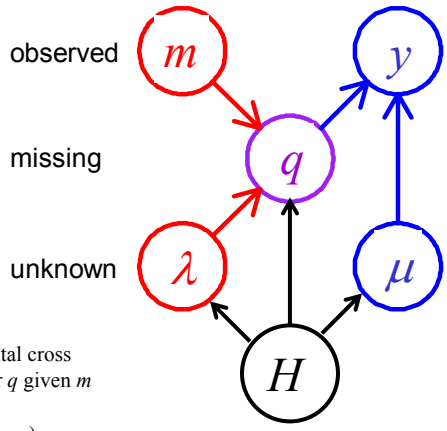
Bayesian Interval Mapping

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



after
Sen Churchill (2001)

1. Bayesian strategy for QTL study

- augment data (y,m) with missing genotypes q
- study unknowns (μ,λ,A) given augmented data (y,m,q)
  - find better genetic architectures A
  - find most likely genomic regions = QTL = λ
  - estimate phenotype parameters = genotype means = μ
- 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)

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

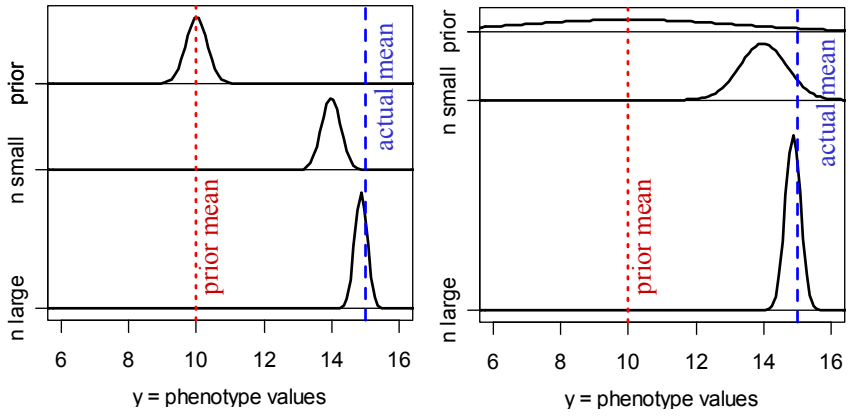
$$\text{posterior for } q,\mu,\lambda,A = \frac{\text{phenotype likelihood} * [\text{prior for } q,\mu,\lambda,A]}{\text{constant}}$$

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

Bayesian idea

- Reverend Thomas Bayes (1702-1761)
  - part-time mathematician
  - buried in Bunhill Cemetary, 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

Bayes posterior for normal data



small prior variance

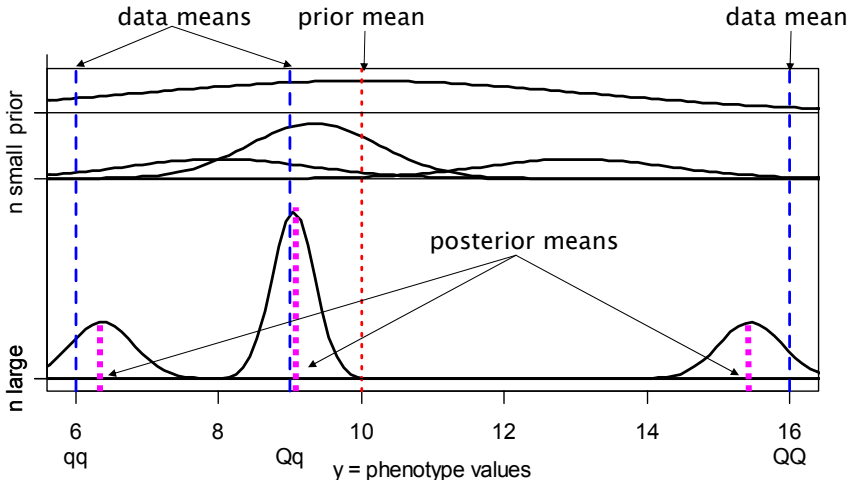
large prior variance

Bayes posterior for 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	$\mu \sim N\left(b_n \bar{y}_\bullet + (1-b_n)\mu_0, b_n\sigma^2/n\right)$ with $\bar{y}_\bullet = \sum_{\{i=1,\ldots,n\}} y_i/n$
fudge factor (shrinks to 1)	$b_n = \frac{\kappa n}{\kappa n + 1} \to 1$

what values are the genotypic means?

