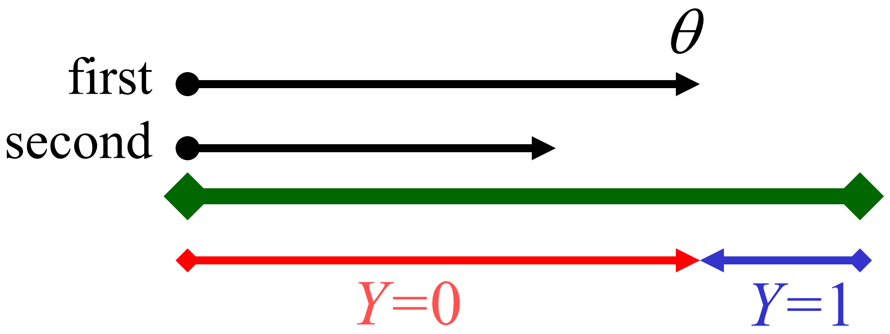
Bayesian Interval Mapping

1. Who was Bayes? 2-6
    - What is Bayes theorem?
2. Bayesian inference for QTL 7-14
3. Markov chain sampling 15-29
    - for fixed number of QTL *m*
4. Sampling across architectures 30-40
    - handling epistasis

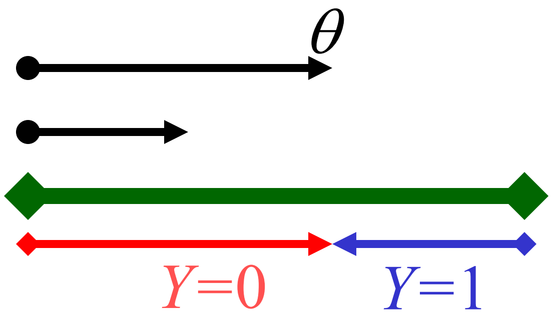
1. who was Bayes? what is Bayes theorem?

- 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 (right)?



prior	$\mathrm{pr}(\theta)$	$= 1$
likelihood	$\mathrm{pr}(Y \mid \theta)$	$= \theta^{1-Y}(1-\theta)^{Y}$
posterior	$\mathrm{pr}(\theta \mid Y)$	$= ?$

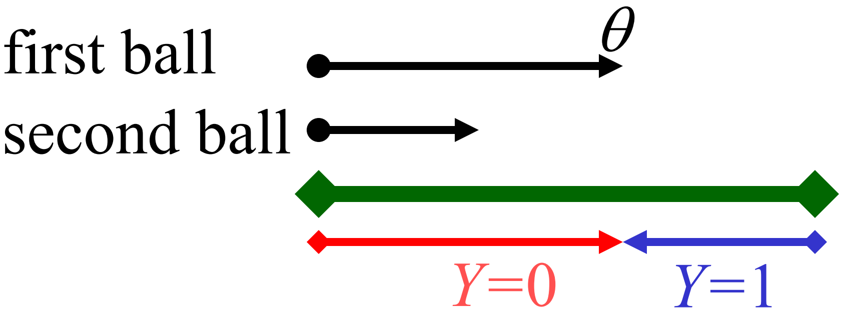
what is Bayes theorem?



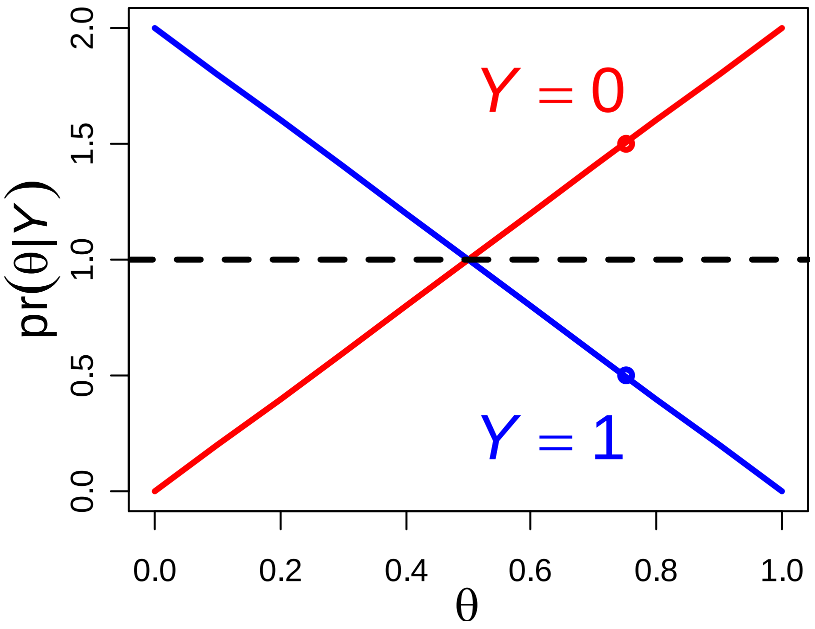
- where is first ball if the second is to its left (right)?
- prior: probability of parameter before observing data
  - pr(θ) = pr(parameter)
  - equal chance of being anywhere on the table
- posterior: probability of parameter after observing data
  - pr(θ | Y) = pr(parameter | data)
  - more likely to left if first ball is toward the right end of table
- likelihood: probability of data given parameters
  - pr(Y | θ) = pr(data | parameter)
  - basis for classical statistical inference
- Bayes theorem
  - posterior = likelihood * prior / pr(data)
  - normalizing constant pr(Y) often drops out of calculation

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

where is the second ball given the first?



prior	$\mathrm{pr}(\theta)$	$= 1$
likelihood	$\mathrm{pr}(Y\mid\theta) =$	$\theta^{1-Y}(1-\theta)^{Y}$
posterior	$\mathrm{pr}(\theta\mid Y) =$	?



$$\text{prior}: \mathrm{pr}(\theta) = 1$$

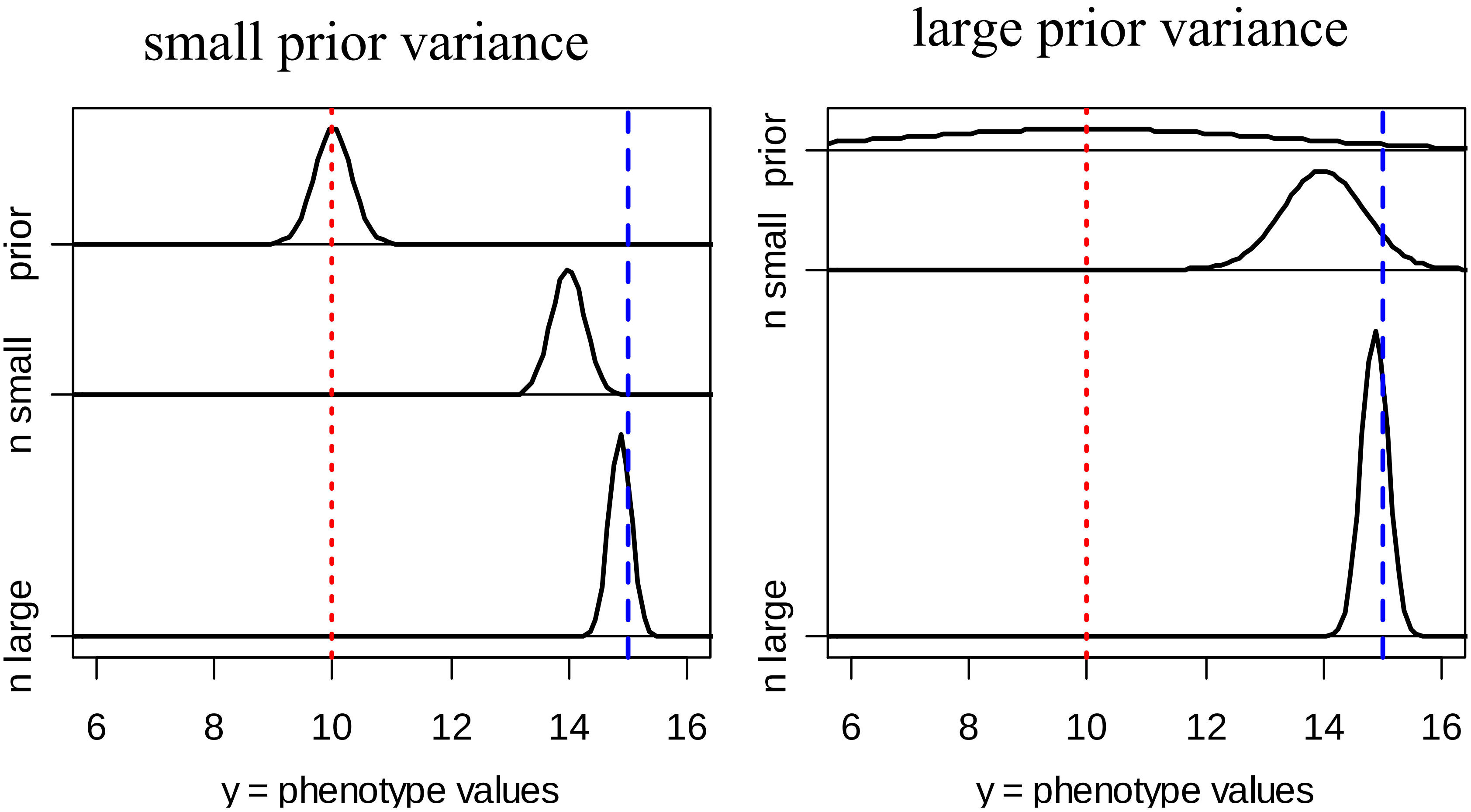
$$\text{likelihood}: \mathrm{pr}(Y \mid \theta) = \begin{cases} \theta & \text{if } Y = 0 \\ 1-\theta & \text{if } Y = 1 \end{cases}$$

