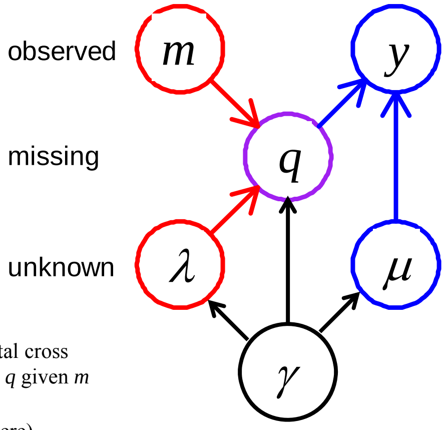
Bayesian Interval Mapping

1. Bayesian strategy
2. Markov chain sampling
3. sampling genetic architectures
4. criteria for model selection

QTL model selection: key players

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



after
Sen Churchill (2001)

1. Bayesian strategy for QTL study

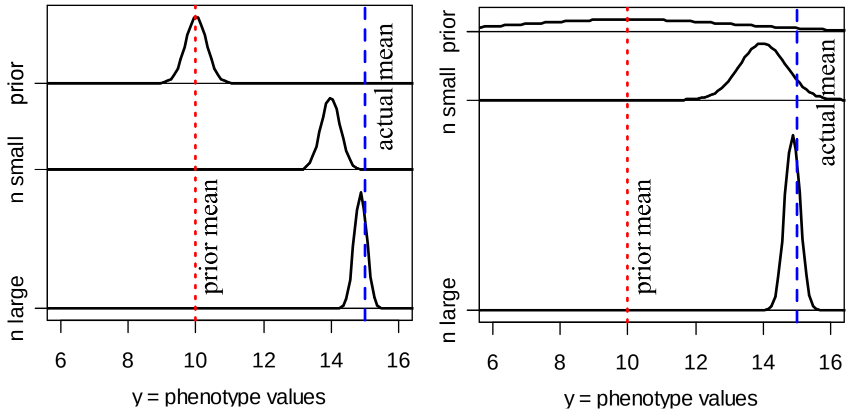
- augment data (*y*,*m*) with missing genotypes *q*
- study unknowns (μ,λ,γ) given augmented data (*y*,*m*,*q*)
  - find better genetic architectures γ
  - find most likely genomic regions = QTL = λ
  - estimate phenotype parameters = genotype means = μ
- sample from posterior in some clever way
  - multiple imputation (Sen Churchill 2002)
  - Markov chain Monte Carlo (MCMC)
    - (Satagopan et al. 1996; Yi et al. 2005, 2007)

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

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

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

Bayes posterior for normal data



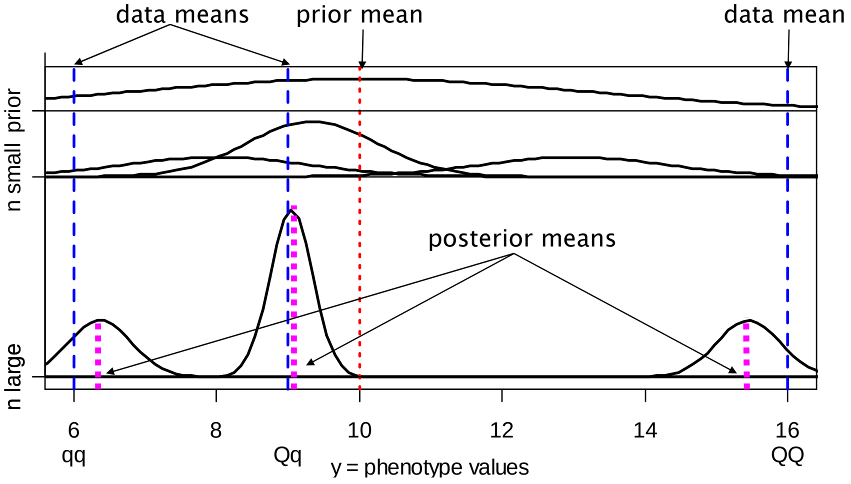
small prior variance large prior variance

