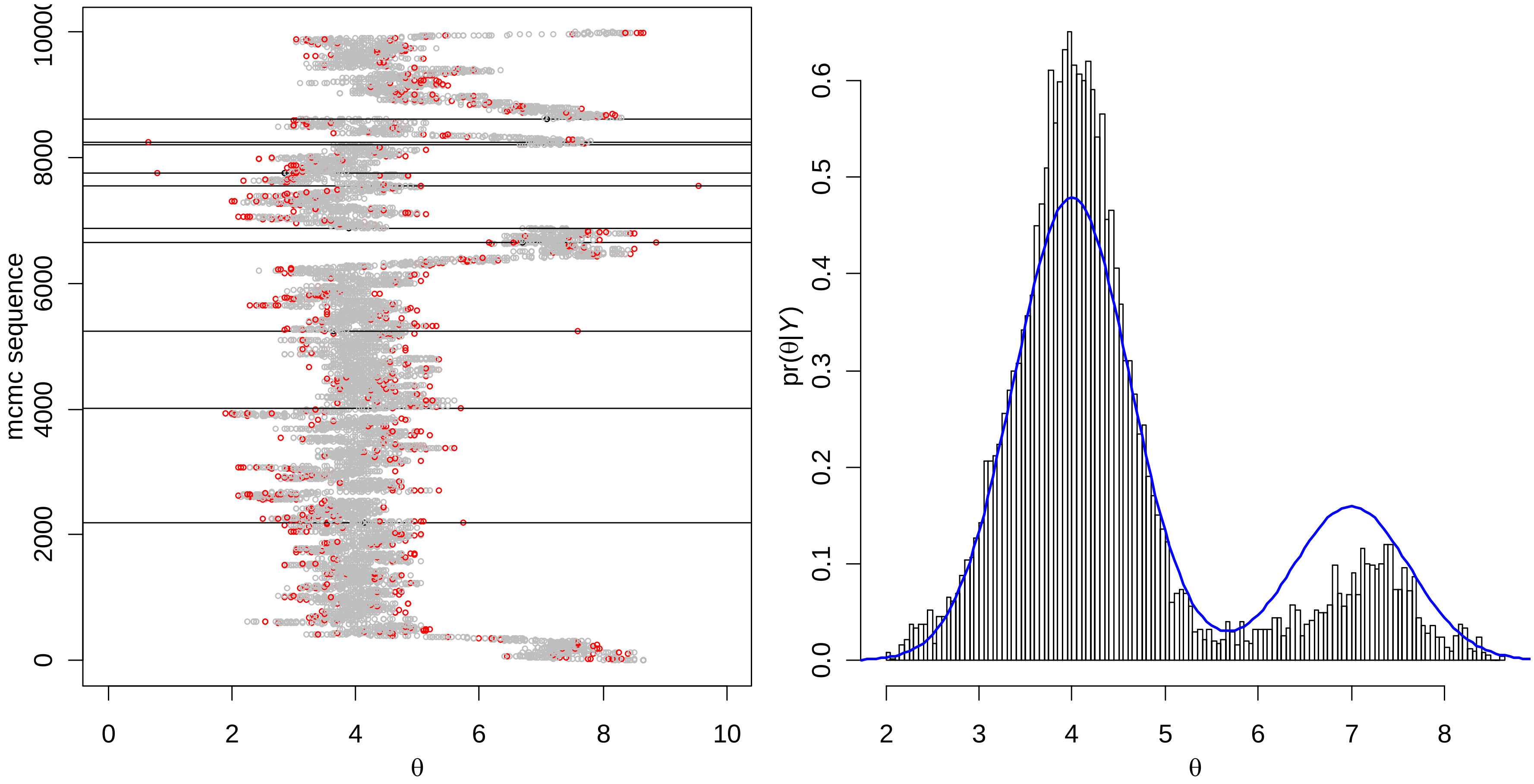
- want to study distribution $f(\theta)$
- take Monte Carlo samples
  - unless too complicated
- Metropolis-Hastings samples:
  - current sample value θ
  - propose new value θ^*
    - from some distribution $g(\theta,\theta^*)$
    - Gibbs sampler: $g(\theta,\theta^*) = f(\theta^*)$
  - accept new value with prob A
    - Gibbs sampler: $A = 1$

$$A = \min\left(1, \frac{f(\theta^*)g(\theta^*,\theta)}{f(\theta)g(\theta,\theta^*)}\right)$$