(phenotype mean for genotype q is μ_q)



Bayes posterior QTL means

posterior centered on sample genotypic mean
but shrunken slightly toward overall mean

prior: $\mu_q \sim N\left(\bar{y}_\bullet, \kappa\sigma^2\right)$

posterior: $\mu_q \sim N\left(b_q\bar{y}_q + (1-b_q)\bar{y}_\bullet, b_q\sigma^2/n_q\right)$

$$n_q = \text{count}\{q_i = q\}, \bar{y}_q = \underset{\{q_i=q\}}{\text{sum}} y_i / n_q$$

fudge factor: $b_q = \dfrac{\kappa n_q}{\kappa n_q + 1} \to 1$

QTL with epistasis

- same phenotype model overview

$$Y = \mu_q + e, \text{var}(e) = \sigma^2$$

- partition of genotypic value with epistasis

$$\mu_q = \mu + \beta_{q1} + \beta_{q2} + \beta_{q12}$$

- partition of genetic variance & heritability

$$\text{var}(\mu_q) = \sigma_q^2 = \sigma_1^2 + \sigma_2^2 + \sigma_{12}^2$$

$$h_q^2 = \frac{\sigma_q^2}{\sigma_q^2 + \sigma^2} = h_1^2 + h_2^2 + h_{12}^2$$

partition of multiple QTL effects

- partition genotype-specific mean into QTL effects
  - μ_q = mean + main effects + epistatic interactions
  - $\mu_q = \mu + \beta_q = \mu + \text{sum}_{j \text{ in } A}\, \beta_{qj}$
- priors on mean and effects

μ	~ $N(\mu_0, \kappa_0\sigma^2)$	grand mean
β_q	~ $N(0, \kappa_1\sigma^2)$	model-independent genotypic effect
β_{qj}	~ $N(0, \kappa_1\sigma^2/\|A\|)$	effects down-weighted by size of A

- determine hyper-parameters via empirical Bayes

$$\mu_0 \approx \overline{Y}_\bullet \text{ and } \kappa_1 \approx \frac{h_q^2}{1 - h_q^2} = \frac{\sigma_q^2}{\sigma^2}$$

posterior mean ≈ LS estimate

$$\mu_q \mid Y, m \sim N(B_q \hat{\mu}_q, B_q C_q \sigma^2)$$

$$\approx N(\hat{\mu}_q, C_q \sigma^2)$$

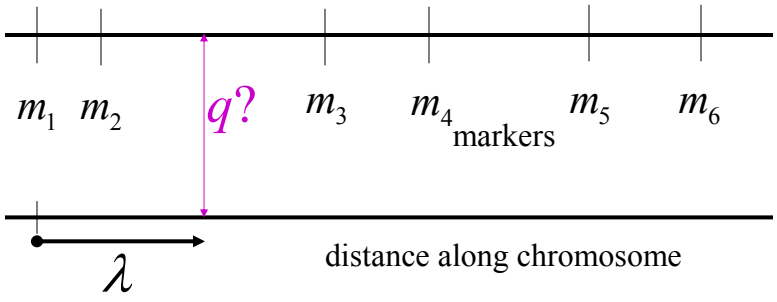
LS estimate $\hat{\mu}_q = \text{sum}_i[\text{sum}_{j \in M} \hat{\beta}_{qji}] = \text{sum}_i w_{qi} Y$

