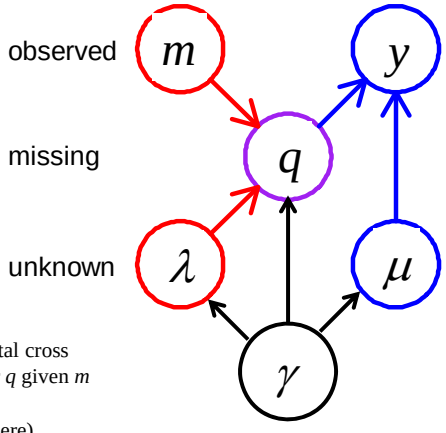
# QTL Model Selection

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



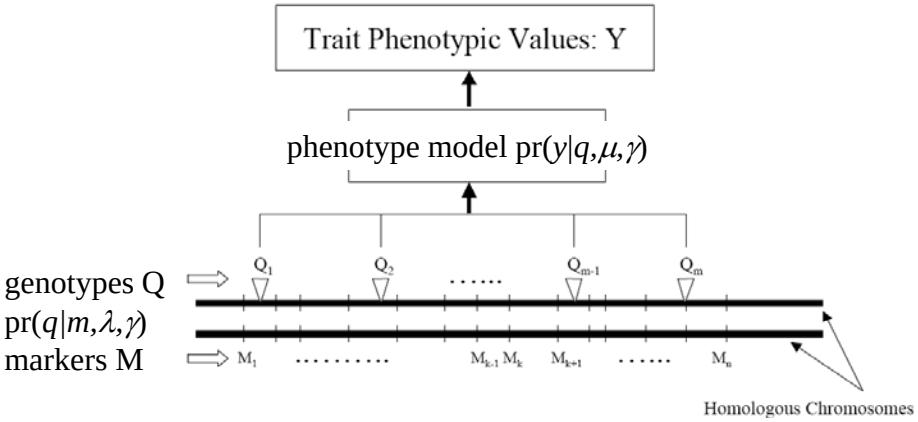
# QTL model selection: key players

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




after
Sen Churchill (2001)

# QTL mapping (from ZB Zeng)






# classical likelihood approach

- genotype model pr($q|m,\lambda,\gamma$)
  - missing genotypes $q$ depend on observed markers $m$ across genome
- phenotype model pr($y|q,\mu,\gamma$)
  - link phenotypes $y$ to genotypes $q$

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

likelihood mixes over missing QTL genotypes:

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

# EM approach

- Iterate E and M steps
  - expectation (E): geno prob's pr($q|m,\lambda,\gamma$)
  - maximization (M): pheno model parameters
    - mean, effects, variance
  - careful attention when many QTL present
    - Multiple papers by Zhao-Bang Zeng and others
  - Start with simple initial model
    - Add QTL, epistatic effects sequentially



# classic model search

- initial model from single QTL analysis
- search for additional QTL
- search for epistasis between pairs of QTL
  - Both in model? One in model? Neither?
- Refine model
  - Update QTL positions
  - Check if existing QTL can be dropped
- Analogous to stepwise regression

# comparing models (details later)

- balance model fit against model complexity
  - want to fit data well (maximum likelihood)
  - without getting too complicated a model

| | **smaller model** | **bigger model** |
|---|---|---|
| **fit model** | miss key features | fits better |
| **estimate phenotype** | may be biased | no bias |
| **predict new data** | may be biased | no bias |
| **interpret model** | easier | more complicated |
| **estimate effects** | low variance | high variance |



# 1. Bayesian strategy for QTL study

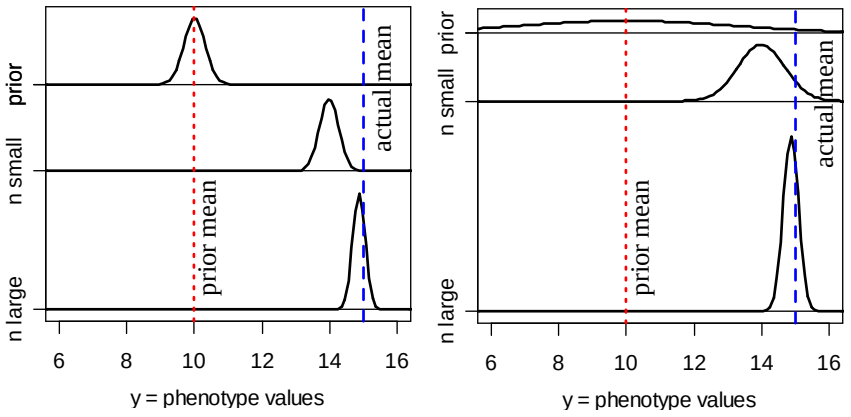
- augment data ($y$,$m$) with missing genotypes $q$
- study unknowns ($\mu$,$\lambda$,$\gamma$) given augmented data ($y$,$m$,$q$)
  - find better genetic architectures $\gamma$
  - find most likely genomic regions = QTL = $\lambda$
  - estimate phenotype parameters = genotype means = $\mu$
- sample from posterior in some clever way
  - multiple imputation (Sen Churchill 2002)
  - Markov chain Monte Carlo (MCMC)
    - (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