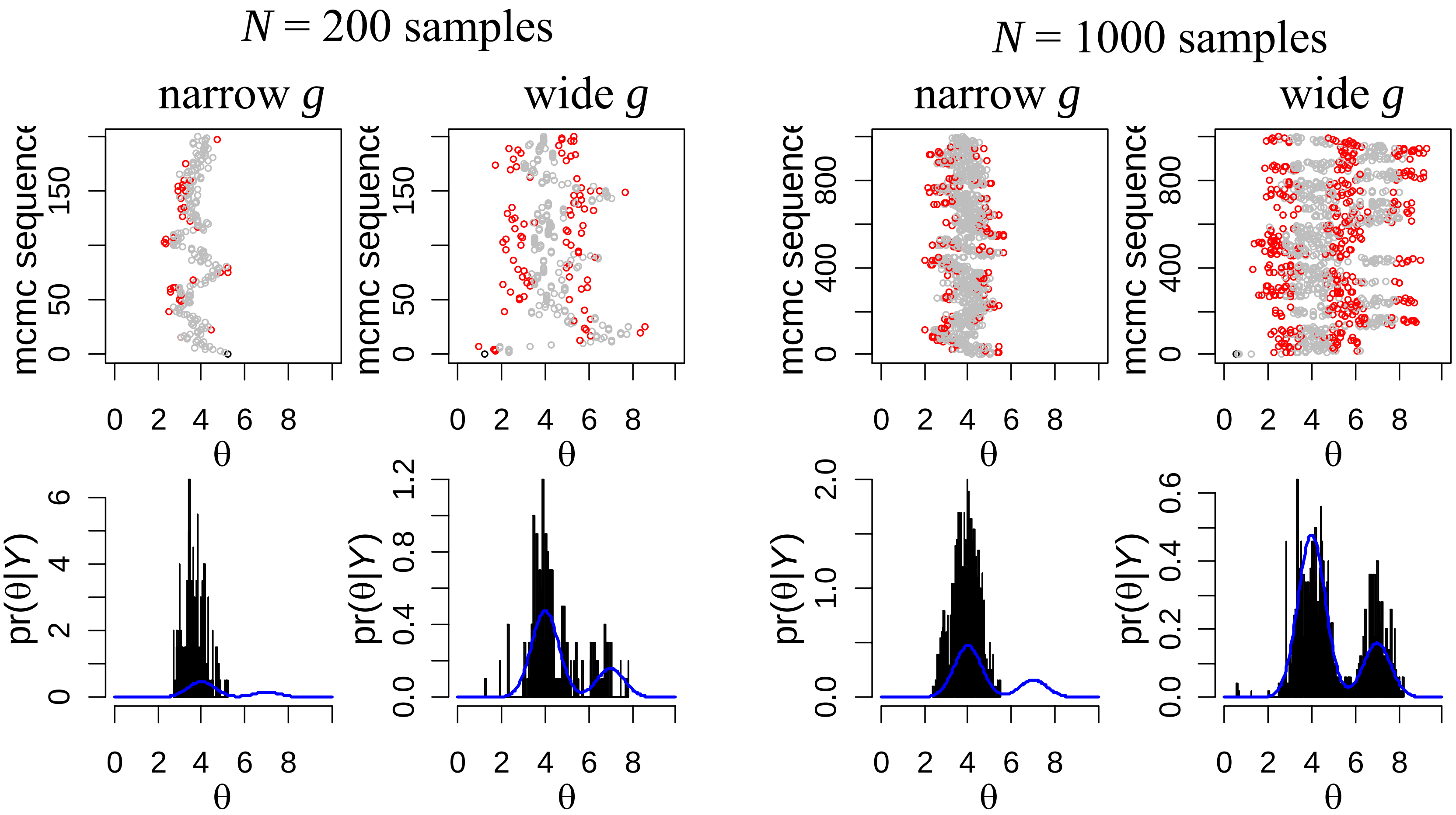


MCMC realization



added twist: occasionally propose from whole domain

Metropolis-Hastings samples



N = 200 samples

N = 1000 samples

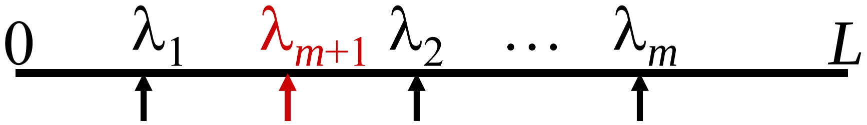
narrow *g*

wide *g*

narrow *g*

wide *g*

sampling multiple loci



action steps: draw one of three choices

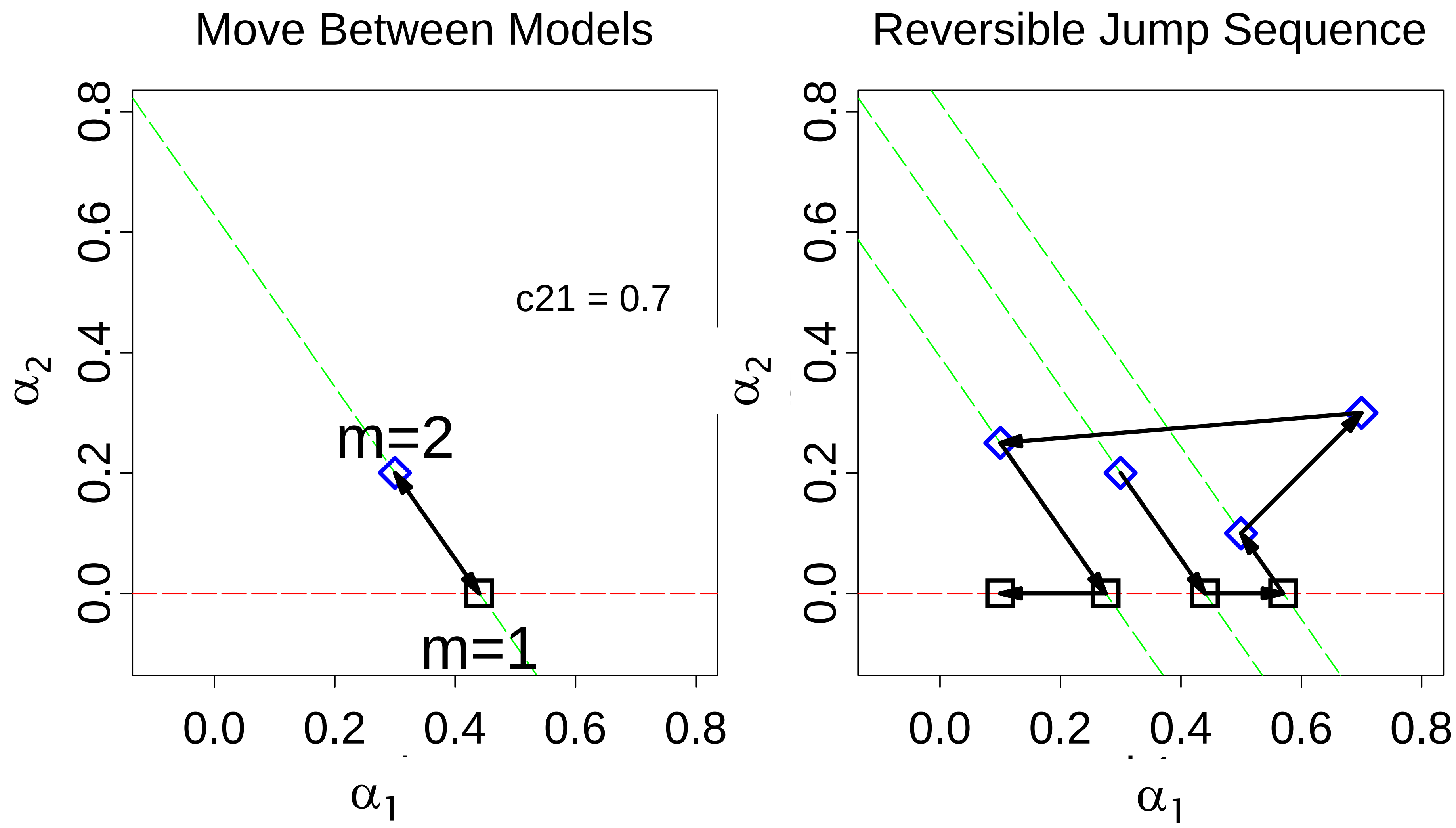
- update *m*-QTL model with probability 1-*b*(*m*+1)-*d*(*m*)
  - update current model using full conditionals
  - sample *m* QTL loci, effects, and genotypes
- add a locus with probability *b*(*m*+1)
  - propose a new locus along genome
  - innovate new genotypes at locus and phenotype effect
  - decide whether to accept the “birth” of new locus
- drop a locus with probability *d*(*m*)
  - propose dropping one of existing loci
  - decide whether to accept the “death” of locus