Bayes posterior for normal data

model	$y_i = \mu + e_i$
environment	$e \sim N(0, \sigma^2), \sigma^2$ known
likelihood	$y \sim N(\mu, \sigma^2)$
prior	$\mu \sim N(\mu_0, \kappa\sigma^2), \kappa$ known
posterior:	mean tends to sample mean
single individual	$\mu \sim N(\mu_0 + b_1(y_1 - \mu_0), b_1\sigma^2)$
sample of n individuals	$\mu \sim N\left(b_n \bar{y}_{\bullet} + (1-b_n)\mu_0, b_n\sigma^2 / n\right)$ with $\bar{y}_{\bullet} = \underset{\{i=1,\ldots,n\}}{\text{sum}} y_i / n$
shrinkage factor (shrinks to 1)	$b_n = \dfrac{\kappa n}{\kappa n + 1} \to 1$

what values are the genotypic means?

phenotype model pr($y|q,\mu$)



Bayes posterior QTL means

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

phenotype mean: $E(y \mid q) \quad = \quad \mu_q \qquad V(y \mid q) = \sigma^2$

genotypic prior: $E(\mu_q) \quad = \quad \bar{y}_{\bullet} \qquad V(\mu_q) = \kappa\sigma^2$

posterior: $E(\mu_q \mid y) \quad = \quad b_q \bar{y}_q + (1 - b_q)\bar{y}_{\bullet} \qquad V(\mu_q \mid y) = b_q \sigma^2 / n_q$

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

shrinkage: $b_q \quad = \quad \dfrac{\kappa n_q}{\kappa n_q + 1} \to 1$

partition genotypic effects on phenotype

- phenotype depends on genotype
- genotypic value partitioned into
  - main effects of single QTL
  - epistasis (interaction) between pairs of QTL

$$
\begin{aligned}
\mu_q &= \beta_0 + \beta_q = E(Y;q) \\
\beta_q &= \beta(q_2) + \beta(q_2) + \beta(q_1, q_2)
\end{aligned}
$$

partitition genotypic variance

- consider same 2 QTL + epistasis

- centering variance $V(\beta_0) = \kappa_0 \sigma^2 = s^2$

- genotypic variance $V(\beta_q) = \kappa_1 \sigma^2 = \sigma_q^2 = \sigma_1^2 + \sigma_2^2 + \sigma_{12}^2$

- heritability $h_q^2 = \dfrac{\sigma_q^2}{\sigma_q^2 + \sigma^2} = h_1^2 + h_2^2 + h_{12}^2$

posterior mean ≈ LS estimate

$$\beta_q \mid y \sim N(b_q \hat{\beta}_q, b_q C_q \sigma^2)$$

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

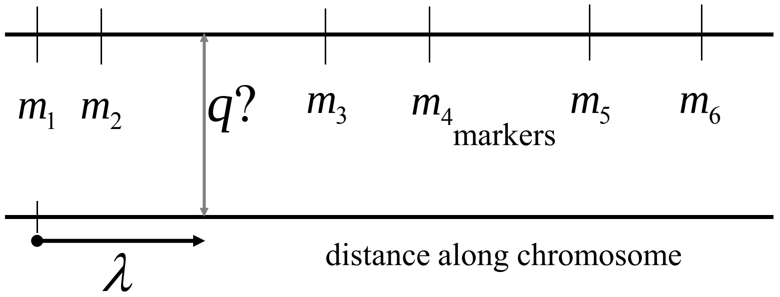
LS estimate $\hat{\beta}_q = \text{sum}_i[\text{sum}_j \hat{\beta}(q_{ij})] = \text{sum}_i w_{qi} y_i$

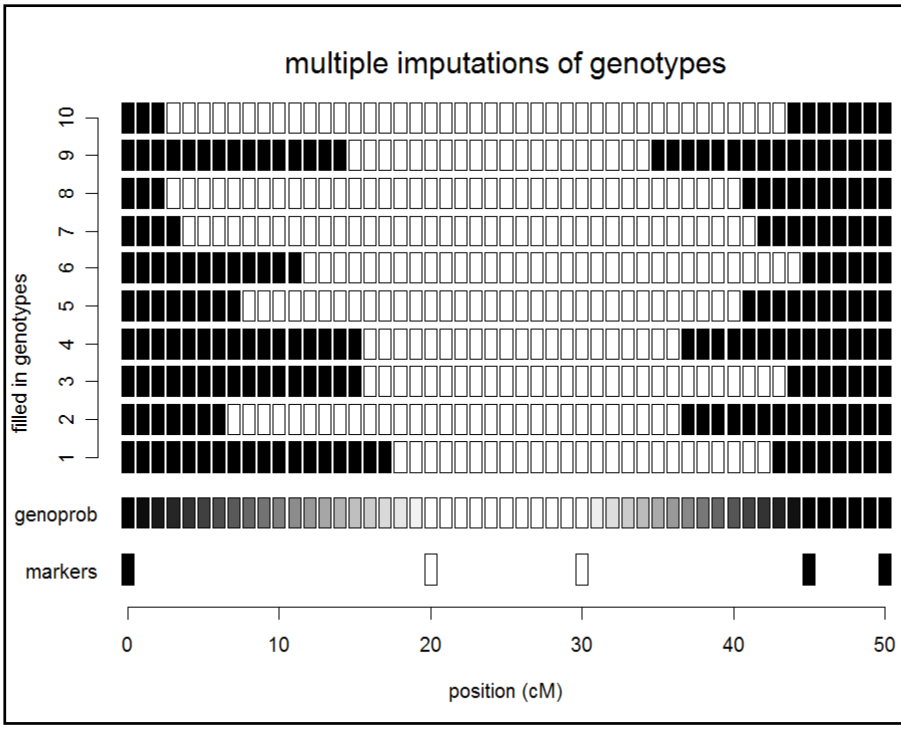
variance $V(\hat{\beta}_q) = \text{sum}_i w_{qi}^2 \sigma^2 = C_q \sigma^2$