variance $V(\hat{\mu}_q) = \text{sum}_i w_{qi}^2 \sigma^2 = C_q \sigma^2$

shrinkage $B_q = \kappa/(\kappa + C_q) \to 1$

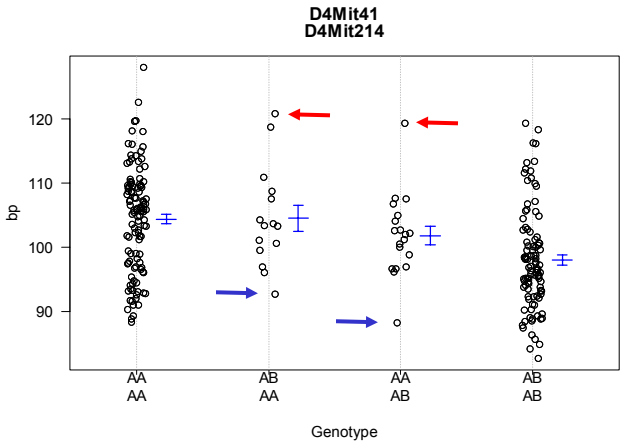
pr($q|m,\lambda$) recombination model

pr($q|m,\lambda$) = pr(geno | map, locus) ≈
pr(geno | flanking markers, locus)



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

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

Where are the 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

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

what is the genetic architecture A?

- 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 μ
  - 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 λ
  - prior is flat across genome (all loci equally likely)
- sampling QTL model A
  - 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 λ given q,A (using Metropolis-Hastings step)
  - sample genotypes q given λ,μ,y,A (using Gibbs sampler)
  - sample effects μ given q,y,A (using Gibbs sampler)
  - sample QTL model A given λ,μ,y,q (using Gibbs or M-H)

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

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

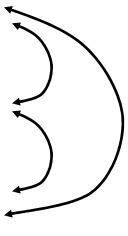
MCMC sampling of (λ, q, μ)

- Gibbs sampler
  - genotypes q
  - effects μ
  - *not* loci λ

$$q \sim \text{pr}(q \mid y_i, m_i, \mu, \lambda)$$

$$\mu \sim \frac{\text{pr}(y \mid q, \mu)\text{pr}(\mu)}{\text{pr}(y \mid q)}$$

$$\lambda \sim \frac{\text{pr}(q \mid m, \lambda)\text{pr}(\lambda \mid m)}{\text{pr}(q \mid m)}$$



- Metropolis-Hastings sampler
  - extension of Gibbs sampler
  - does not require normalization
    - pr(q | m) = sum$_\lambda$ pr(q | m, λ) pr(λ)

Gibbs sampler
for two genotypic means

- want to study two correlated effects
  - could sample directly from their bivariate distribution
  - assume correlation ρ 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

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

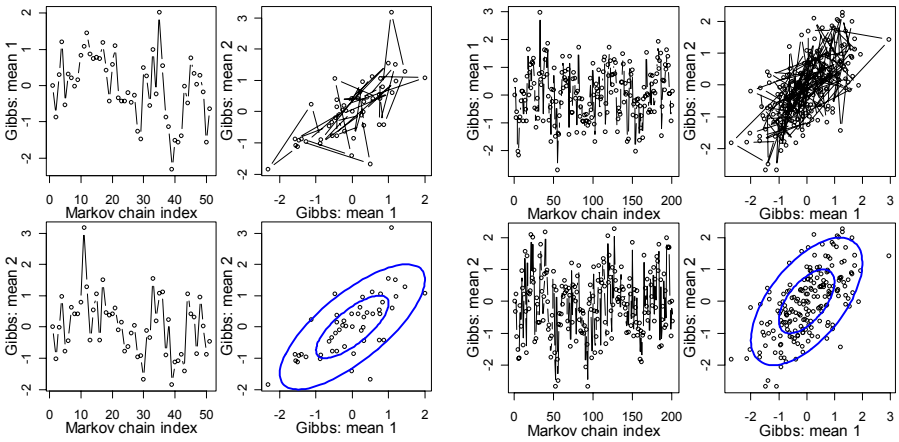
$$\mu_1 \sim N\left(\rho\mu_2, 1-\rho^2\right)$$