$$\text{marginal}: \mathrm{pr}(Y) = \frac{1}{2}$$

$$\text{posterior}: \mathrm{pr}(\theta \mid Y) = \frac{\mathrm{pr}(Y \mid \theta)\mathrm{pr}(\theta)}{\mathrm{pr}(Y)}$$

$$= \begin{cases} 2\theta & \text{if } Y = 0 \\ 2(1-\theta) & \text{if } Y = 1 \end{cases}$$

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	$\mu \sim N\left(B_n\overline{Y}_{\bullet} + (1 - B_n)\mu_0, B_n\dfrac{\sigma^2}{n}\right)$
	with $\overline{Y}_{\bullet} = \text{sum}\dfrac{Y_i}{n}$
fudge factor (shrinks to 1)	$B_n = \dfrac{\kappa n}{\kappa n + 1} \to 1$

2. Bayesian inference for QTL

- develop priors on unknowns
  - unknowns:
    - missing genotypes Q
    - effects $\theta = (G_Q, \sigma^2)$
    - loci λ (see next section)
  - use empirical Bayes to set useful priors
- study posterior for unknowns given data
  - data:
    - phenotypes Y
    - markers & linkage map X
  - marginal posteriors for effects θ, loci λ

Bayesian priors for QTL

- missing genotypes Q
  - pr(Q | X, λ)
  - recombination model is formally a prior
- effects $\theta = (G_Q, \sigma^2)$
  - 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
- each locus λ may be uniform over genome
  - pr(λ | X) = 1 / length of genome
- combined prior
  - pr(Q, θ, λ | X) = pr(Q | X, λ) pr(θ) pr(λ | X)

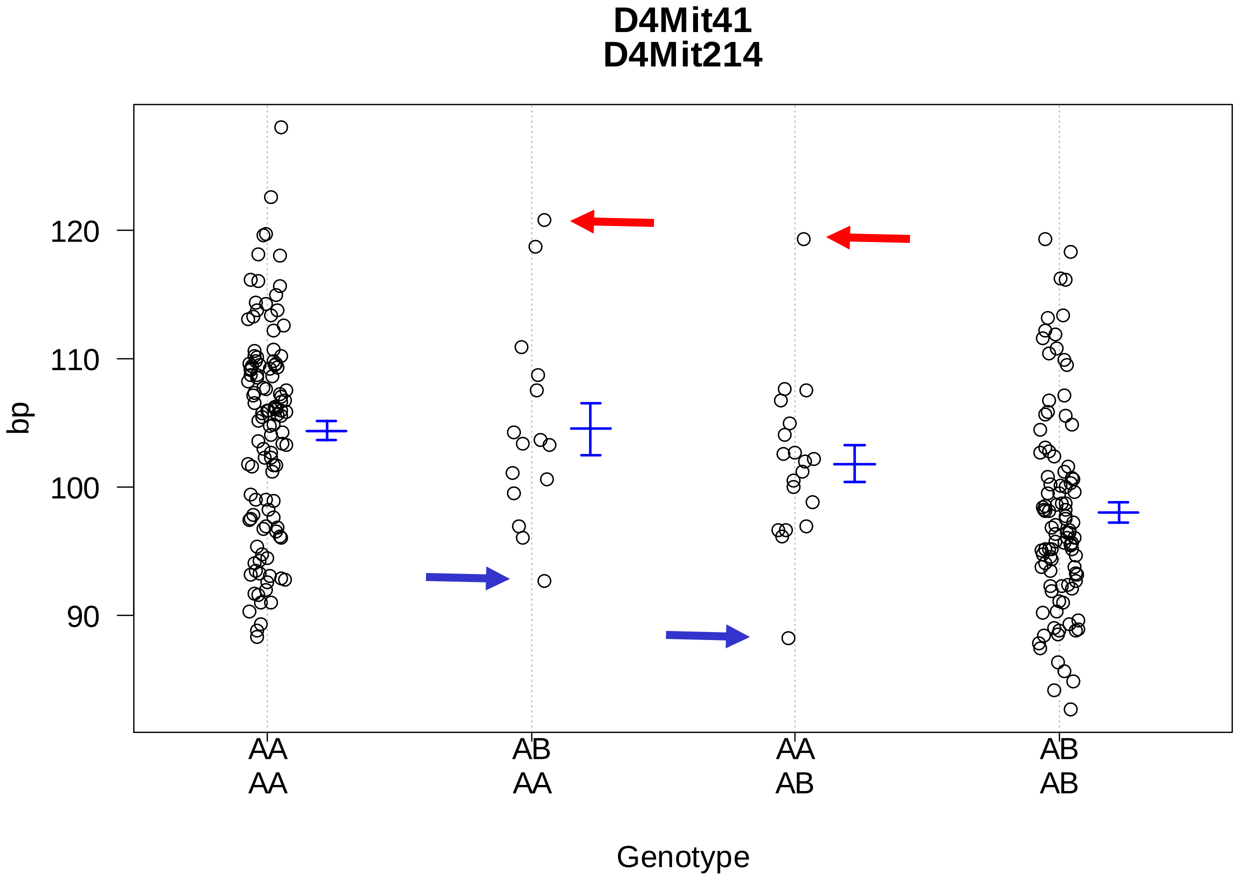
Bayesian model posterior

- augment data (Y,X) with unknowns Q
- study unknowns (θ,λ,Q) given data (Y,X)
  - – properties of posterior $\text{pr}(\theta,\lambda,Q \mid Y, X)$
- sample from posterior in some clever way
  - – multiple imputation or MCMC

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

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

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



what are probabilities for genotype *Q* between markers?