reversible jump MCMC

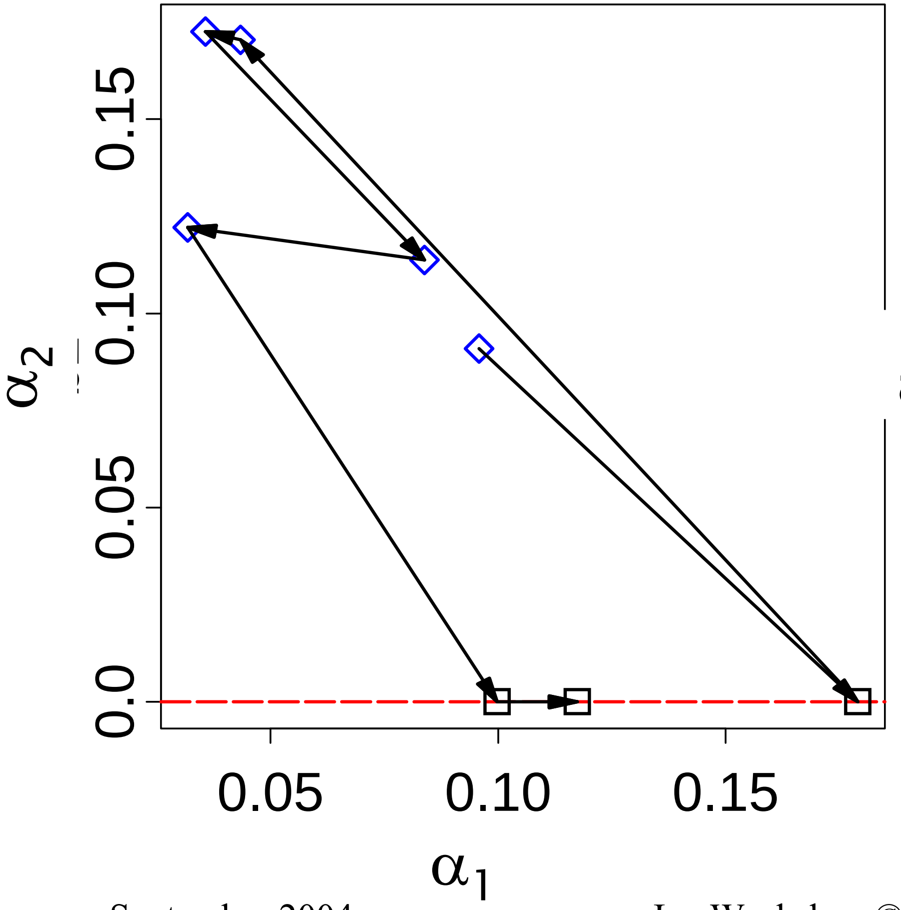
- consider known genotypes Q at 2 known loci λ
  - models with 1 or 2 QTL
- jump between 1-QTL and 2-QTL models
  - adjust parameters when model changes
  - α and α_1 differ due to collinearity of QTL genotypes

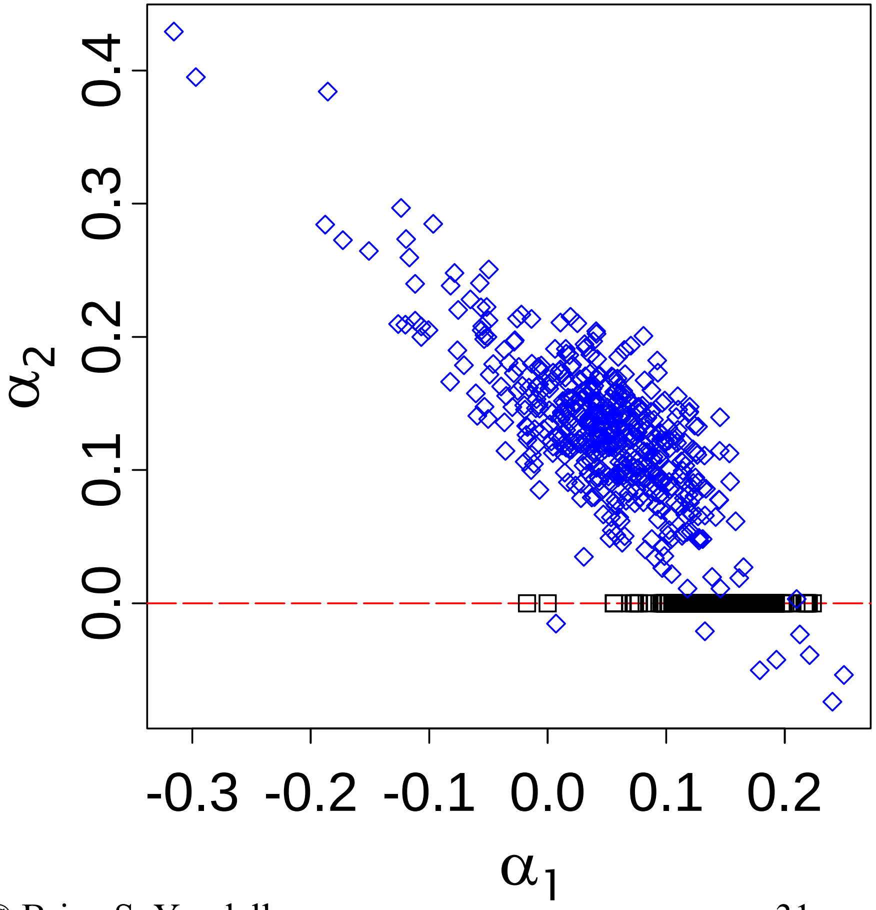
$$\begin{tikzcd}
m = 1 : Y = \mu + \beta_{1Q} + e \arrow[d, leftrightarrow, bend right] \\
m = 2 : Y = \mu + \beta_{1Q} + \beta_{2Q} + e
\end{tikzcd}$$

geometry of reversible jump



geometry allowing *Q* and *λ* to change


a short sequence


first 1000 with m<3

Gibbs sampler with loci indicators

- partition genome into intervals
  - at most one QTL per interval
  - interval = marker interval or large chromosome region
- use loci indicators in each interval
  - $\delta = 1$ if QTL in interval
  - $\delta = 0$ if no QTL
- Gibbs sampler on loci indicators
  - still need to adjust genetic effects for collinearity of *Q*
  - see work of Nengjun Yi (and earlier work of Ina Hoeschele)

$$Y = \mu + \delta_1 \alpha_1 (Q_1 - \overline{Q_1}) + \delta_2 \alpha_2 (Q_1 - \overline{Q_1}) + e$$

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)

limits of epistatic inference

- power to detect effects
  - epistatic model size grows exponentially
    - $|M| = 3^m$ for general interactions
  - power depends on ratio of n to model size
    - want n / $|M|$ to be fairly large (say > 5)
    - $n = 100$, $m = 3$, $n / |M| \approx 4$
- empty cells mess up adjusted (Type 3) tests
  - missing q_1Q_2 / q_1Q_2 or $q_1Q_2q_3$ / $q_1Q_2q_3$ genotype
  - null hypotheses not what you would expect
  - can confound main effects and interactions
  - can bias AA, AD, DA, DD partition