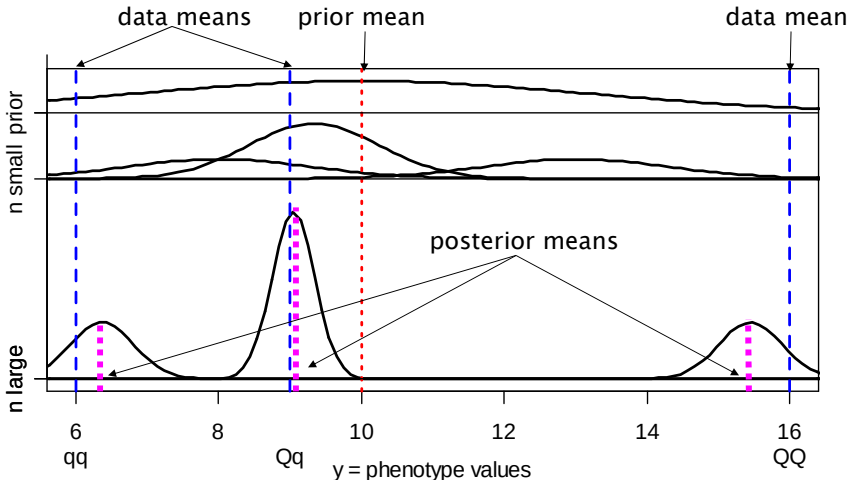

small prior variance

large prior variance



# Posterior on 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: $E(y \mid q) = \mu_q \qquad V(y \mid q) = \sigma^2$

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

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

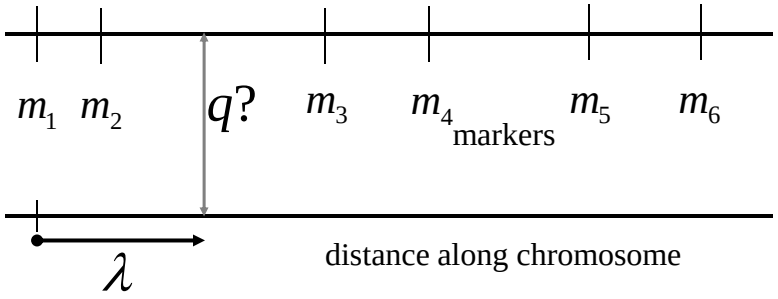
$n_q = \text{count}\{q_i = q\} \qquad \bar{y}_q = \underset{\{q_i = q\}}{\text{sum}} y_i / n_q$