recombinants AA:AB

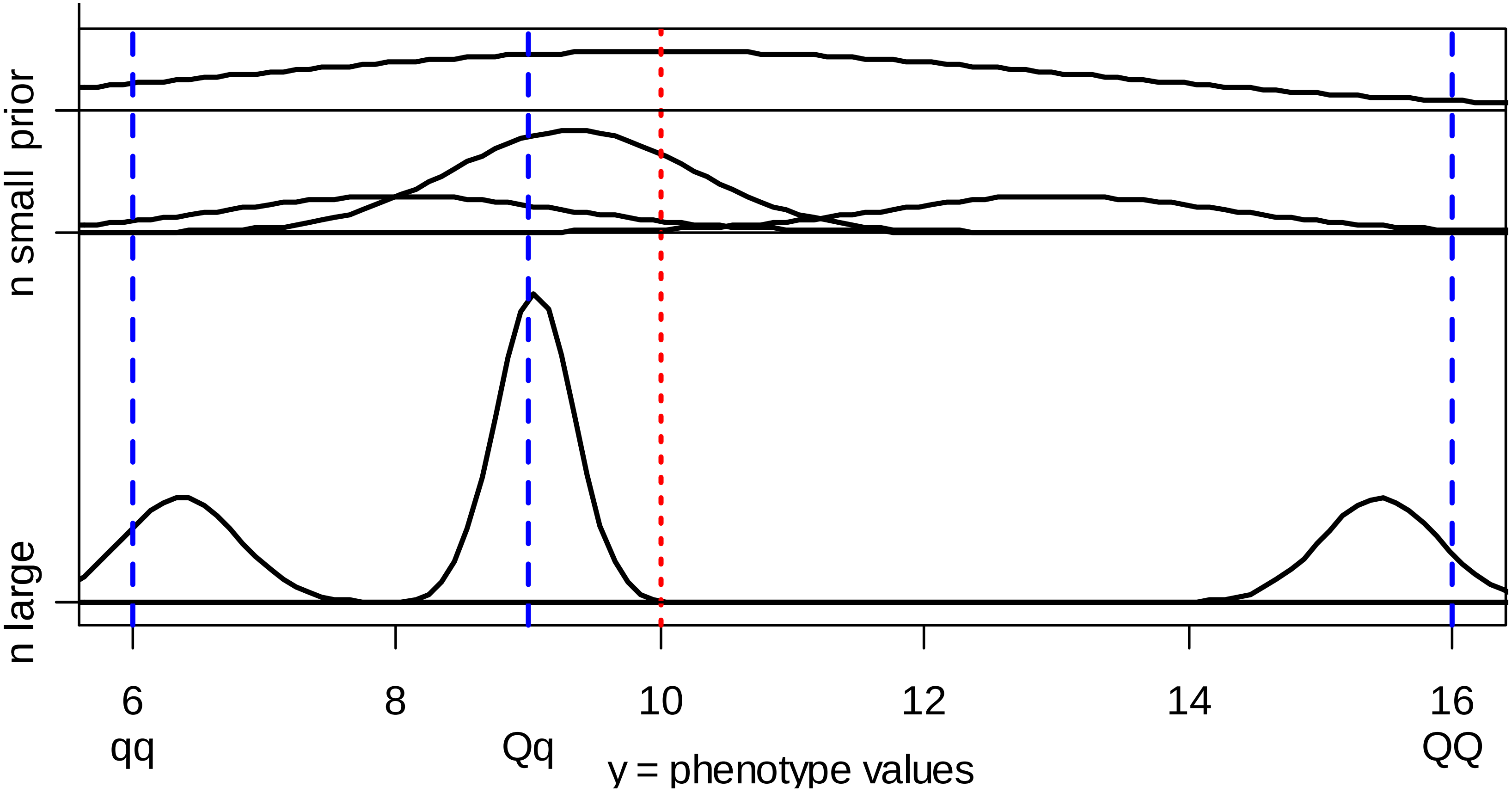
all 1:1 if ignore *Y* and if we use *Y*?

posterior on QTL genotypes

- full conditional of Q given data, parameters
  - proportional to prior pr($Q \mid X_i, \lambda$)
    - weight toward Q that agrees with flanking markers
  - proportional to likelihood 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

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

posterior genotypic means G_Q



genetic effect posterior given Q

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

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
  - $\tau^2 = s^2/3$, $\nu=6$
  - $E(\sigma^2 \mid \nu,\tau^2) = s^2/2$, $\mathrm{Var}(\sigma^2 \mid \nu,\tau^2) = s^4/4$

3. Markov chain sampling of architectures

- 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
- hard to sample (λ, Q, θ, m) from joint posterior
  - update (λ, Q, θ) from full conditionals for m-QTL model
  - update m using reversible jump technology

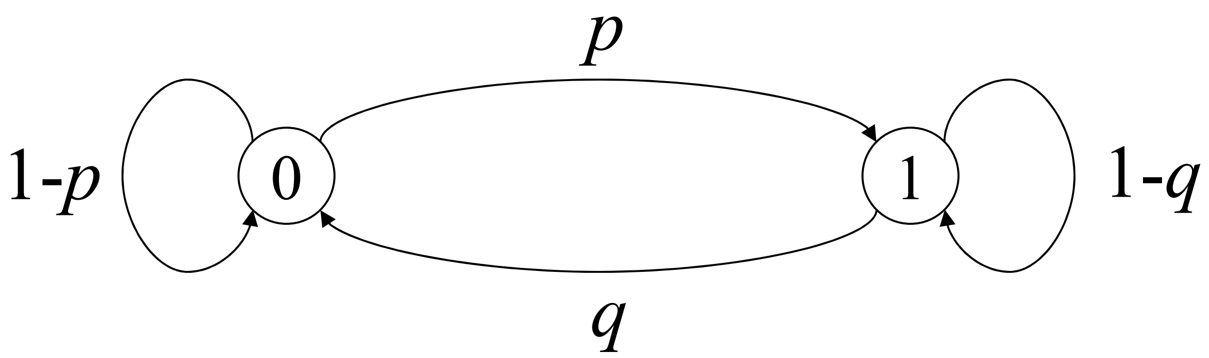
$$(\lambda,Q,\theta,m) \sim \mathrm{pr}(\lambda,Q,\theta,m \mid Y,X)$$

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

What is a Markov chain?

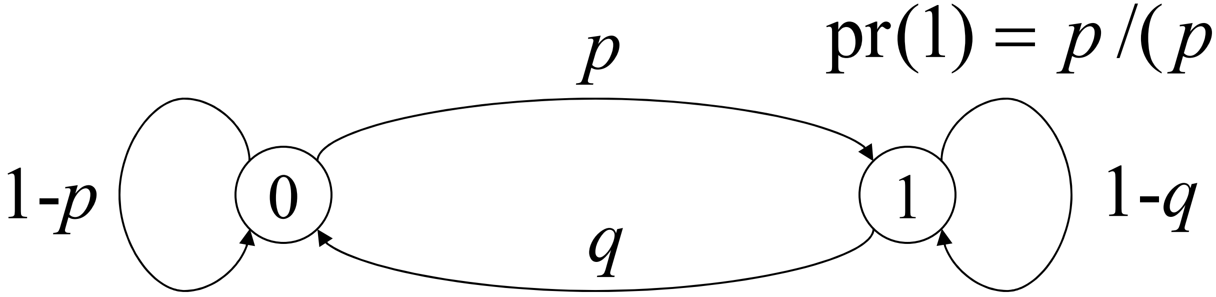
- future given present is independent of past
- update chain based on current value
  - can make chain arbitrarily complicated
  - chain converges to stable pattern $\pi()$ we wish to study
- toy problem
  - two states (0,1)
  - move chances depend on current state
  - what is the chance of being in state 1?