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

Gibbs sampler samples: $\rho = 0.6$

$N = 50$ samples

$N = 200$ samples



full conditional for locus

- cannot easily sample from locus full conditional

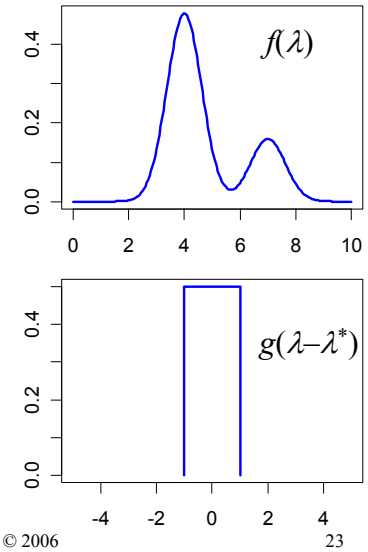
  $\text{pr}(\lambda \mid y,m,\mu,q) = \text{pr}(\lambda \mid m,q)$

  $= \text{pr}(q \mid m, \lambda)\ \text{pr}(\lambda) / \text{constant}$
- constant is very difficult to compute explicitly
  - must average over all possible loci λ 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

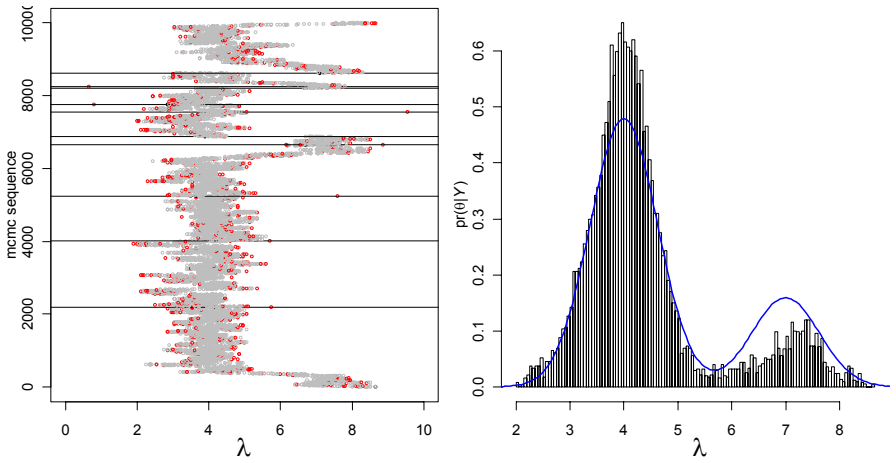
Metropolis-Hastings idea

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

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

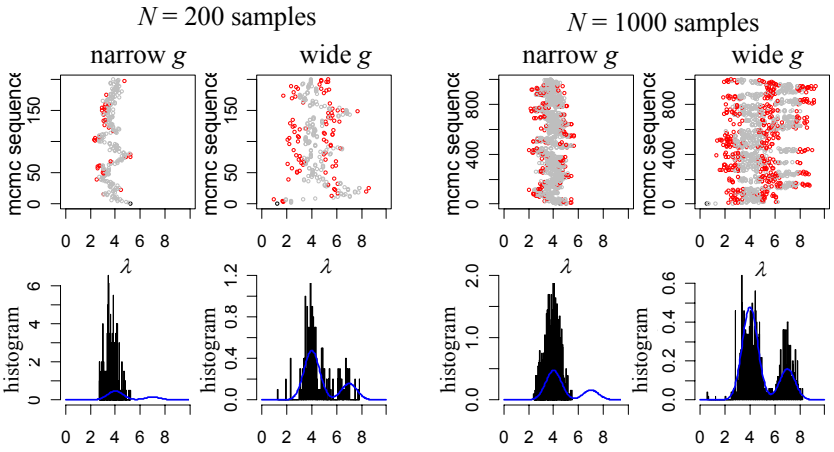


Metropolis-Hastings for locus λ



added twist: occasionally propose from entire genome