4. Model Assessment

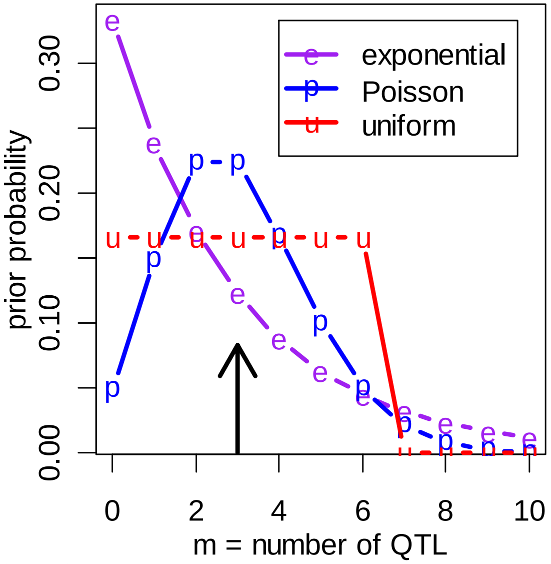
- balance model fit against model complexity

	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 $|M|$
  - compare IC = $-2 \log L(M \mid Y) + \text{penalty}(M)$
- Bayes factors: balance posterior by prior choice
  - compare pr(data $Y \mid$ model M)

QTL Bayes factors

- BF = posterior odds / prior odds
- BF equivalent to BIC
  - simple comparison: 1 vs 2 QTL
    - same as LOD test
  - general comparison of models
  - want Bayes factor >> 1
- m = number of QTL
  - indexes model complexity
  - genetic architecture also important



$$BF_{m,m+1} = \frac{\text{pr}(m|\text{data})/\text{pr}(m)}{\text{pr}(m+1|\text{data})/\text{pr}(m+1)}$$

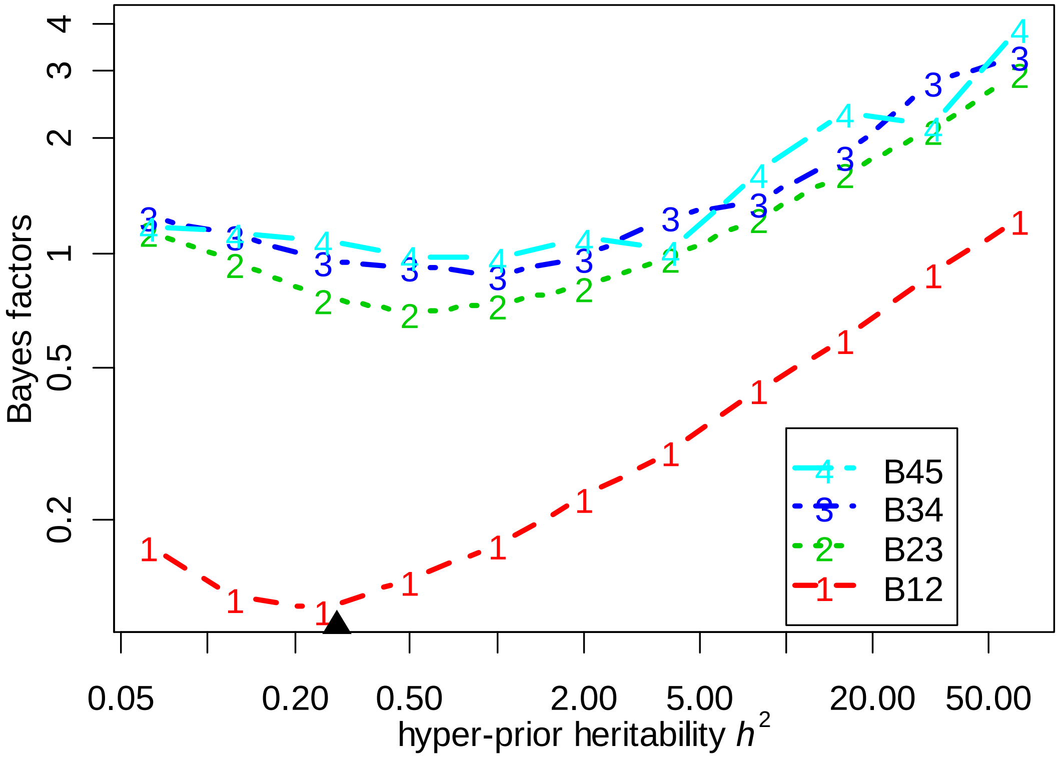
Bayes factors to assess models

- Bayes factor: which model best supports the data?
  - ratio of posterior odds to prior odds
  - ratio of model likelihoods
- equivalent to *LR* statistic when
  - comparing two nested models
  - simple hypotheses (e.g. 1 vs 2 QTL)
- Bayes Information Criteria (BIC)
  - Schwartz introduced for model selection in general settings
  - penalty to balance model size (*p* = number of parameters)

$$B_{12} = \frac{\mathrm{pr}(\text{ model}_1 \mid Y) / \mathrm{pr}(\text{ model}_2 \mid Y)}{\mathrm{pr}(\text{ model}_1) / \mathrm{pr}(\text{ model}_2)} = \frac{\mathrm{pr}(Y \mid \text{model}_1)}{\mathrm{pr}(Y \mid \text{model}_2)}$$

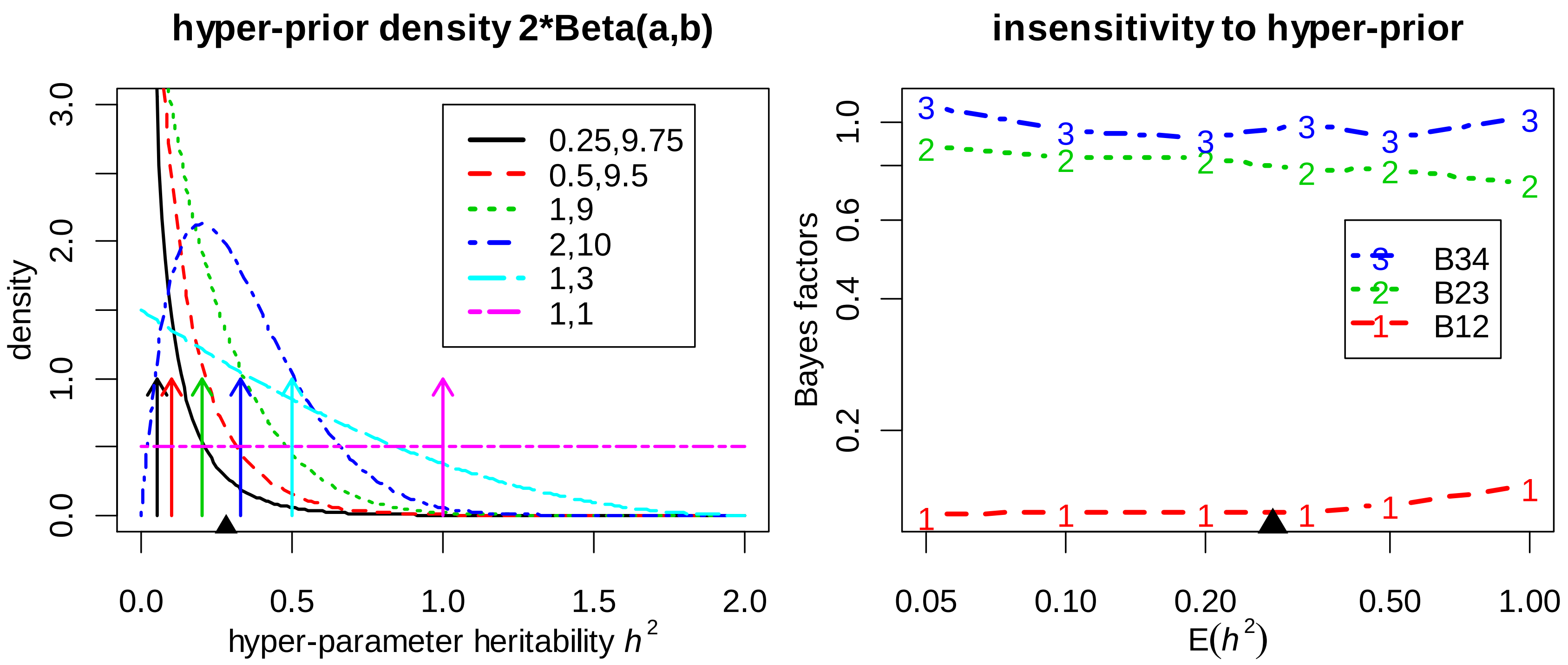
$$-2\log(B_{12}) = -2\log(LR) - (p_2 - p_1)\log(n)$$