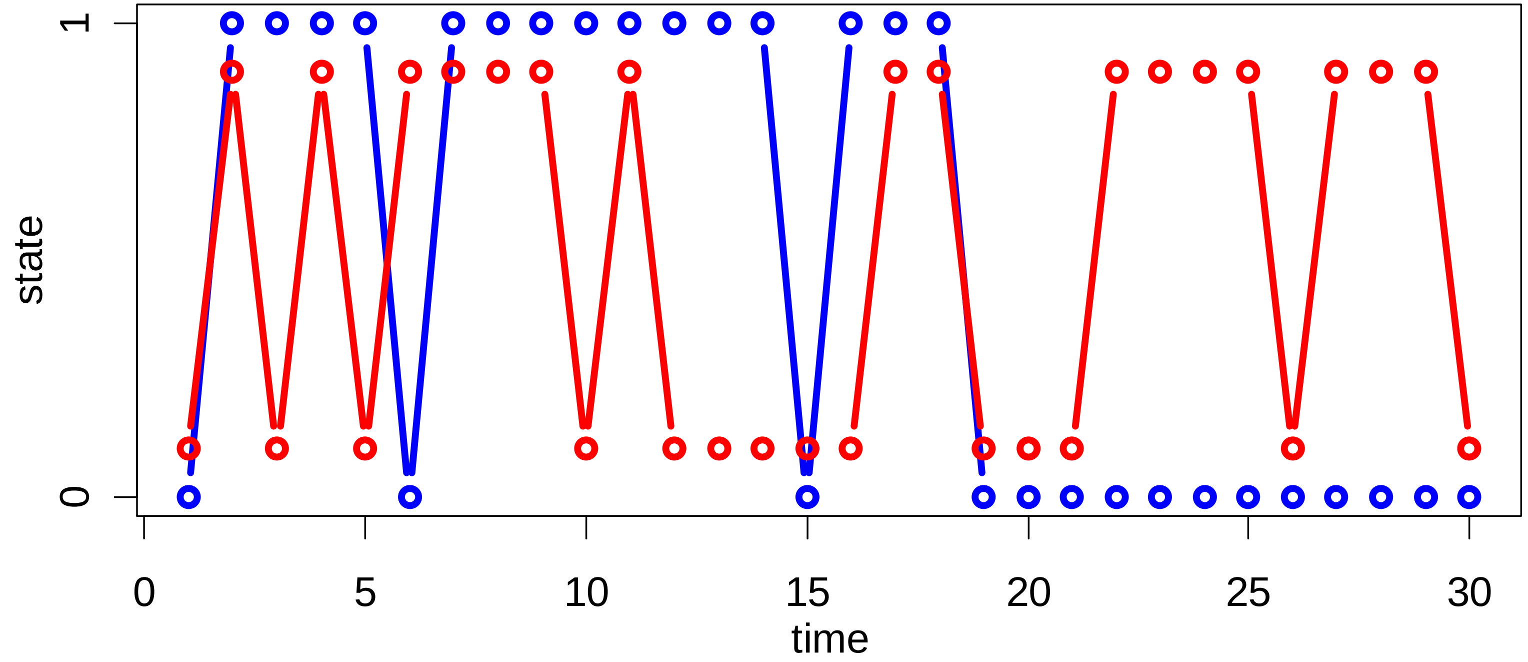
$$\mathrm{pr}(1) = p/(p+q)$$



Markov chain idea

$$\mathrm{pr}(1) = p/(p+q)$$





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

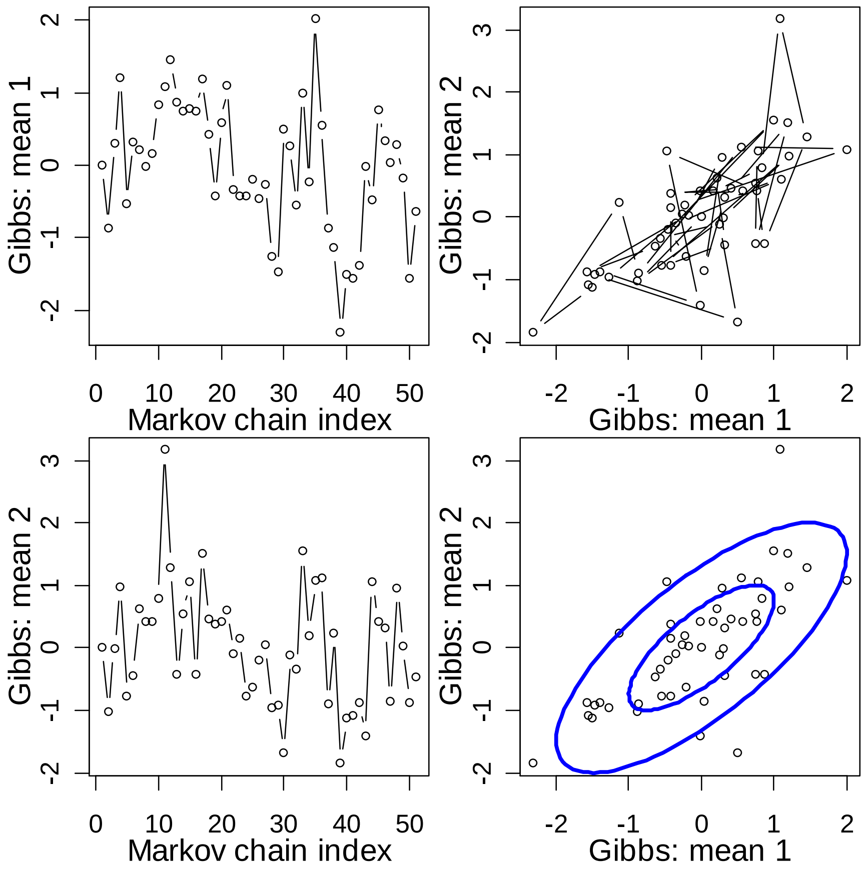
$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

$$\theta_1 \sim N\left(\mu_1 + \rho(\theta_2 - \mu_2), 1 - \rho^2\right)$$

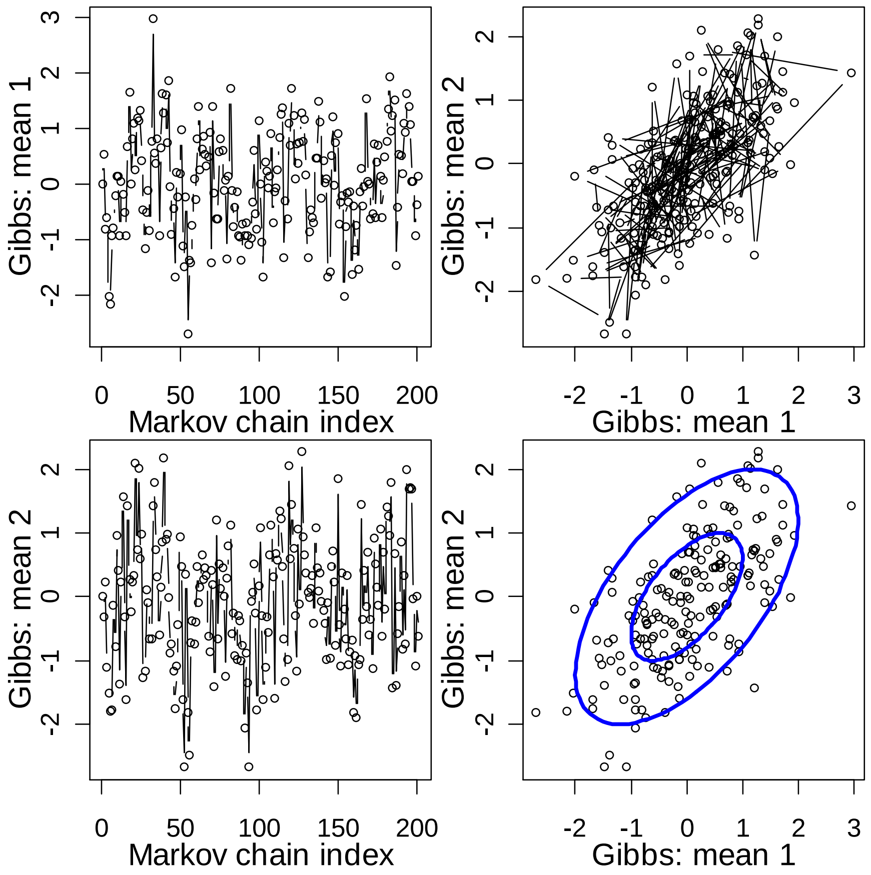
$$\theta_2 \sim N\left(\mu_2 + \rho(\theta_1 - \mu_1), 1 - \rho^2\right)$$

Gibbs sampler samples: $\rho = 0.6$

$N = 50$ samples

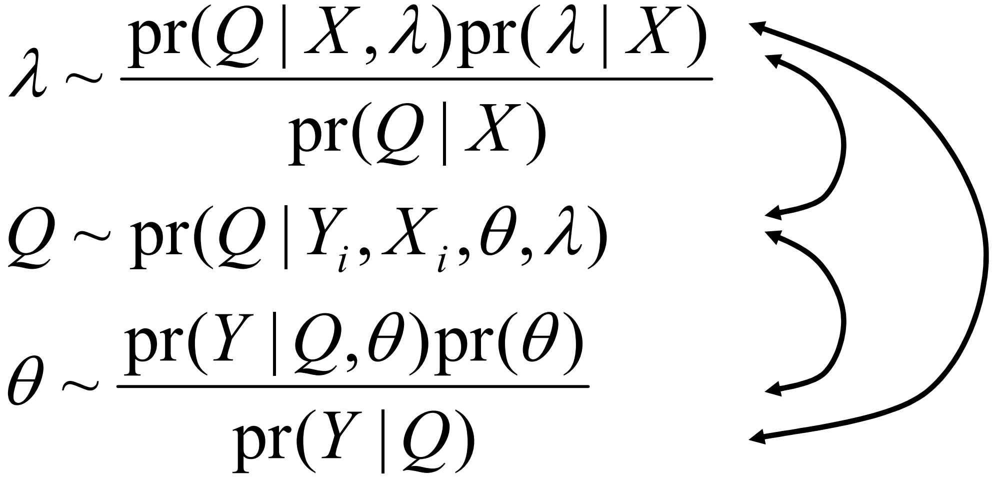


$N = 200$ samples



MCMC sampling of (λ, Q, θ)