Metropolis-Hastings samples



3. sampling genetic architectures

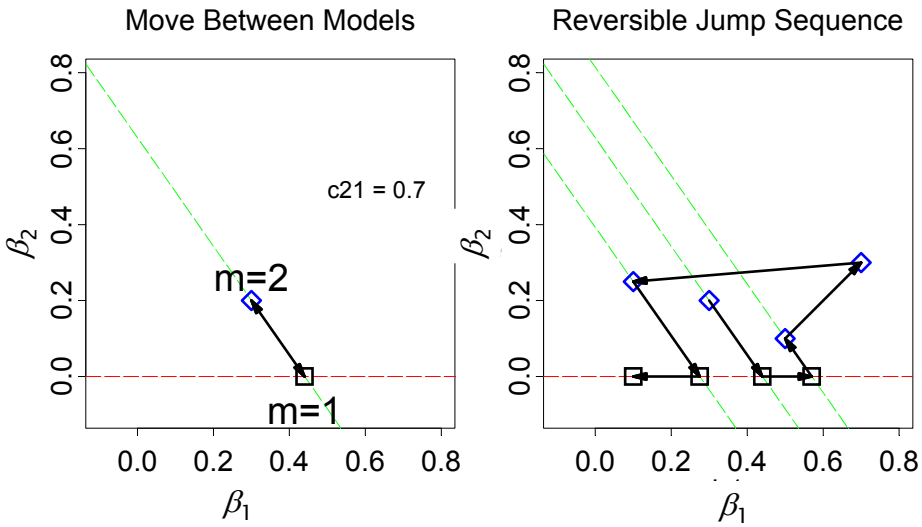
- search across genetic architectures *A* of various sizes
  - allow change in number of QTL
  - allow change in types of epistatic interactions
- methods for search
  - reversible jump MCMC
  - Gibbs sampler with loci indicators
- complexity of epistasis
  - Fisher-Cockerham effects model
  - general multi-QTL interaction & limits of inference

reversible jump MCMC

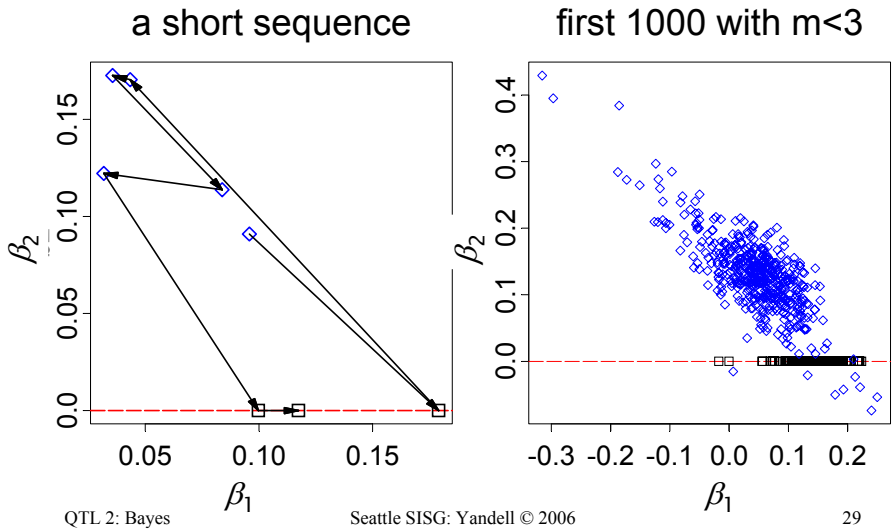
- consider known genotypes q at 2 known loci λ
  - models with 1 or 2 QTL
- M-H step between 1-QTL and 2-QTL models
  - model changes dimension (via careful bookkeeping)
  - consider mixture over QTL models H

$$n.qtl = 1 : Y = \beta_0 + \beta_{q1} + e$$
$$n.qtl = 2 : Y = \beta_0 + \beta_{q1} + \beta_{q2} + e$$

