4 Interval Mapping for a Single QTL

- basic idea of interval mapping
- interval mapping by maximum likelihood
  - maximum likelihood using EM, MCMC
- Bayesian interval mapping
  - "natural" Bayesian priors
  - multiple imputation, MCMC
- bootstrapped variance estimates
- advantages & shortcomings of IM
- Haley-Knott regression approximation

4.1 basic idea of interval mapping

- study properties of likelihood at each possible QTL
  - treating QTL as missing data
  - assuming only a single QTL (for now)
- recall likelihood as mixture over unknown QTL
  - likelihood = product of sum of products
  - complicated to evaluate--requires iteration

$$
\begin{aligned}
L(\theta,\lambda \mid Y,X) &= \mathrm{pr}(Y \mid X,\theta,\lambda) \\
&= \mathrm{prod}_i\ \mathrm{pr}(Y_i \mid X_i,\theta,\lambda) \\
&= \mathrm{prod}_i\ \mathrm{sum}_Q\ \mathrm{pr}(Q \mid X_i,\lambda)\mathrm{pr}(Y_i \mid Q,\theta)
\end{aligned}
$$

uncertainty in QTL genotype *Q*

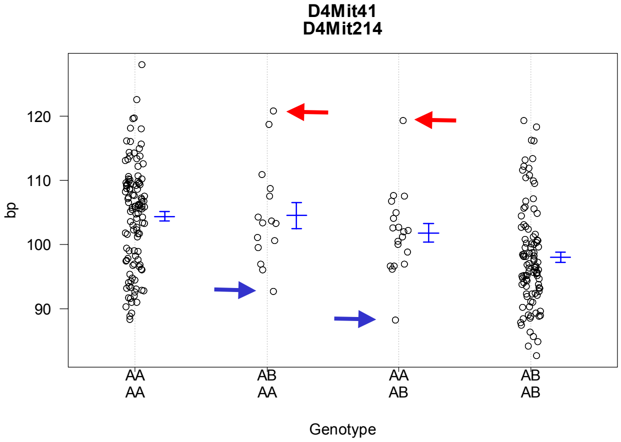
- how to improve guess on *Q* with data, parameters?
  - prior recombination: pr($Q \mid X_i, \lambda$)
  - posterior recombination: pr($Q \mid Y_i, X_i, \theta, \lambda$)
- main philosophies for assessing likelihood
  - maximum likelihood: study peak(s)
  - Bayesian analysis: study whole shape
- implementation methodologies
  - Expectation-Maximization (EM)
  - Markov chain Monte Carlo (MCMC)
  - multiple imputation
  - genetic algorithms, GEE, …

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($Y_i \mid Q, \theta$)
    - weight toward *Q* so that group mean $G_Q \approx Y_i$
- phenotype and flanking markers may conflict
  - posterior recombination balances these two weights

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

how does phenotype *Y* affect *Q*?



what are probabilities for genotype *Q* between markers?

recombinants AA:AB

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

maximum likelihood (ML) idea

- pick QTL locus λ (usually scan whole genome)
- find ML estimates of gene action θ given λ
- maximum likelihood at peak of likelihood
  - slope (derivative with respect to θ) is zero
  - sometimes maximum is at a boundary (non-zero slope)
- slope is weighted average using posteriors for *Q*
  - cannot write estimate in "closed form"
  - need to know θ to estimate it!
  - iterate toward the maximum in some clever way

$$\frac{dL(\theta,\lambda\,|\,Y,X)}{d\theta} = \text{sum}_{i,Q}\ \text{pr}(Q\,|\,Y_i,X_i,\theta,\lambda)\frac{d\log(\text{pr}(Y_i\,|\,Q,\theta))}{d\theta}$$

Bayesian model posterior

- augment data (*Y,X*) with unknowns *Q*
  - study unknowns (θ,λ,Q) given data (*Y,X*)
  - $Q \sim \text{pr}(Q \mid Y_i, X_i, \theta, \lambda)$
- no longer need weighted average over *Q*
  - instead we average over *Q* to study parameters
  - $\text{pr}(\theta,\lambda|Y,X) = \text{sum}_Q\ \text{pr}(\theta,\lambda,Q|Y,X)$
- study properties of posterior
  - need to specify priors for (θ,λ)
  - denominator is very difficult to compute in practice
  - drawing samples from posterior in some clever way