- Gibbs sampler
  - effects $\theta = (\ G_Q,\ \sigma^2\ )$
  - genotypes Q
  - *not* loci λ



$$\lambda \sim \frac{\mathrm{pr}(Q \mid X, \lambda)\mathrm{pr}(\lambda \mid X)}{\mathrm{pr}(Q \mid X)}$$

$$Q \sim \mathrm{pr}(Q \mid Y_i, X_i, \theta, \lambda)$$

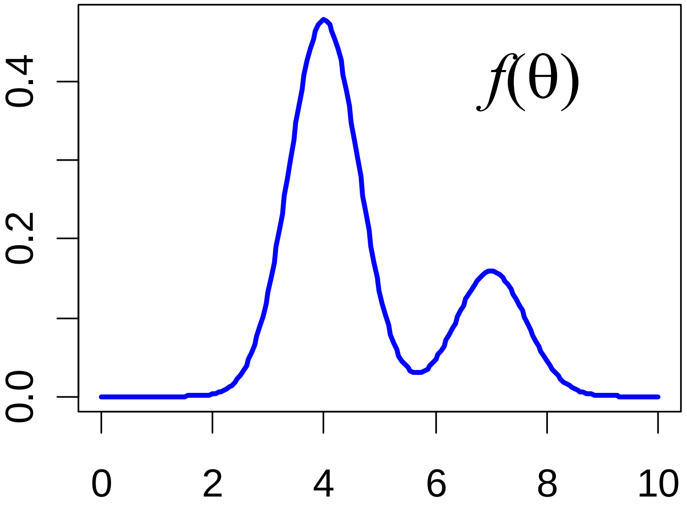
$$\theta \sim \frac{\mathrm{pr}(Y \mid Q, \theta)\mathrm{pr}(\theta)}{\mathrm{pr}(Y \mid Q)}$$

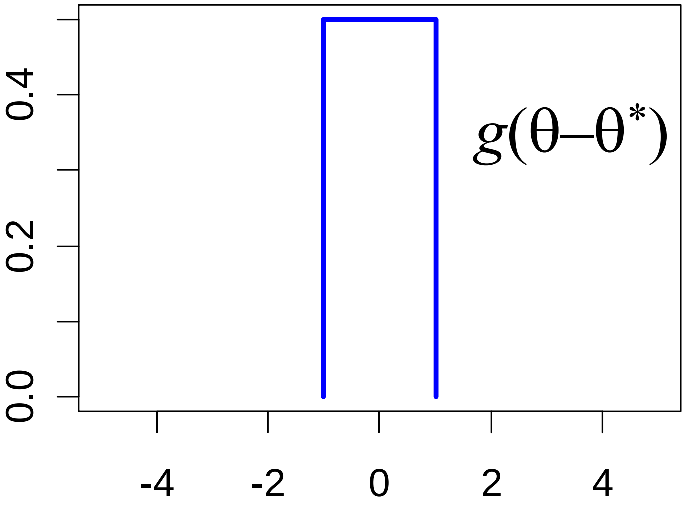
- extension of Gibbs sampler
  - Metropolis-Hastings sampler
  - does not require normalization
  - loci λ: pr($Q \mid X$) difficult to compute

Metropolis-Hastings idea

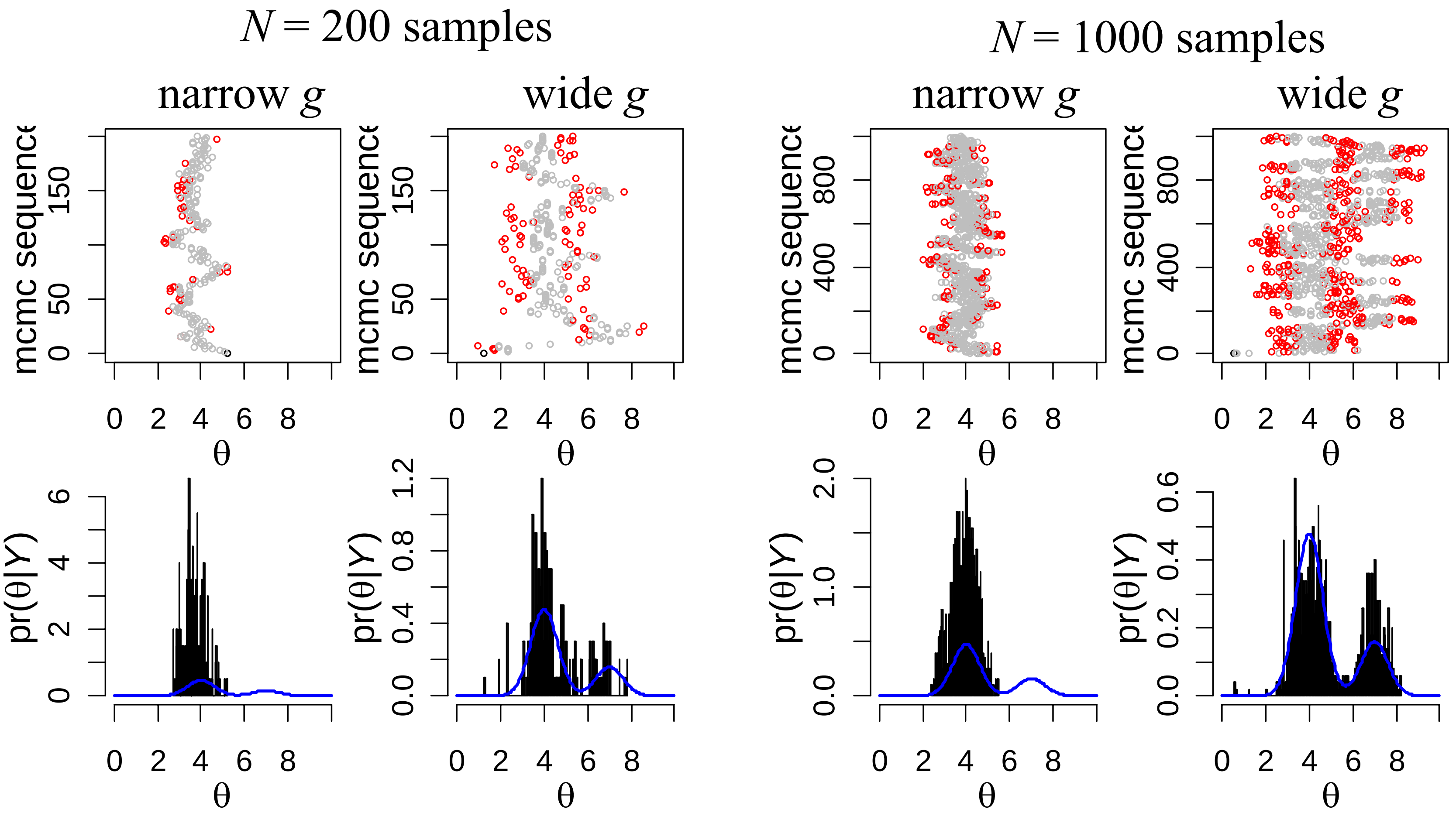
- want to study distribution $f(\theta)$
  - take Monte Carlo samples
    - unless too complicated
  - take samples using ratios of f
- Metropolis-Hastings samples:
  - current sample value θ
  - propose new value θ^*
    - from some distribution $g(\theta,\theta^*)$
    - Gibbs sampler: $g(\theta,\theta^*)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

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





Metropolis-Hastings samples



full conditional for locus

- cannot easily sample from locus full conditional

$$\text{pr}(\lambda \mid Y, X, \theta, Q) = \text{pr}(\lambda \mid X, Q)$$
$$= \text{pr}(\lambda)\, \text{pr}(Q \mid X, \lambda) / \text{constant}$$

- 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 Step

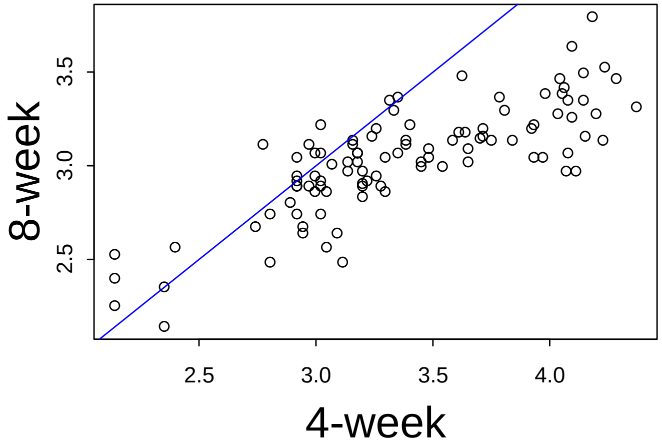
- pick new locus based upon current locus
  - propose new locus from some distribution *g*( )
    - pick value near current one? (usually)
    - pick uniformly across genome? (sometimes)
  - accept new locus with probability *A*
    - otherwise stick with current value