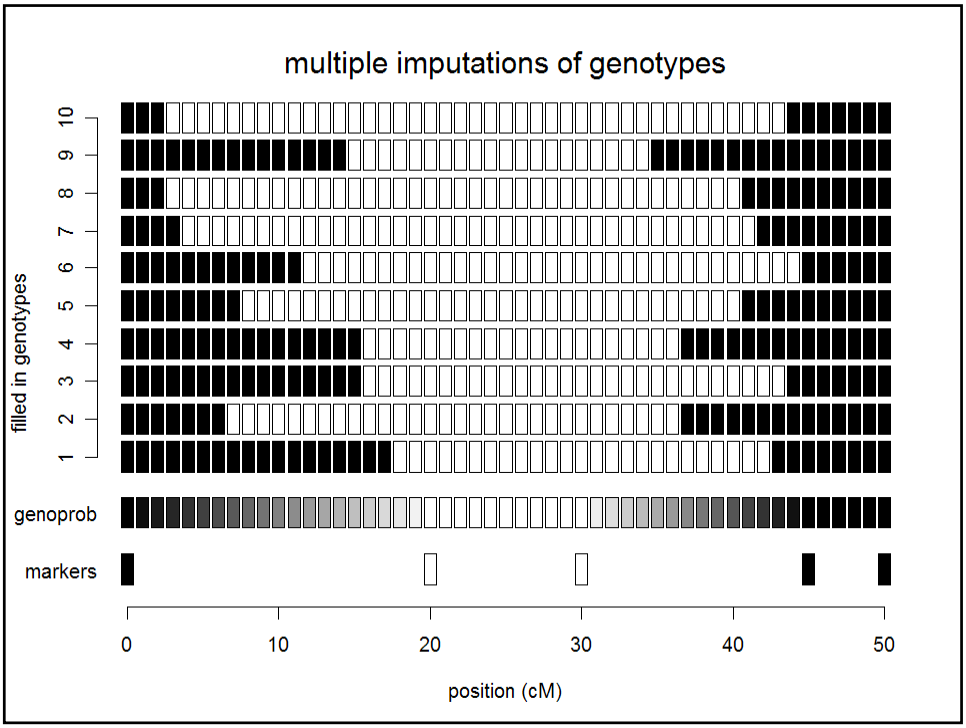
shrinkage: $b_q = \dfrac{\kappa n_q}{\kappa n_q + 1} \to 1$



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

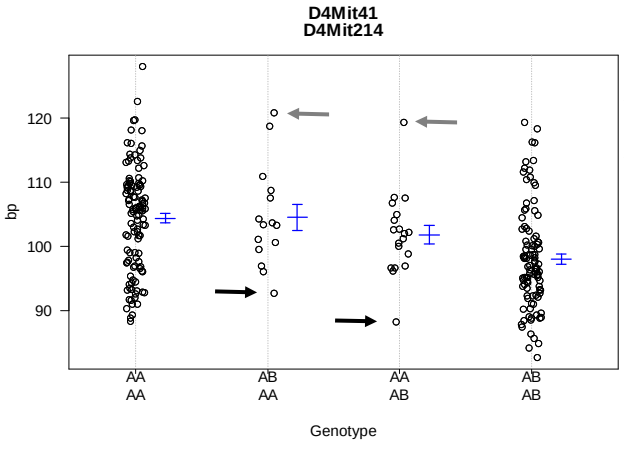
pr($q|m,\lambda$) = pr(geno | map, locus) ≈
pr(geno | flanking markers, locus)






# 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($q$ | $m$, $\lambda$)
    - 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 $\lambda$ 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 $q$

  pr($\lambda$ | $m$,$q$) = pr($\lambda$) pr($q$ | $m$,$\lambda$) / 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 $\gamma$?

- 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



$\gamma$ = genetic architecture:

loci:

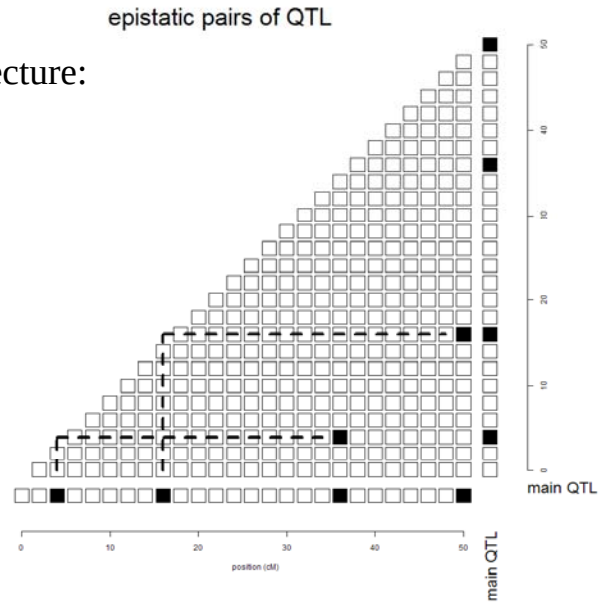
main QTL

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

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

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

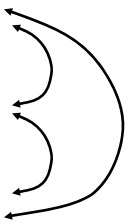
# MCMC sampling of unknowns ($q$,$\mu$,$\lambda$)
## for given genetic architecture $\gamma$