BF sensitivity to fixed prior for effects



$$\beta_{jq} \sim \mathrm{N}\left(0, \frac{h^2 s^2}{|M|}\right), h^2 \text{ fixed}$$

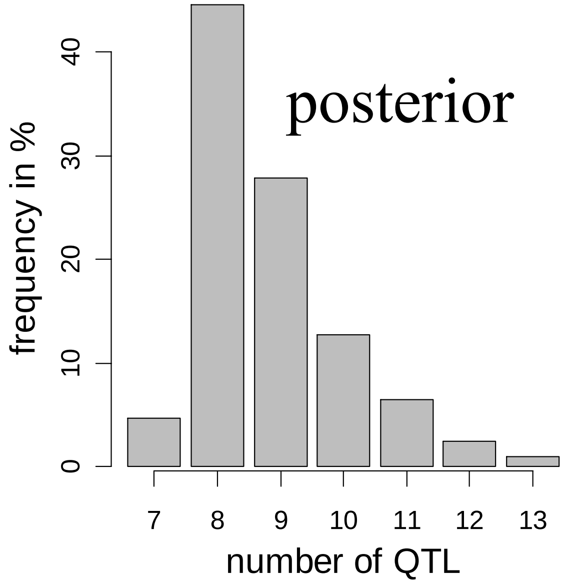
BF insensitivity to random effects prior



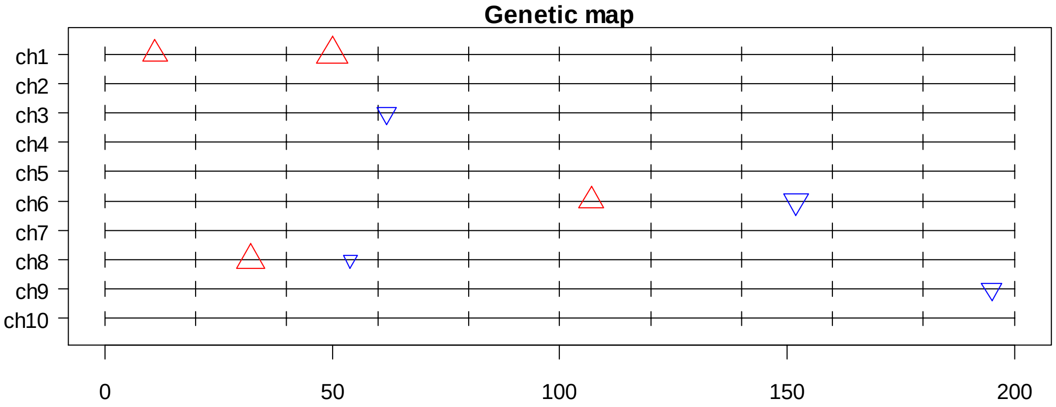
$$\beta_{jq} \sim \mathrm{N}\left(0, \frac{h^2 s^2}{|M|}\right), \; \frac{h^2}{2} \sim \mathrm{Beta}(a,b)$$

simulations and data studies

- 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%



QTL	chr	loci	effect
1	1	11	–3
2	1	50	–5
3	3	62	+2
4	6	107	–3
5	6	152	+3
6	8	32	–4
7	8	54	+1
8	9	195	+2



loci pattern across genome

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

	Chromosome										
m	**1**	2	**3**	4	5	**6**	7	**8**	**9**	10	**Count of 8000**
8	**2**	**0**	**1**	**0**	**0**	**2**	**0**	**2**	**1**	**0**	3371
9	*3*	0	1	0	0	2	0	2	1	0	751
7	2	0	1	0	0	2	0	*1*	1	0	377
9	2	0	1	0	0	2	0	2	1	0	218
9	2	0	1	0	0	*3*	0	2	1	0	218
9	2	0	1	0	0	2	0	2	*2*	0	198