shrinkage $b_q = \kappa_1 / (\kappa_1 + C_q) \to 1$

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

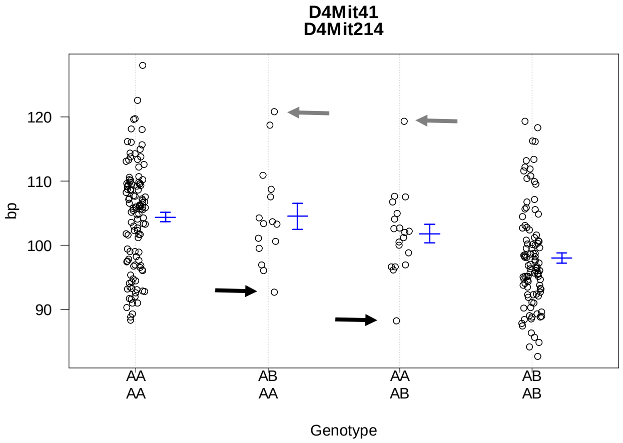
pr($q|m,\lambda$) = pr(geno | map, locus) ≈
pr(geno | flanking markers, locus)





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 | m, λ)
    - weight toward *q* that agrees with flanking markers
  - proportional to likelihood pr(y | q, μ)
    - 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*

  $\mathrm{pr}(\lambda \mid m,q) = \mathrm{pr}(\lambda)\ \mathrm{pr}(q \mid m,\lambda)$ / 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 γ?

- 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

γ = genetic architecture:

loci:

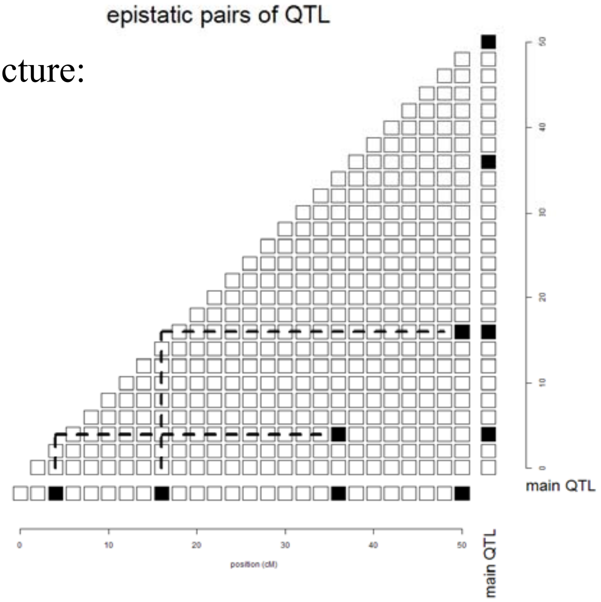
main QTL

epistatic pairs

effects:

add, dom

aa, ad, dd



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

$$(\lambda,q,\mu,\gamma) \sim \mathrm{pr}(\lambda,q,\mu,\gamma \mid y,m)$$

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

MCMC sampling of unknowns (q,μ,λ)

for given genetic architecture γ

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

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

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

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

- Metropolis-Hastings sampler
  - extension of Gibbs sampler
  - does not require normalization
    - pr(q | m) = sum_λ 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