- Gibbs sampler
  - genotypes $q$
  - effects $\mu$
  - *not* loci $\lambda$

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

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

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



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



# Gibbs sampler
## for two genotypic means

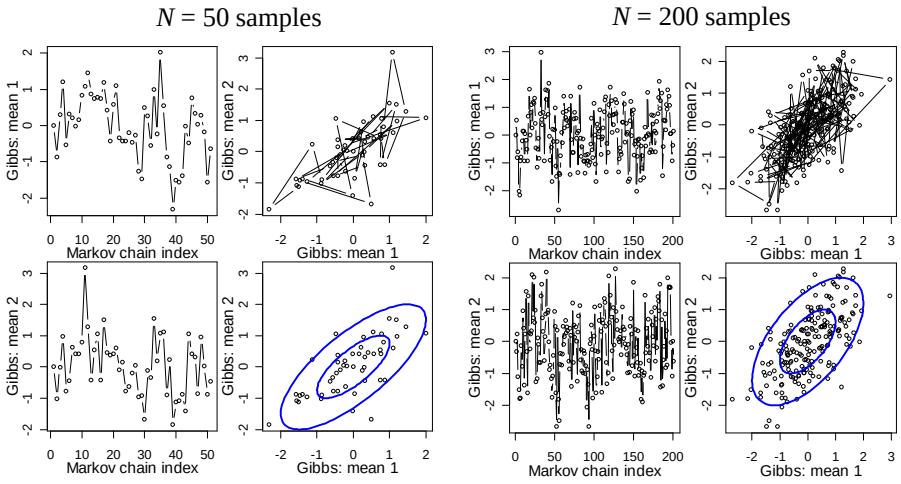
- want to study two correlated effects
  - could sample directly from their bivariate distribution
  - assume correlation $\rho$ is known
- instead use Gibbs sampler:
  - sample each effect from its full conditional given the other
  - pick order of sampling at random
  - repeat many times

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

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

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

# Gibbs sampler samples: $\rho = 0.6$





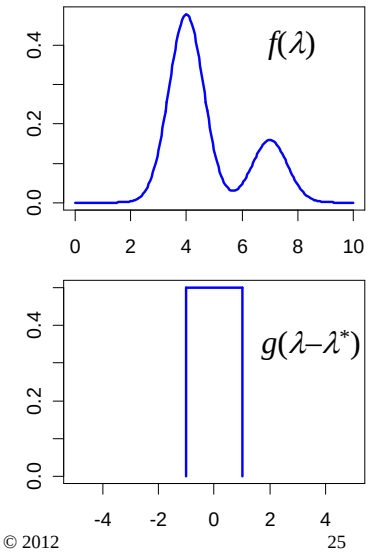

# full conditional for locus

- 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 $\lambda$ 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

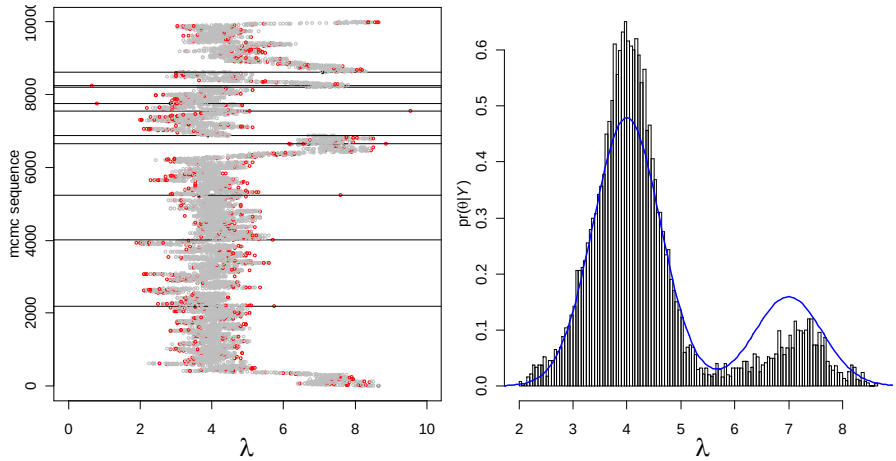
- want to study distribution $f(\lambda)$
  - take Monte Carlo samples
    - unless too complicated
  - take samples using ratios of $f$
- Metropolis-Hastings samples:
  - propose new value $\lambda^*$
    - near (?) current value $\lambda$
    - from some distribution $g$
  - accept new value with prob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