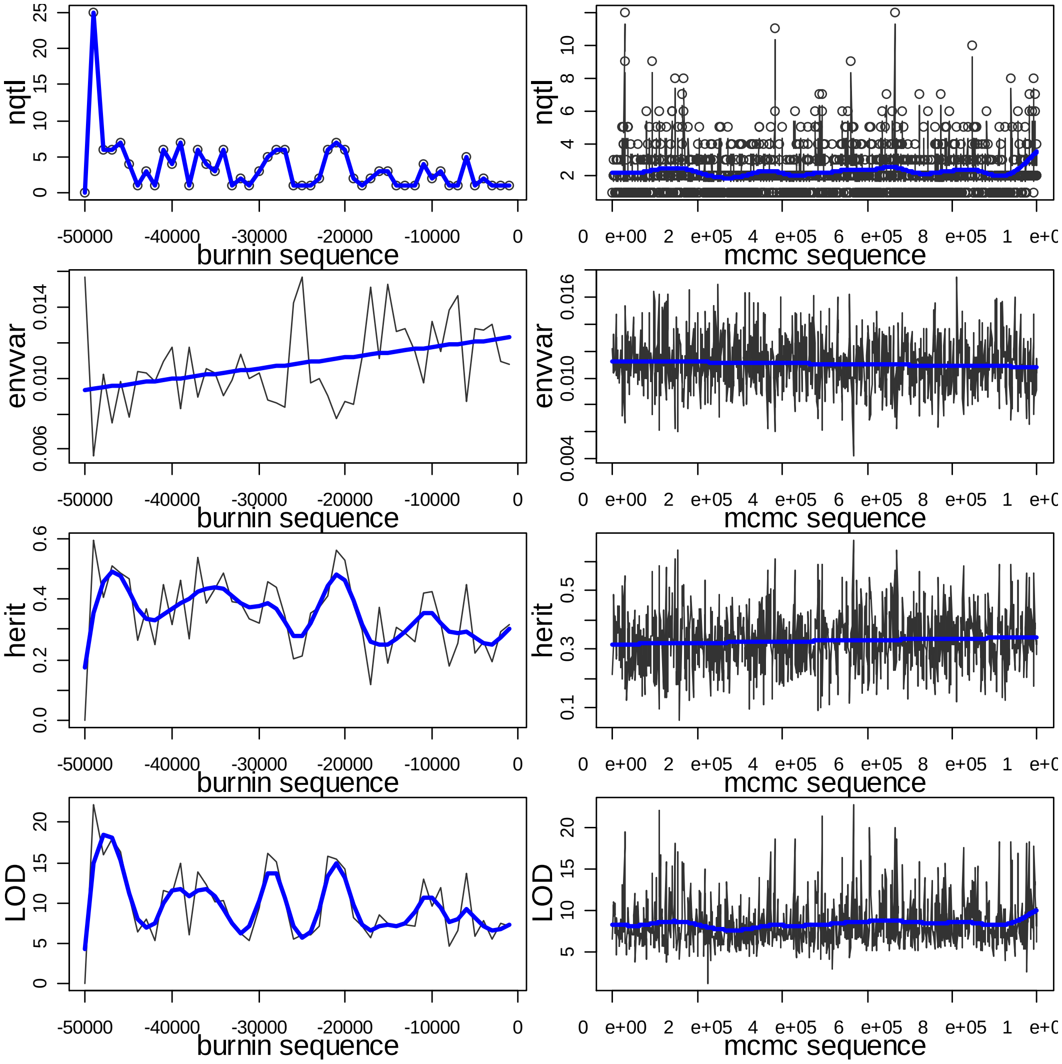
B. napus 8-week vernalization whole genome study

- 108 plants from double haploid
  - similar genetics to backcross: follow 1 gamete
  - parents are Major (biennial) and Stellar (annual)
- 300 markers across genome
  - 19 chromosomes
  - average 6cM between markers
    - median 3.8cM, max 34cM
  - 83% markers genotyped
- phenotype is days to flowering
  - after 8 weeks of vernalization (cooling)
  - Stellar parent requires vernalization to flower
- available in R/bim package
- Ferreira et al. (1994); Kole et al. (2001); Schranz et al. (2002)

Markov chain Monte Carlo sequence

burnin (sets up chain)
mcmc sequence

number of QTL
environmental variance
h^2 = heritability
(genetic/total variance)
LOD = likelihood