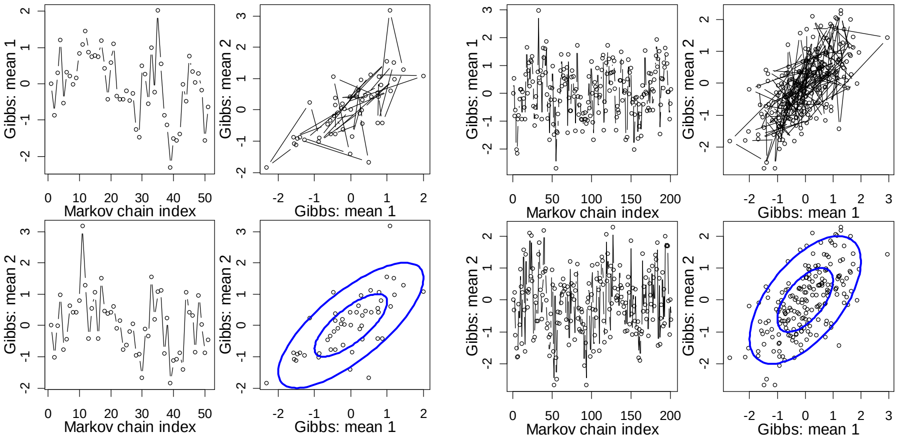
$$\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} \sim N\left(\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$

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

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

Gibbs sampler samples: $\rho = 0.6$

$N = 50$ samples $N = 200$ samples



full conditional for locus

- cannot easily sample from locus full conditional

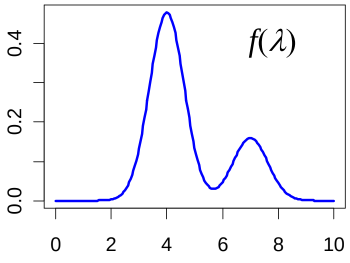
  $\text{pr}(\lambda \,|y,m,\mu,q) \quad = \text{pr}(\ \lambda \mid m,q)$

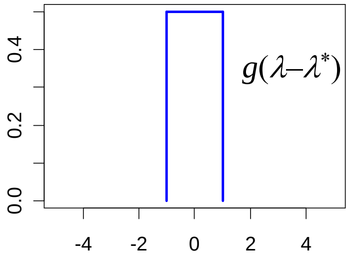
  $\qquad\qquad\qquad\qquad = \text{pr}(\ q \mid m,\ \lambda\ )\ \text{pr}(\lambda\ )\ /\ \text{constant}$
- constant is very difficult to compute explicitly
  - must average over all possible loci λ over genome
  - must do this for every possible genotype q
- Gibbs sampler will not work in general
  - but can use method based on ratios of probabilities
  - Metropolis-Hastings is extension of Gibbs sampler

Metropolis-Hastings idea

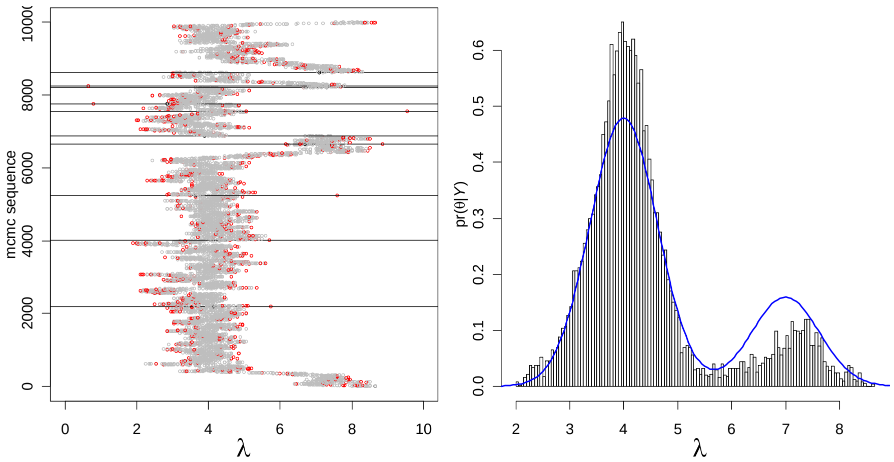
- want to study distribution $f(\lambda)$
  - take Monte Carlo samples
    - unless too complicated
  - take samples using ratios of f
- Metropolis-Hastings samples:
  - propose new value λ^*
    - near (?) current value λ
    - from some distribution g
  - accept new value with prob a
    - Gibbs sampler: a = 1 always