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



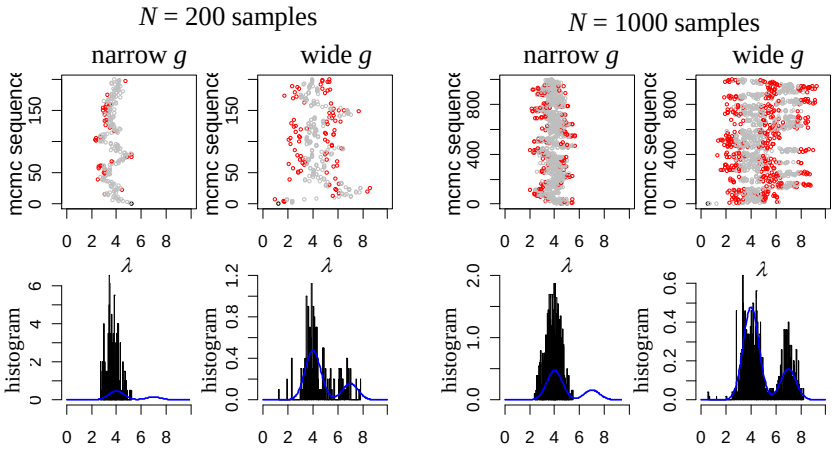



# Metropolis-Hastings for locus $\lambda$




added twist: occasionally propose from entire genome

# Metropolis-Hastings samples






# 3. sampling genetic architectures

- search across genetic architectures $\gamma$ 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

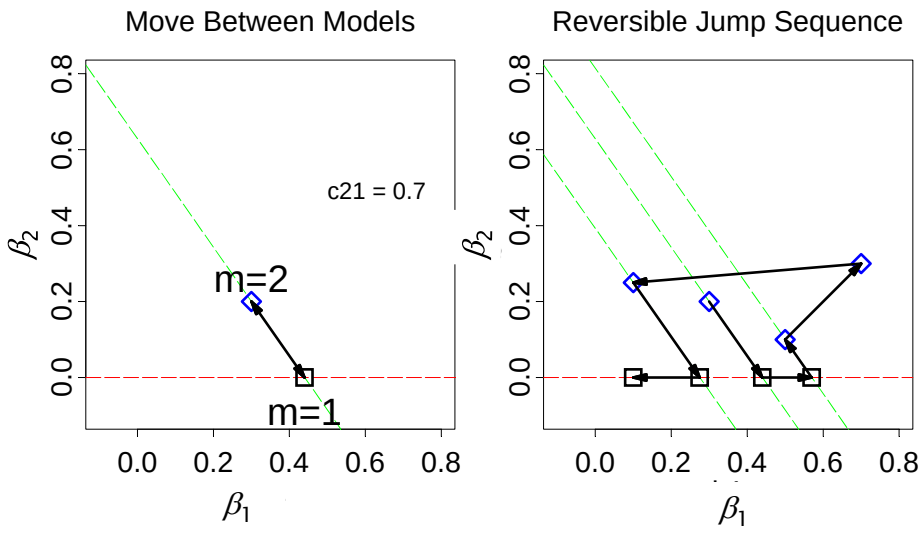
- consider known genotypes $q$ at 2 known loci $\lambda$
  - models with 1 or 2 QTL
- M-H step between 1-QTL and 2-QTL models
  - model changes dimension (via careful bookkeeping)
  - consider mixture over QTL models $H$

$$\gamma = 1\,\text{QTL} : Y = \beta_0 + \beta(q_1) + e$$

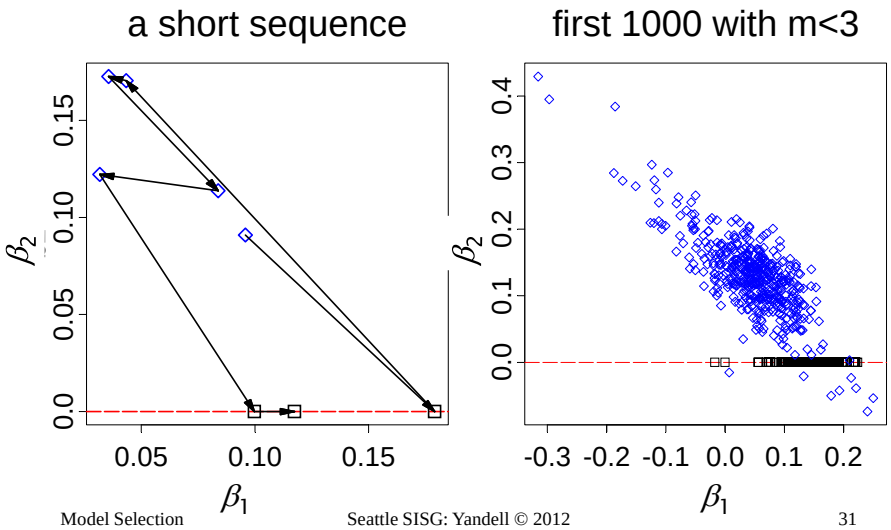
$$\gamma = 2\,\text{QTL} : Y = \beta_0 + \beta_1(q_1) + \beta_2(q_2) + e$$