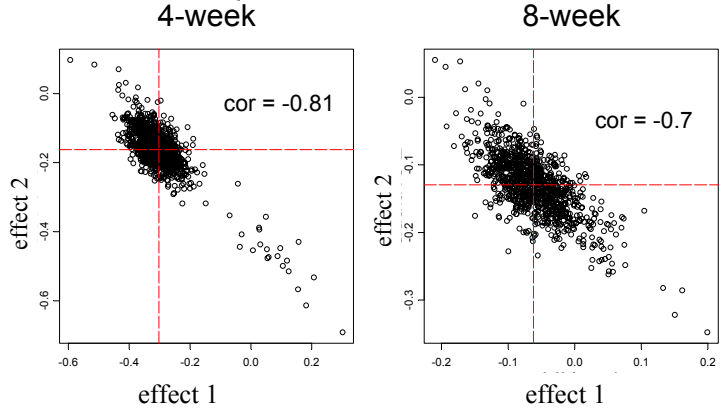
geometry of reversible jump



geometry allowing q and λ to change



collinear QTL = correlated effects



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

sampling across QTL models *A*

action steps: draw one of three choices

- update QTL model *A* with probability 1-*b*(*A*)-*d*(*A*)
  - update current model using full conditionals
  - sample QTL loci, effects, and genotypes
- add a locus with probability *b*(*A*)
  - 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*(*A*)
  - 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
  - $\delta = 1$ if QTL present
  - $\delta = 0$ if no QTL present
- Gibbs sampler on loci indicators δ
  - relatively easy to incorporate epistasis
  - Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005 *Genetics*)
    - (see earlier work of Nengjun Yi and Ina Hoeschele)

$$\mu_q = \mu + \delta_1 \beta_{q1} + \delta_2 \beta_{q2}$$

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

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

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

4. Bayesian QTL model selection

- Bayes factor details
- Bayesian model averaging
- false discovery rate (FDR)

Bayes factors

- ratio of model likelihoods
  - ratio of posterior to prior odds for architectures
  - averaged over unknowns

$$B_{12} = \frac{\mathrm{pr}(A_1 \mid y, m) / \mathrm{pr}(A_2 \mid y, m)}{\mathrm{pr}(A_1) / \mathrm{pr}(A_2)} = \frac{\mathrm{pr}(y \mid m, A_1)}{\mathrm{pr}(y \mid m, A_2)}$$

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

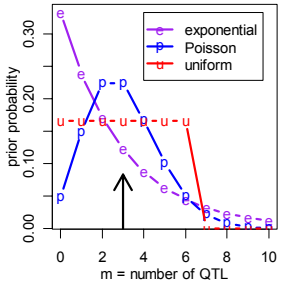
$$-2\log(B_{12}) = -2\log(LR) - (p_2 - p_1)\log(n)$$

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
  - sample posterior using MCMC
  - posterior pr(A | *y, m*) is marginal histogram

Bayes factors and genetic model A

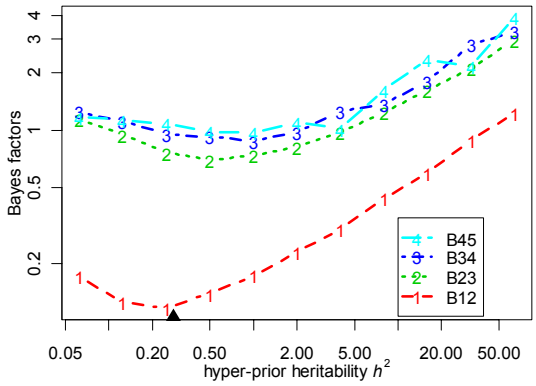
- $|A|$ = number of QTL
  - prior pr(A) chosen by user
  - posterior pr($A|y,m$)
    - sampled marginal histogram
    - shape affected by prior pr(A)