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



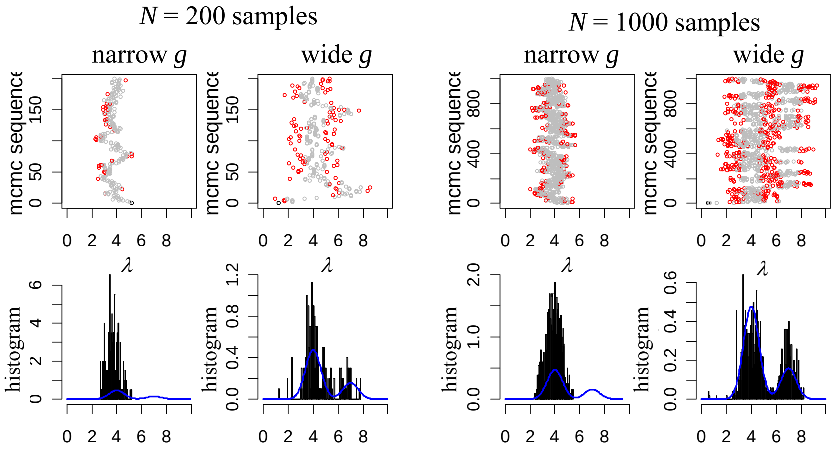


Metropolis-Hastings for locus λ



added twist: occasionally propose from entire genome

Metropolis-Hastings samples



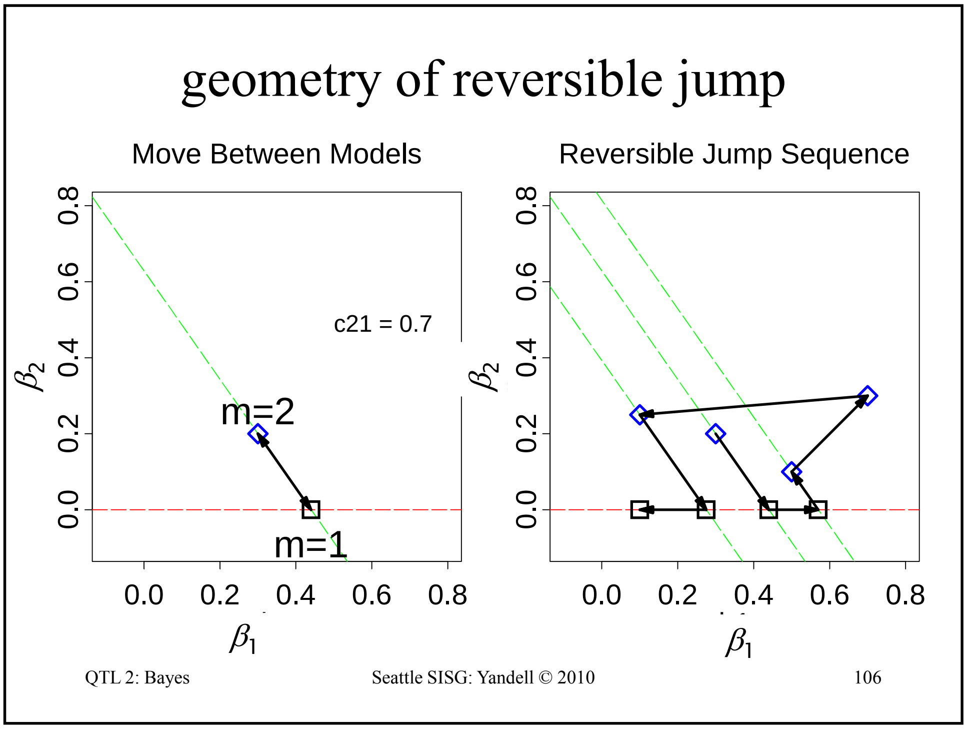
3. sampling genetic architectures

- search across genetic architectures γ 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