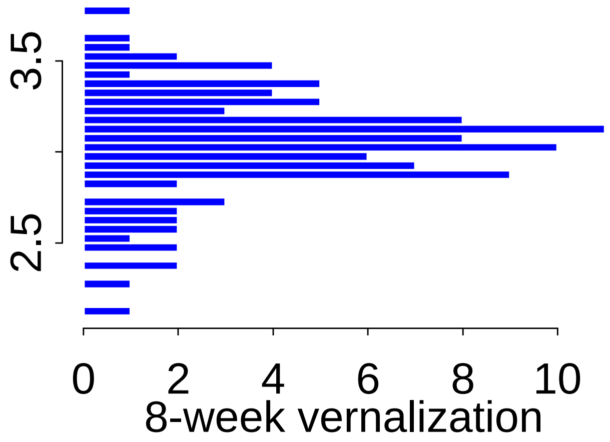
$$A(\lambda_{old},\lambda_{new}) = \min\left(1, \frac{\mathrm{pr}(\lambda_{new})\mathrm{pr}(Q \mid X,\lambda_{new})g(\lambda_{new},\lambda_{old})}{\mathrm{pr}(\lambda_{old})\mathrm{pr}(Q \mid X,\lambda_{old})g(\lambda_{old},\lambda_{new})}\right)$$

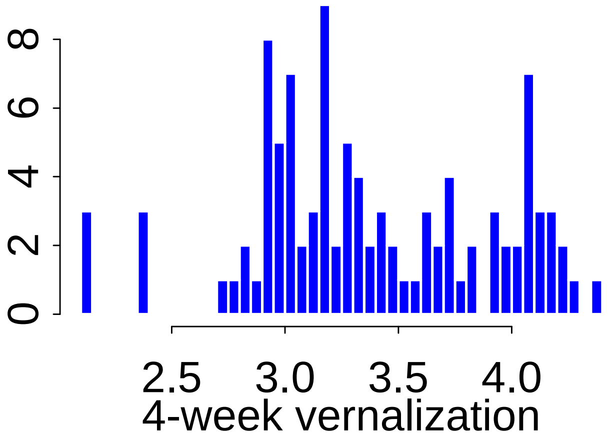
Brassica napus data

- 4-week & 8-week vernalization effect
  - log(days to flower)
- genetic cross of
  - Stellar (annual canola)
  - Major (biennial rapeseed)
- 105 F1-derived double haploid (DH) lines
  - homozygous at every locus (*QQ* or *qq*)
- 10 molecular markers (RFLPs) on LG9
  - two QTLs inferred on LG9 (now chromosome N2)
  - corroborated by Butruille (1998)
  - exploiting synteny with *Arabidopsis thaliana*

Brassica 4- & 8-week data

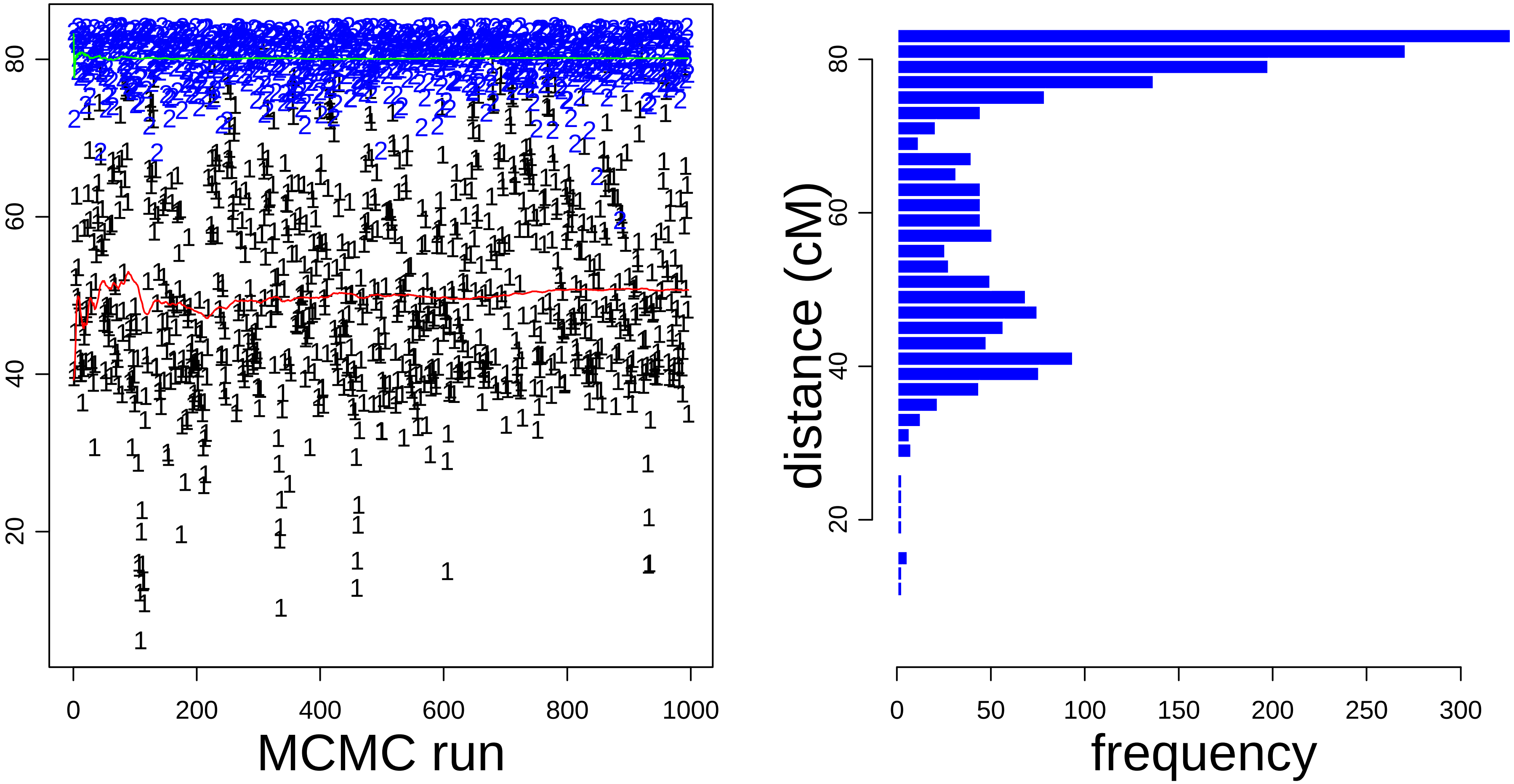






summaries of raw data
joint scatter plots
(identity line)
separate histograms

Brassica 8-week data locus MCMC with *m*=2



4-week vs 8-week vernalization

4-week vernalization

- longer time to flower
- larger LOD at 40cM
- modest LOD at 80cM
- loci well determined