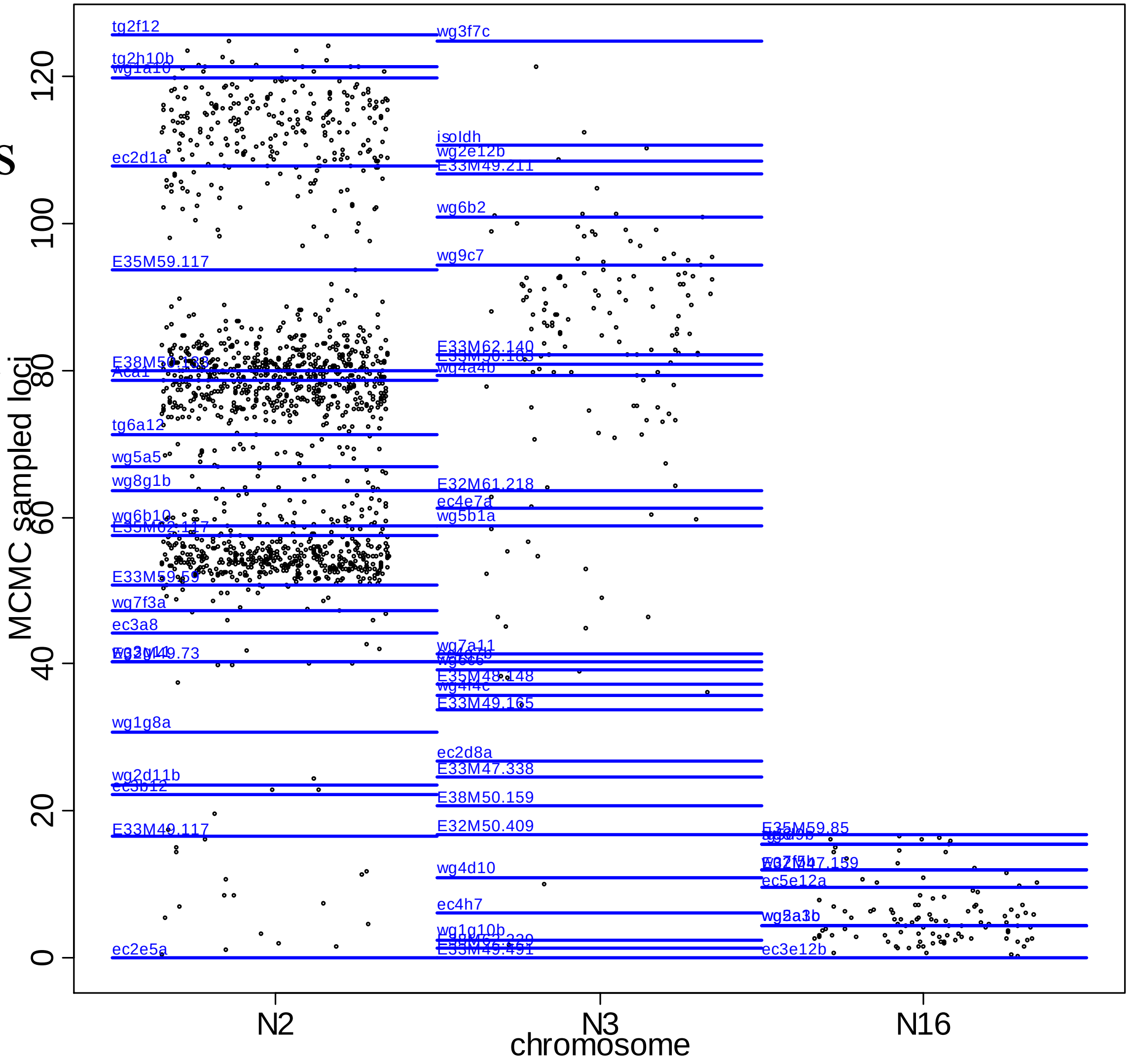
MCMC sampled loci

subset of chromosomes
  N2, N3, N16

points jittered for view
blue lines at markers

note concentration
on chromosome N2

includes all models

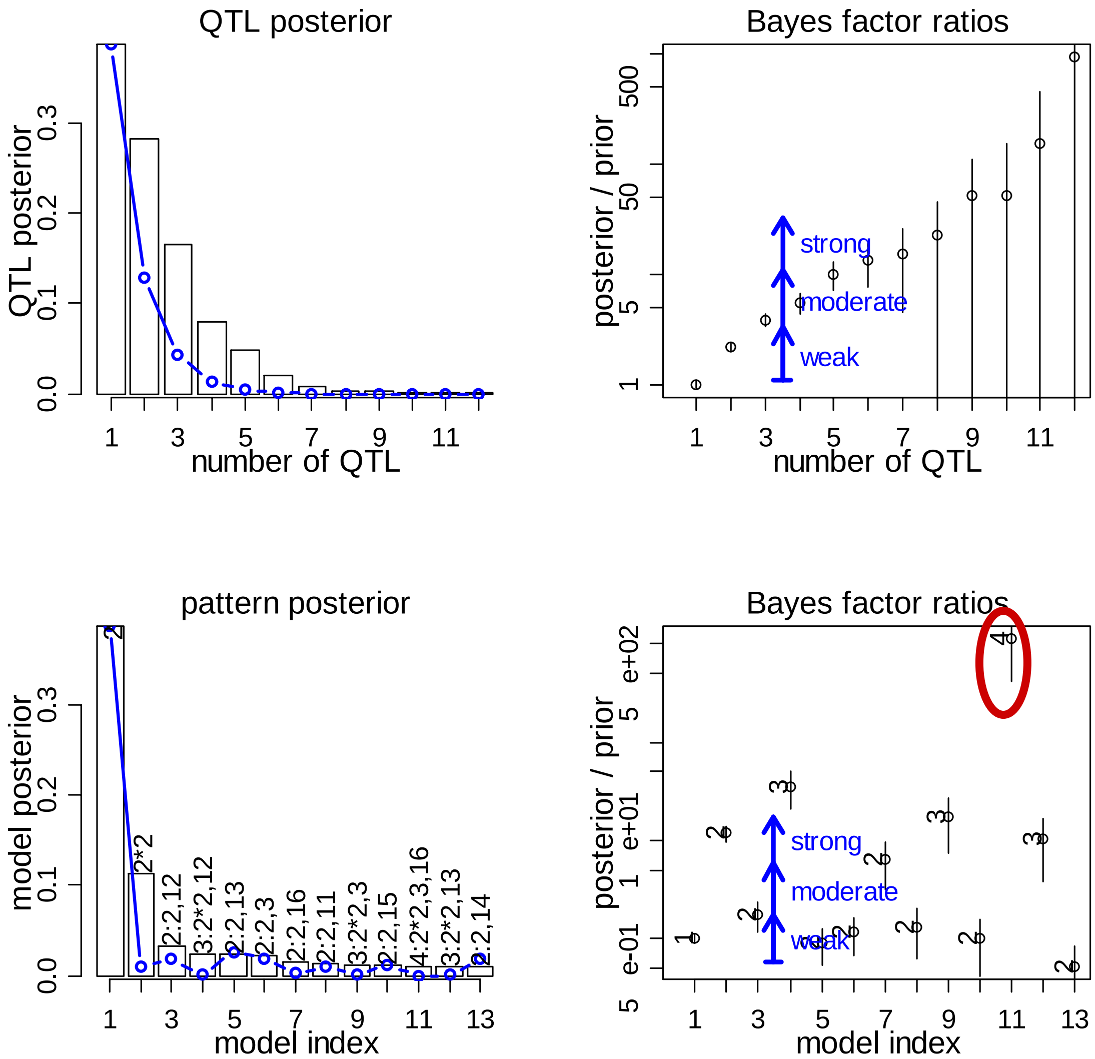


Bayesian model assessment

row 1: # QTL
row 2: pattern

col 1: posterior
col 2: Bayes factor
note error bars on bf

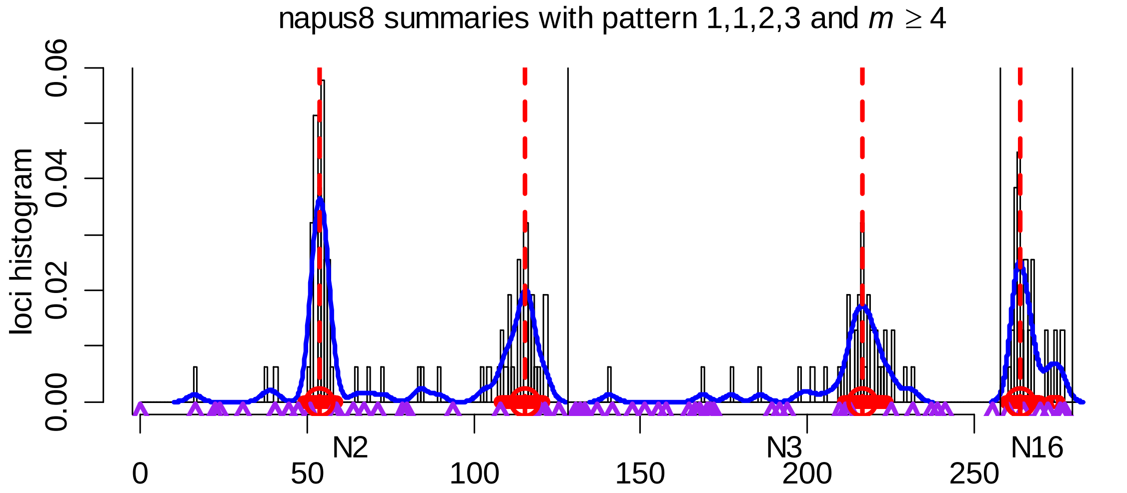
evidence suggests
 4-5 QTL
 N2(2-3),N3,N16



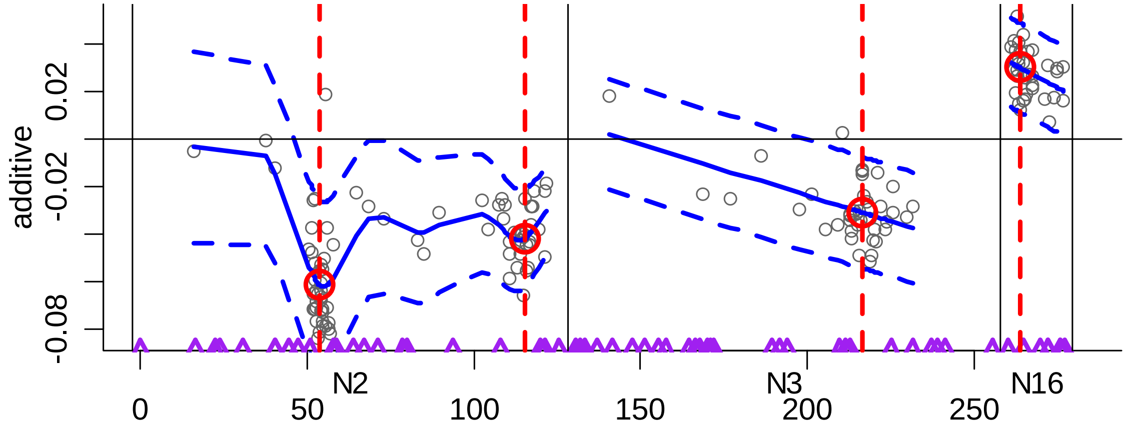
Bayesian estimates of loci & effects

model averaging: at least 4 QTL

histogram of loci
blue line is density
red lines at estimates



estimate additive effects
  (red circles)
grey points sampled
  from posterior
blue line is cubic spline
dashed line for 2 SD