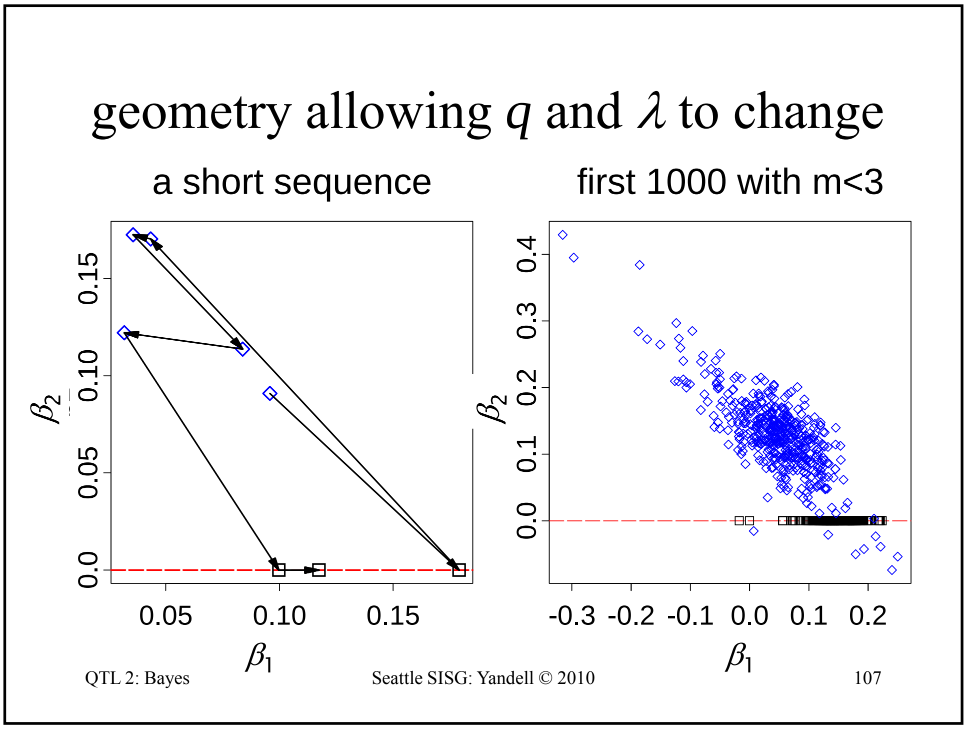
reversible jump MCMC

- 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

$$\begin{tikzcd} \gamma = 1\,\text{QTL} : Y = \beta_0 + \beta(q_1) + e \arrow[d, bend right, leftrightarrow] \\ \gamma = 2\,\text{QTL} : Y = \beta_0 + \beta(q_1) + \beta(q_2) + e \end{tikzcd}$$

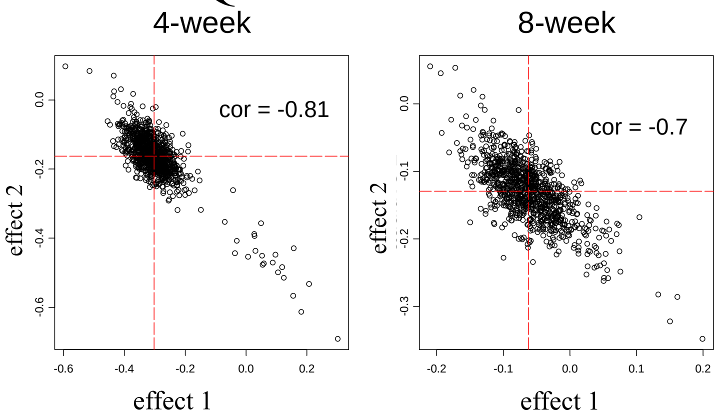


geometry of reversible jump



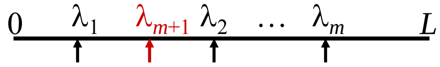
geometry allowing q and λ to change

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 γ



action steps: draw one of three choices

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

Gibbs sampler with loci indicators

- consider only QTL at pseudomarkers
  - every 1-2 cM
  - modest approximation with little bias
- use loci indicators in each pseudomarker
  - $\gamma = 1$ if QTL present
  - $\gamma = 0$ if no QTL present
- Gibbs sampler on loci indicators γ
  - relatively easy to incorporate epistasis
  - Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005 *Genetics*)
    - (see earlier work of Nengjun Yi and Ina Hoeschele)

$$\mu_q = \mu + \gamma_1 \beta(q_1) + \gamma_2 \beta(q_2),\ \ \gamma_k = 0{,}1$$

Bayesian shrinkage estimation

- soft loci indicators
  - strength of evidence for λ_j depends on γ
  - $0 \le \gamma \le 1$ (grey scale)
  - shrink most γs to zero
- Wang et al. (2005 *Genetics*)
  - Shizhong Xu group at U CA Riverside

$$\mu_q = \beta_0 + \gamma_1 \beta_1(q_1) + \gamma_2 \beta_2(q_1),\ \ 0 \le \gamma_k \le 1$$

other model selection approaches

- include all potential loci in model
- assume “true” model is “sparse” in some sense
- Sparse partial least squares
  - Chun, Keles (2009 *Genetics*; 2010 *JRSSB*)
- LASSO model selection
  - Foster (2006); Foster Verbyla Pitchford (2007 *JABES*)
  - Xu (2007 *Biometrics*); Yi Xu (2007 *Genetics*)
  - Shi Wahba Wright Klein Klein (2008 *Stat & Infer*)