# geometry of reversible jump

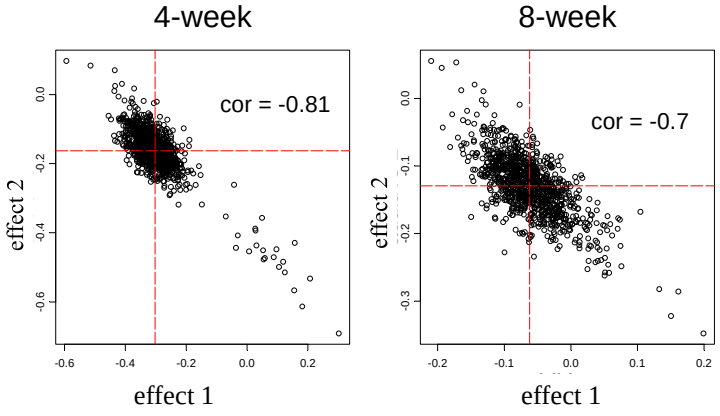


# geometry allowing $q$ and $\lambda$ to change



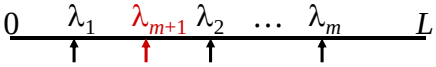



# collinear QTL = correlated effects




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

# sampling across QTL models $\gamma$




action steps: draw one of three choices

- update QTL model $\gamma$ 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 $\gamma$
  - relatively easy to incorporate epistasis
  - Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005 *Genetics*)
    - (see earlier work of Nengjun Yi and Ina Hoeschele)

$$\mu_q = \mu + \gamma_1 \beta_1(q_1) + \gamma_2 \beta_2(q_2), \quad \gamma_k = 0,1$$

# Bayesian shrinkage estimation

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

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



# 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($\gamma$)
  - 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 ($\gamma$) to phenotype ($y$)
    - given marker data ($m$)
  - genetic architecture ($\gamma$) depends on parameters
    - have to estimate loci ($\mu$) and effects ($\lambda$)
- 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: $|\gamma| = 1 + 2n.qtl + 4n.qtl(n.qtl - 1) = 1 + 8 + 48 = 57$

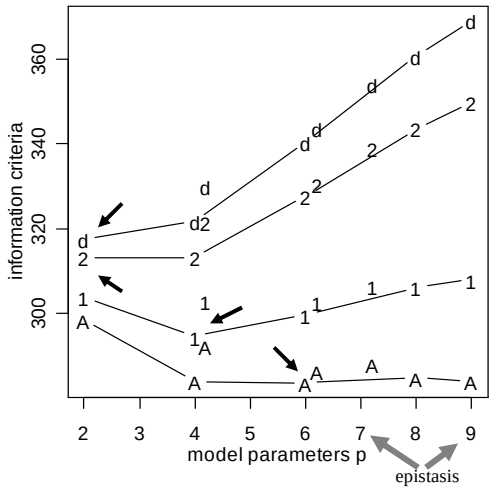
# classical information criteria

- 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($\delta$)
- models
  - 1,2,3,4 QTL
    - 2+5+9+2
  - epistasis
    - 2:2 AD