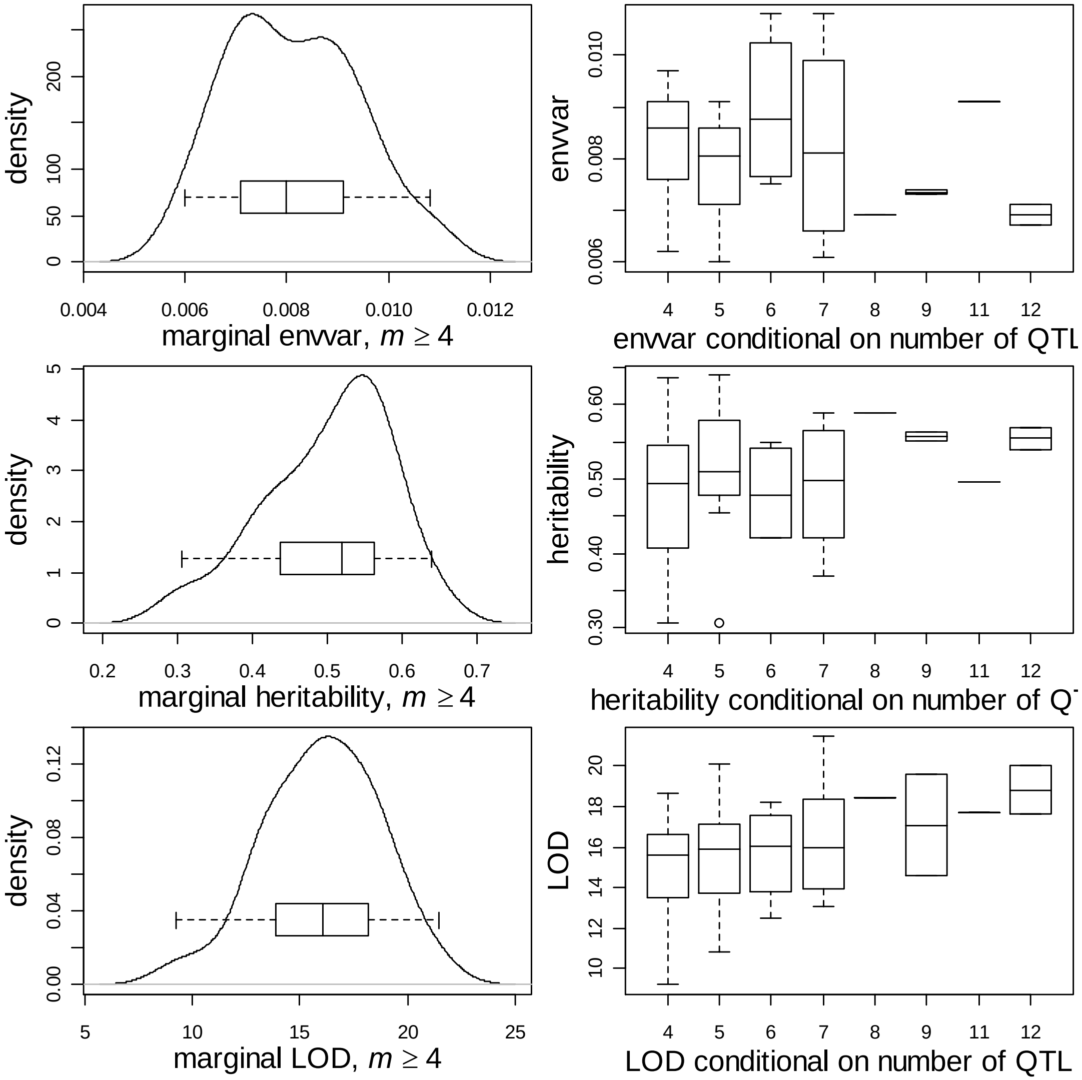


Bayesian model diagnostics

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

environmental variance
$\sigma^2 = .008, \sigma = .09$
heritability
$h^2 = 52\%$
LOD = 16
(highly significant)

but note change with *m*



Bayesian software for QTLs

- R/bim (Satagopan Yandell 1996; Gaffney 2001)
  - www.stat.wisc.edu/~yandell/qtl/software/Bmapqtl
  - www.r-project.org contributed package
  - version available within WinQTLCart (statgen.ncsu.edu/qtlcart)
- Bayesian IM with epistasis (Nengjun Yi, U AB)
  - separate C++ software (papers with Xu)
  - plans in progress to incorporate into R/bim
- R/qtl (Broman et al. 2003)
  - biosun01.biostat.jhsph.edu/~kbroman/software
  - www.r-project.org contributed package
- Pseudomarker (Sen Churchill 2002)
  - www.jax.org/staff/churchill/labsite/software
- Bayesian QTL / Multimapper
  - Sillanpää Arjas (1998)
  - www.rni.helsinki.fi/~mjs
- Stephens & Fisch (email)

R/bim: our software

- www.stat.wisc.edu/~yandell/qtl/software/Bmapqtl
  - R contributed library (www.r-project.org)
    - library(bim) is cross-compatible with library(qtl)
  - Bayesian module within WinQTLCart
    - WinQTLCart output can be processed using R library
- Software history
  - initially designed by JM Satagopan (1996)
  - major revision and extension by PJ Gaffney (2001)
    - whole genome
    - multivariate update of effects; long range position updates
    - substantial improvements in speed, efficiency
    - pre-burnin: initial prior number of QTL very large
  - upgrade (H Wu, PJ Gaffney, CF Jin, BS Yandell 2003)
  - epistasis in progress (H Wu, BS Yandell, N Yi 2004)

many thanks

Michael Newton
Daniel Sorensen
Daniel Gianola
Liang Li
Hong Lan
Hao Wu
Nengjun Yi
David Allison

Tom Osborn
David Butruille
Marcio Ferrera
Josh Udahl
Pablo Quijada
Alan Attie
Jonathan Stoehr
Gary Churchill

Jaya Satagopan
Fei Zou
Patrick Gaffney
Chunfang Jin
Yang Song
Elias Chaibub Neto
Xiaodan Wei

USDA Hatch, NIH/NIDDK, Jackson Labs