4. criteria for model selection

balance fit against complexity

- classical information criteria
  - penalize likelihood L by model size $|\gamma|$
  - IC = $-2 \log L(\gamma \mid y)$ + penalty(γ)
  - maximize over unknowns
- Bayes factors
  - marginal posteriors pr($y \mid \gamma$)
  - average over unknowns

classical information criteria

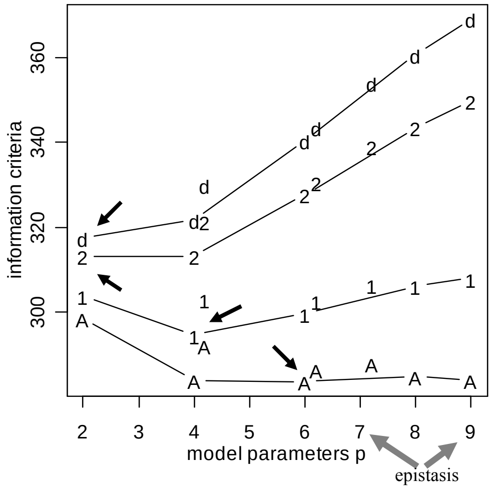
- start with likelihood $L(\gamma \mid y, m)$
  - measures fit of architecture (γ) to phenotype (y)
    - given marker data (m)
  - genetic architecture (γ) depends on parameters
    - have to estimate loci (μ) and effects (λ)
- complexity related to number of parameters
  - $|\gamma|$ = size of genetic architecture
    - BC: $|\gamma| = 1 + n.qtl + n.qtl(n.qtl - 1) = 1 + 4 + 12 = 17$
    - F2: $|\gamma| = 1 + 2n.qtl + 4n.qtl(n.qtl - 1) = 1 + 8 + 48 = 57$

classical information criteria

- construct information criteria
  - balance fit to complexity
  - Akaike AIC $= -2 \log(L) + 2\,|\gamma|$
  - Bayes/Schwartz BIC $= -2 \log(L) + |\gamma| \log(n)$
  - Broman $\text{BIC}_\delta = -2 \log(L) + \delta\,|\gamma| \log(n)$
  - general form: IC $= -2 \log(L) + |\gamma|\, D(n)$
- compare models
  - hypothesis testing: designed for one comparison
    - $2 \log[LR(\gamma_1, \gamma_2)] = L(y|m, \gamma_2) - L(y|m, \gamma_1)$
  - model selection: penalize complexity
    - $\text{IC}(\gamma_1, \gamma_2) = 2 \log[LR(\gamma_1, \gamma_2)] + (|\gamma_2| - |\gamma_1|)\, D(n)$

information criteria vs. model size

- WinQTL 2.0
- SCD data on F2
- A=AIC
- 1=BIC(1)
- 2=BIC(2)
- d=BIC(δ)
- models
  - 1,2,3,4 QTL
    - 2+5+9+2
  - epistasis
    - 2:2 AD



epistasis

Bayes factors

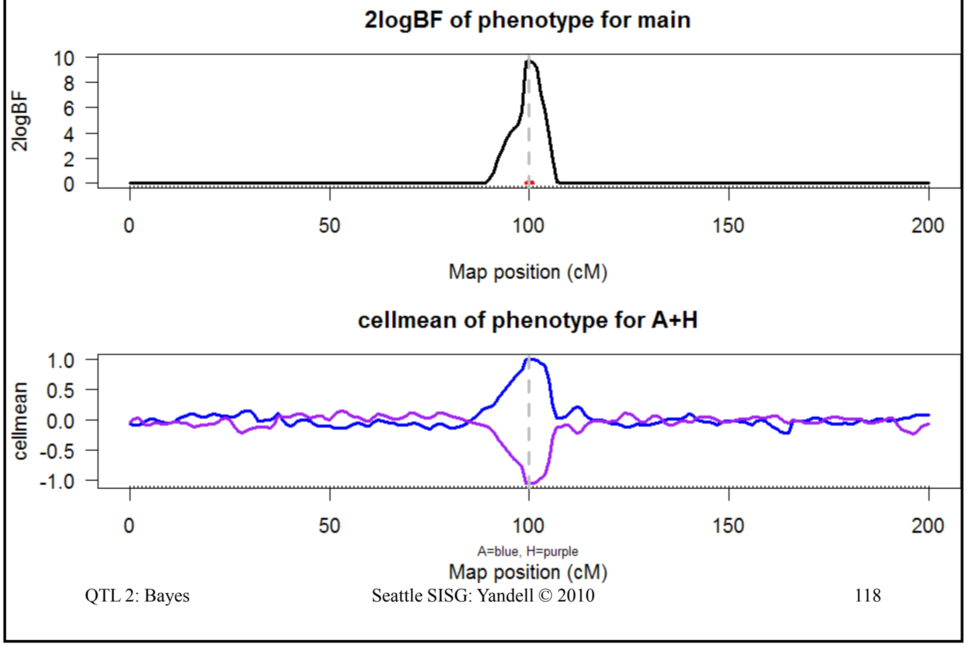
- ratio of model likelihoods
  - ratio of posterior to prior odds for architectures
  - averaged over unknowns