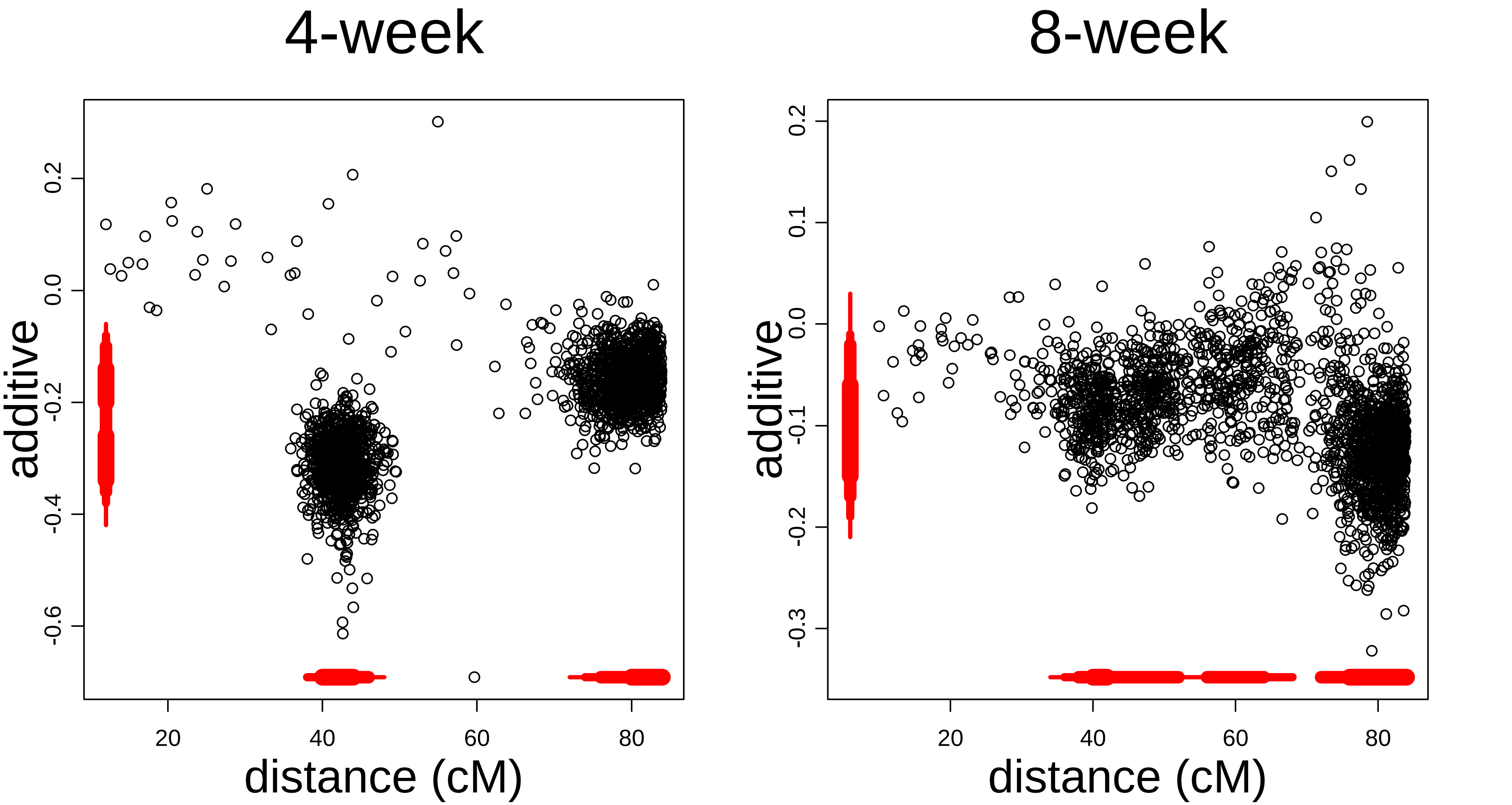
cM	add
40	.30
80	.16

8-week vernalization

- shorter time to flower
- larger LOD at 80cM
- modest LOD at 40cM
- loci poorly determined

cM	add
40	.06
80	.13

Brassica credible regions



4. sampling across architectures

- search across genetic architectures *M* of various sizes
  - allow change in *m* = number of QTL
  - allow change in types of epistatic interactions
- compare architectures
  - Bayes factors: previous talk
- 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 issues

- use reversible jump MCMC to change *m*
  - adjust to change of variables between models
    - bookkeeping helps in comparing models
  - Green (1995); Richardson Green (1997)
- think model selection in multiple regression
  - but regressors (QTL genotypes) are unknown
  - linked loci = collinear regressors = correlated effects
  - consider only additive genetic effects here
    - genotype coding *Q*= –1, 0, 1 centered on average genotype

$$G(Q) = \mu + \beta(Q) \text{ with } \beta(Q) = \alpha \times (Q - \overline{Q})$$

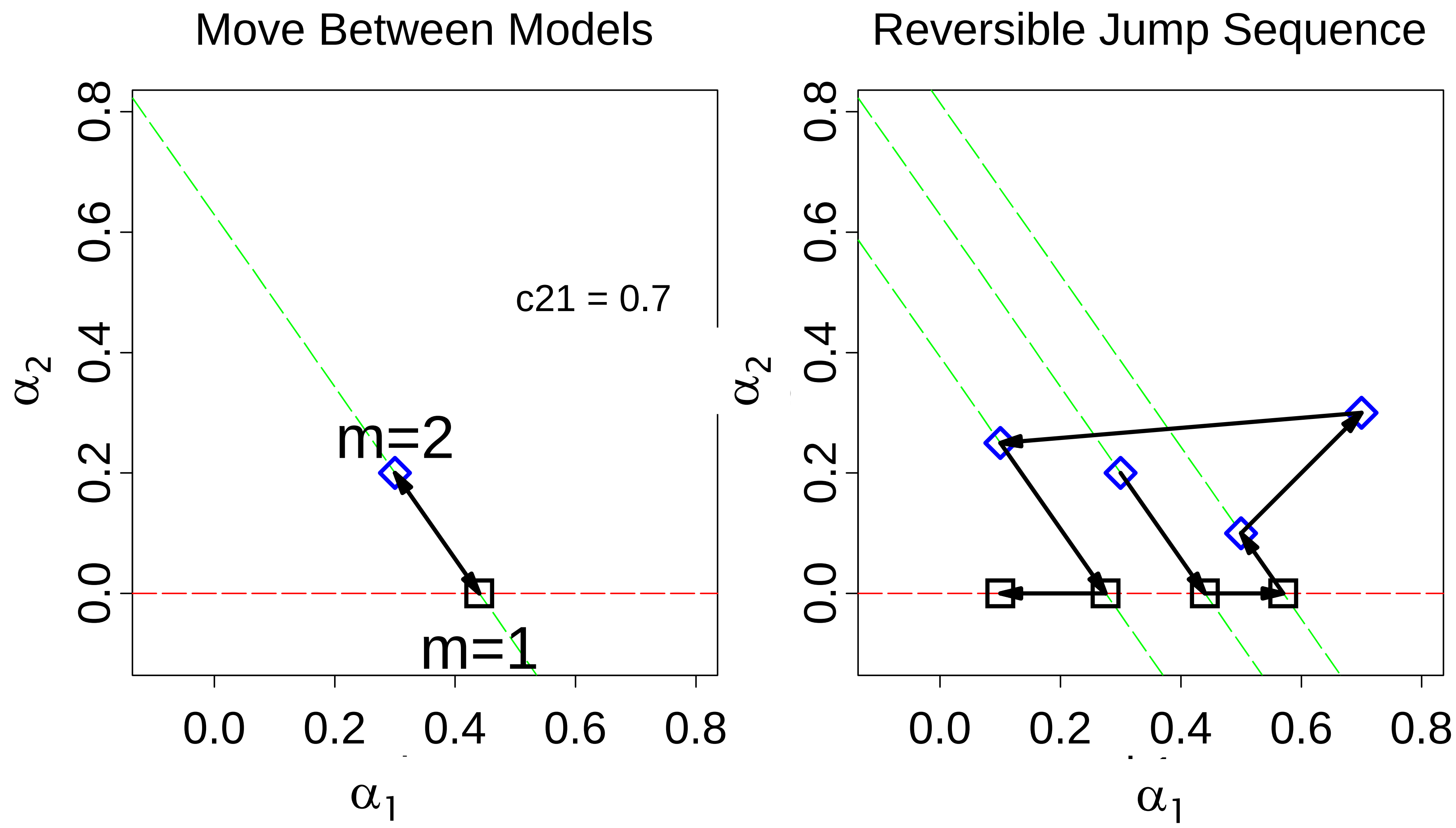
model selection in regression

- 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

$$m = 1 : Y = \mu + \alpha(Q_1 - \overline{Q}_1) + e$$

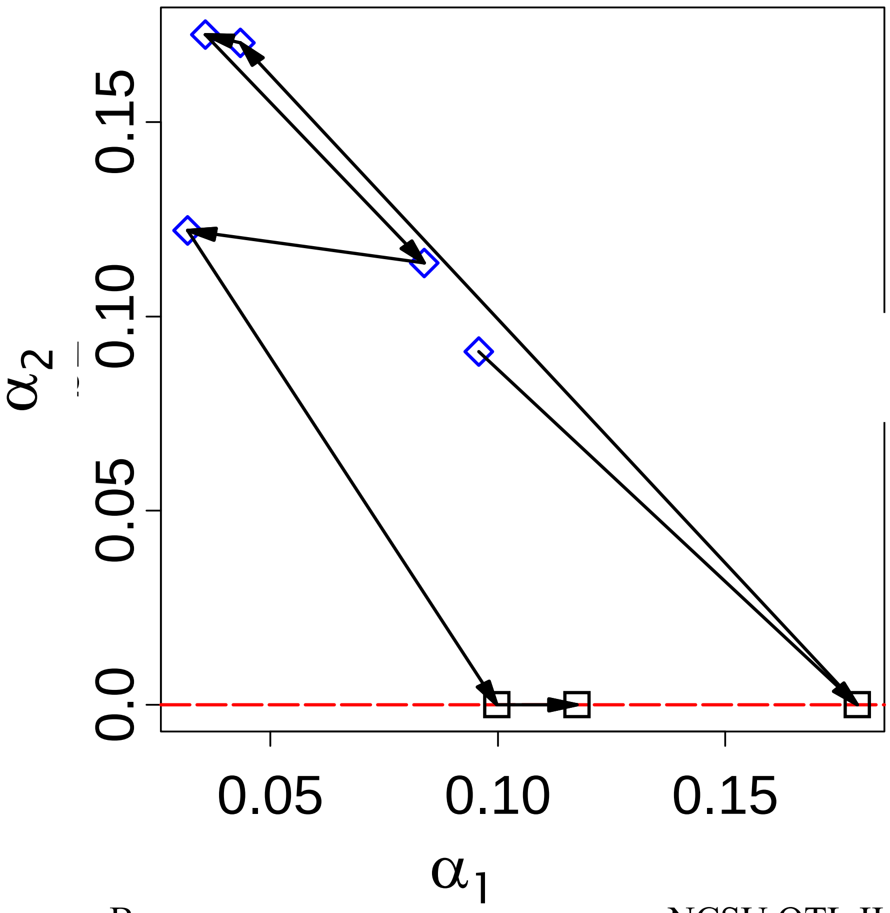
$$m = 2 : Y = \mu + \alpha_1(Q_1 - \overline{Q}_1) + \alpha_2(Q_1 - \overline{Q}_1) + e$$