# Bayes factors

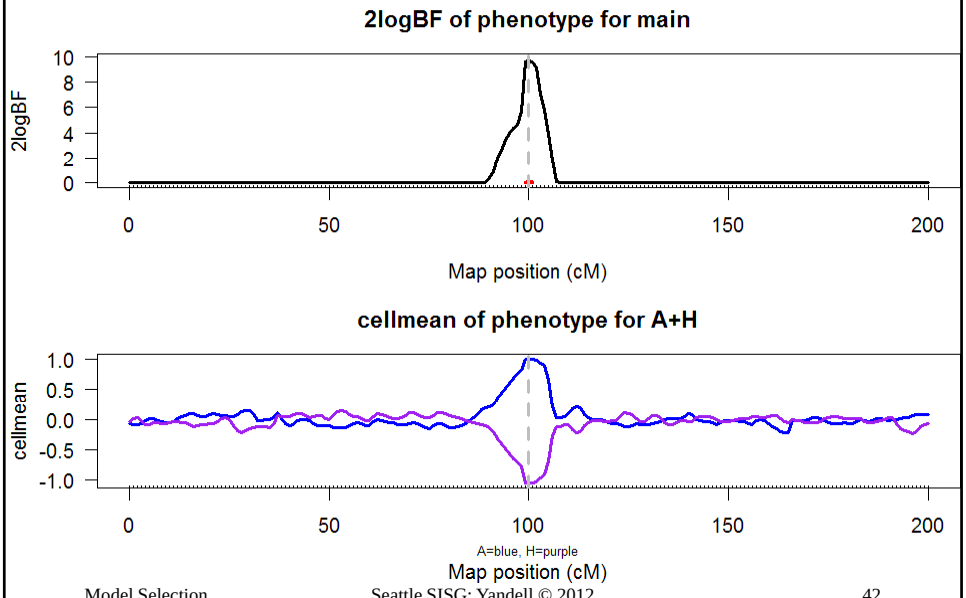
- ratio of model likelihoods
  - ratio of posterior to prior odds for architectures
  - averaged over unknowns

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

- 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

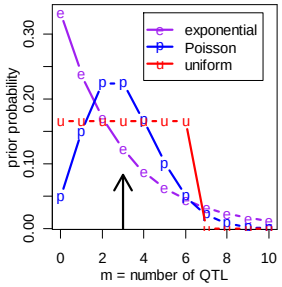
- *BF* insensitive to shape of prior on $\gamma$
  - geometric, Poisson, uniform
  - precision improves when prior mimics posterior
- *BF* sensitivity to prior variance on effects $\theta$
  - prior variance should reflect data variability
  - resolved by using hyper-priors
    - automatic algorithm; no need for user tuning
- easy to compute Bayes factors from samples
  - sample posterior using MCMC
  - posterior pr($\gamma \mid y, m$) is marginal histogram



# Bayes factors & genetic architecture $\gamma$

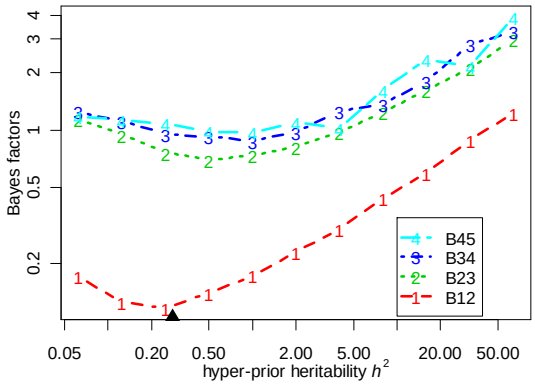
- $|\gamma|$ = number of QTL
  - prior pr($\gamma$) chosen by user
  - posterior pr($\gamma \mid y,m$)
    - sampled marginal histogram
    - shape affected by prior pr(*A*)

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




- pattern of QTL across genome
- gene action and epistasis

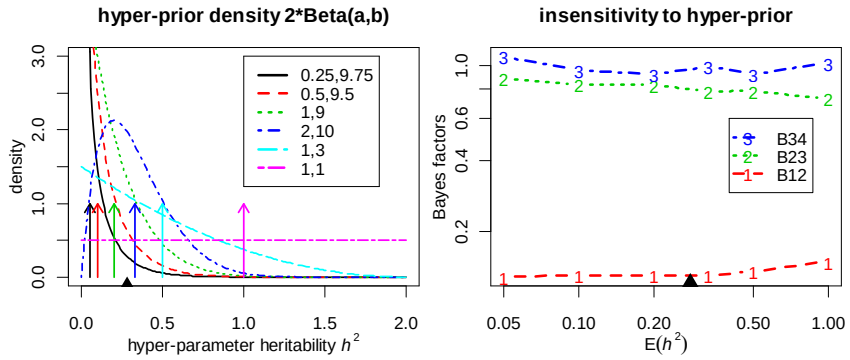
# BF sensitivity to fixed prior for effects




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



# BF insensitivity to random effects prior




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