$$B_{12} = \frac{\mathrm{pr}(\gamma_1 \mid y, m) / \mathrm{pr}(\gamma_2 \mid y, m)}{\mathrm{pr}(\gamma_1) / \mathrm{pr}(\gamma_2)} = \frac{\mathrm{pr}(y \mid m, \gamma_1)}{\mathrm{pr}(y \mid m, \gamma_2)}$$

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

$$-2\log(B_{12}) = -2\log(LR) - (|\gamma_2| - |\gamma_1|)\log(n)$$

scan of marginal Bayes factor & effect



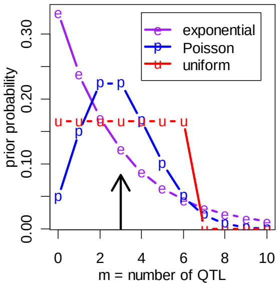
issues in computing Bayes factors

- *BF* insensitive to shape of prior on γ
  - 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(γ | y, m) is marginal histogram

Bayes factors & genetic architecture γ

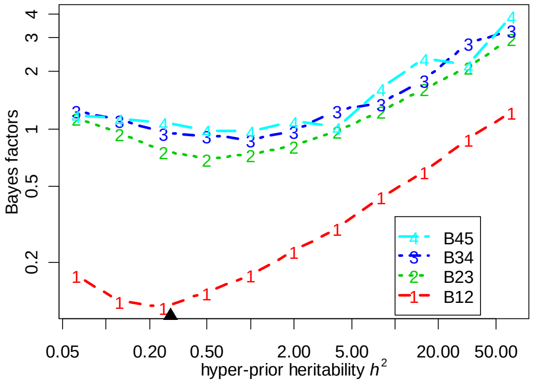
- $|\gamma|$ = number of QTL
  - prior $\text{pr}(\gamma)$ chosen by user
  - posterior $\text{pr}(\gamma|y,m)$
    - sampled marginal histogram
    - shape affected by prior $\text{pr}(A)$

$$BF_{\gamma_1,\gamma_2} = \frac{\text{pr}(\gamma_1|y,m)/\text{pr}(\gamma_1)}{\text{pr}(\gamma_2|y,m)/\text{pr}(\gamma_2)}$$



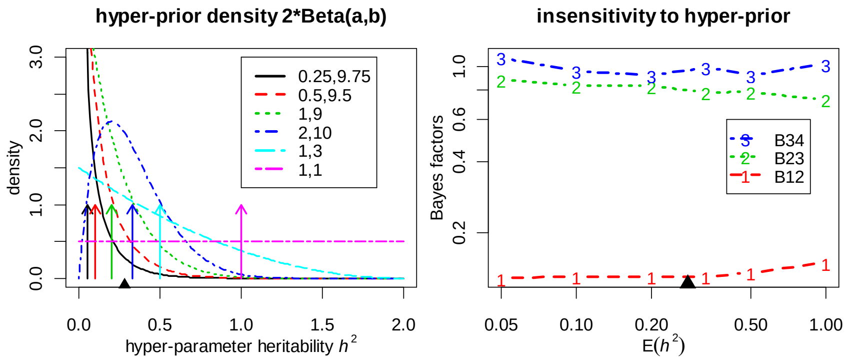
- pattern of QTL across genome
- gene action and epistasis

BF sensitivity to fixed prior for effects



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

BF insensitivity to random effects prior



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