geometry of reversible jump

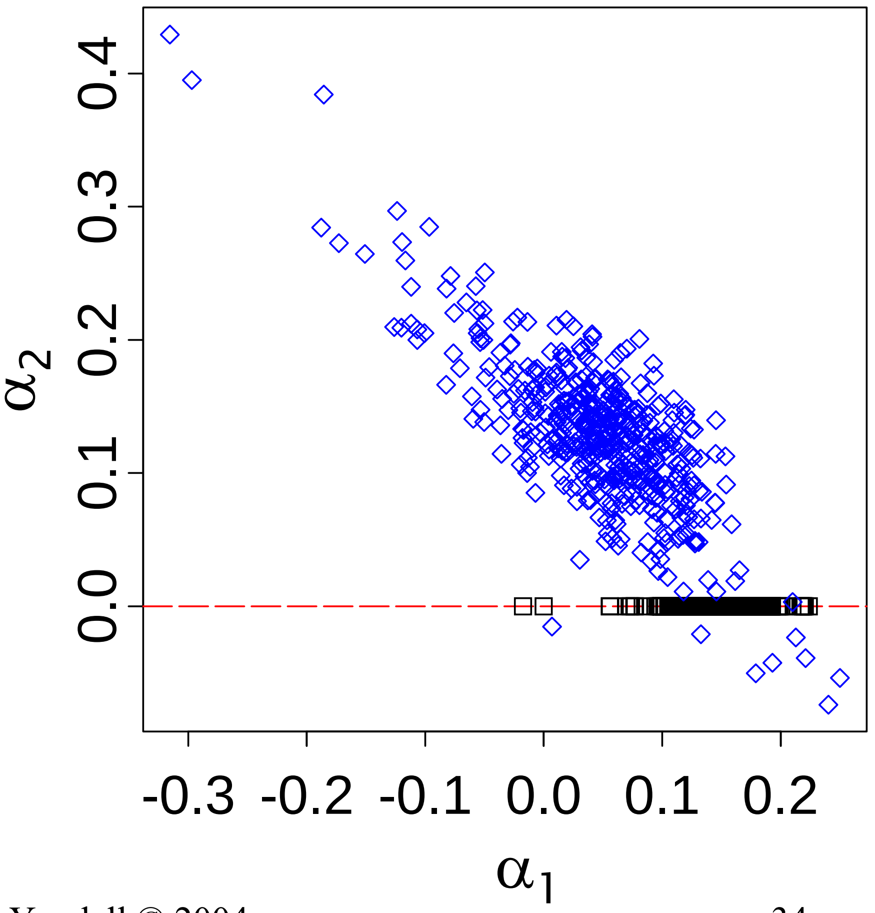


geometry allowing Q and λ to change

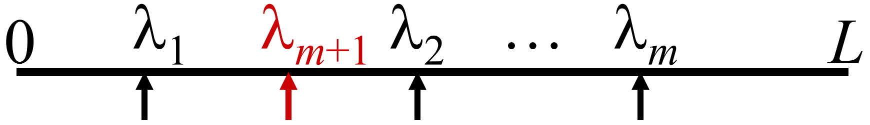
a short sequence



first 1000 with m<3



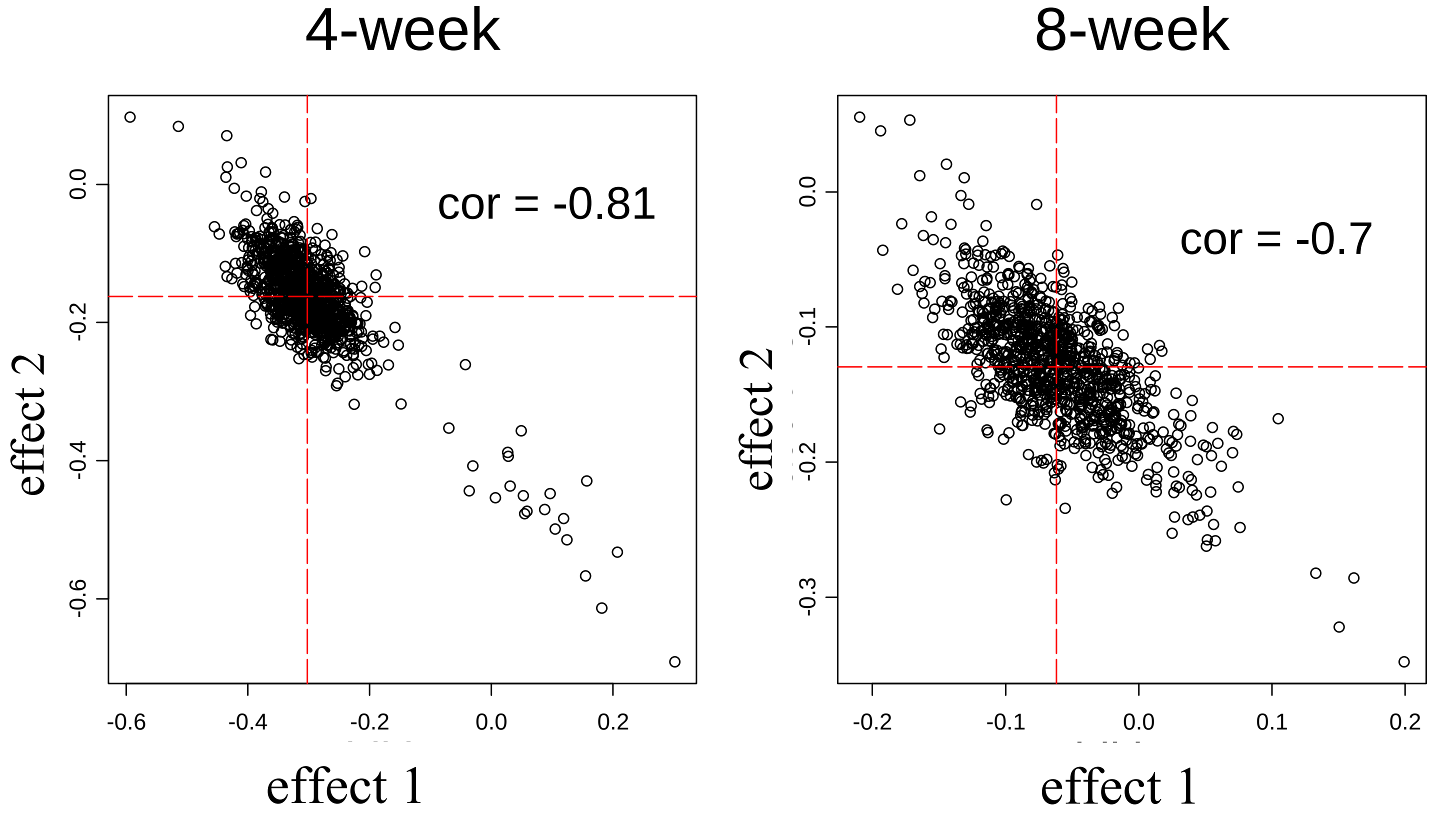
reversible jump MCMC



Metropolis-Hastings updates: 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

collinear QTL = correlated effects



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

R/bim: our RJ-MCMC software

- R: www.r-project.org
  - freely available statistical computing application R
  - library(bim) builds on Broman's library(qtl)
- QTLCart: statgen.ncsu.edu/qtlcart
- www.stat.wisc.edu/~yandell/qtl/software/Bmapqtl
- genesis
  - 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
  - incorporated into QTLCart (S Wang 2003)
  - built as official R library (H Wu, Yandell, Gaffney, CF Jin 2003)

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
  - δ = 1 if QTL in interval
  - δ = 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