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

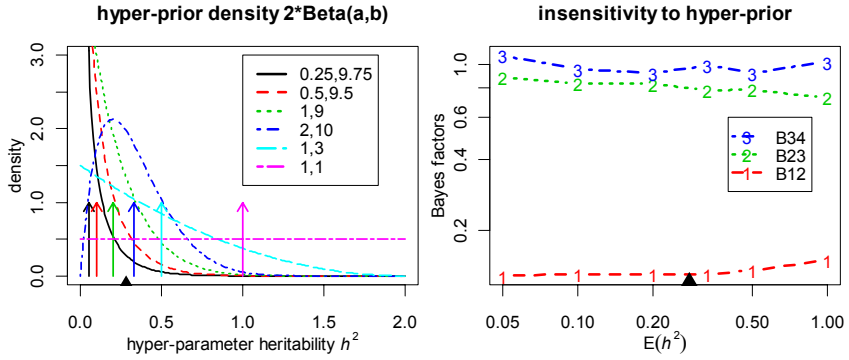
- pattern of QTL across genome
- gene action and epistasis

BF sensitivity to fixed prior for effects



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

BF insensitivity to random effects prior



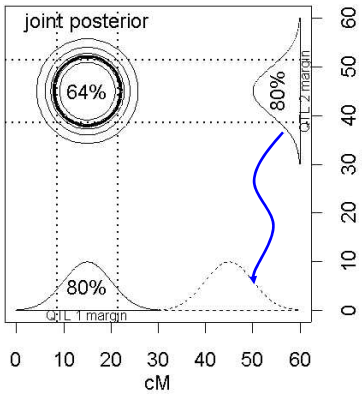
$$\beta_{qj} \sim \mathrm{N}\left(0, \sigma_G^2 / m\right), \sigma_G^2 = h^2 \sigma_{\text{total}}^2, \tfrac{1}{2} h^2 \sim \mathrm{Beta}(a,b)$$

Bayesian model averaging

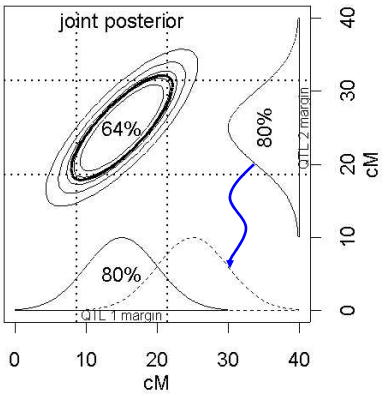
- average summaries over multiple architectures
- avoid selection of "best" model
- focus on "better" models
- examples in data talk later

1-D and 2-D marginals
pr(QTL at λ | *Y,X, m*)

unlinked loci



linked loci



false detection rates and thresholds

- multiple comparisons: test QTL across genome
  - size = pr(LOD(λ) > threshold | no QTL at λ)
  - threshold guards against a single false detection
    - very conservative on genome-wide basis
  - difficult to extend to multiple QTL
- positive false discovery rate (Storey 2001)
  - pFDR = pr(no QTL at λ | LOD(λ) > threshold)
  - Bayesian posterior HPD region based on threshold
    - $\Lambda = \{\lambda \mid \text{LOD}(\lambda) > \text{threshold}\} \approx \{\lambda \mid \text{pr}(\lambda \mid Y,X,m) \text{ large}\}$
  - extends naturally to multiple QTL

pFDR and QTL posterior

- positive false detection rate
  - pFDR = pr(no QTL at λ | *Y,X,* λ in Λ)
  - pFDR = $\frac{\text{pr}(H=0)*\text{size}}{\text{pr}(m=0)*\text{size}+\text{pr}(m>0)*\text{power}}$
  - power = posterior = pr(QTL in Λ | *Y,X, m*>0)
  - size = (length of Λ) / (length of genome)
- extends to other model comparisons
  - m = 1 vs. m = 2 or more QTL
  - pattern = ch1,ch2,ch3 vs. pattern > 2*ch1,ch2,ch3

pFDR for SCD1 analysis