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

4.2 interval mapping by ML

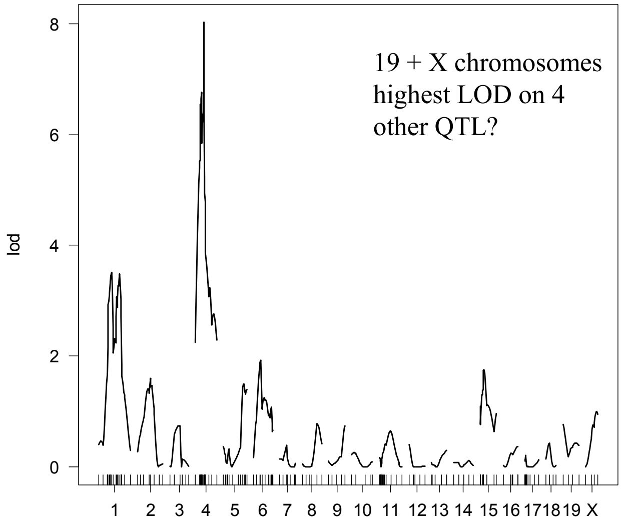
- search whole genome for putative QTL
- "profile" likelihood across all possible λ
  - find ML estimate of θ given λ
  - ML estimate of (θ,λ) at maximum over genome

$$L_0(\hat{\theta}_0 \mid Y) = \text{prod}_i f(Y_i \mid \hat{\mu}, s^2)$$

$$L(\hat{\theta},\lambda \mid Y,X) = \text{prod}_i\ \text{sum}_Q \text{pr}(Q \mid X,\lambda) f(Y_i \mid \hat{G}_Q, \hat{\sigma}^2_{\text{pool}})$$

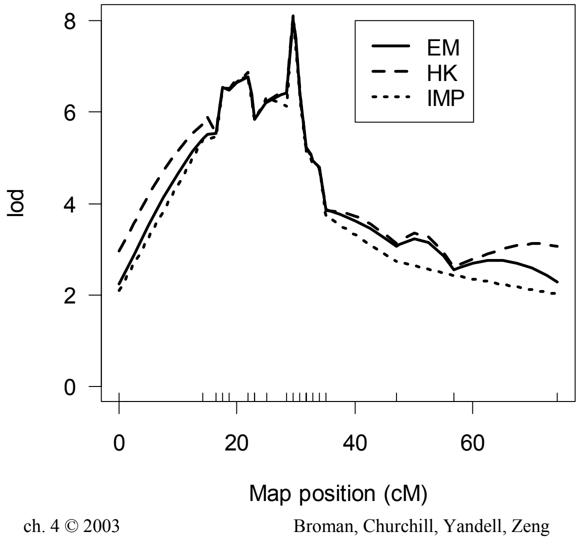
$$LOD(\lambda) = \log_{10}\left(\frac{L(\hat{\theta},\lambda \mid Y,X)}{L_0(\hat{\theta}_0 \mid Y)}\right)$$

LOD for hyper dataset



19 + X chromosomes
highest LOD on 4
other QTL?

LOD(λ) on chr 4 of hyper



EM "exact"
Haley-Knott regression
single imputation

all agree at the peak
and mostly at markers

note marker spacing

EM method for interval mapping

- fix a possible QTL λ
- iterate between expectation & maximization
  - – likelihood increases with each iteration
  - – stop iterating when the change is "negligible"
- initial values
  - – $P_{Qi} = \mathrm{pr}(Q \mid X_i, \lambda)$
    - recombination model in the absence of data
  - – or use Haley-Knott regression estimates of θ

EM method for interval mapping

- E-step: estimate posterior recombination
  - – $P_{Qi} = \mathrm{pr}(Q \mid Y_i, X_i, \theta, \lambda)$
  - – estimate for every individual *i*, genotype *Q*
  - – depends on effects θ
- M-steps: maximize likelihood for θ
  - – may be many parameters
  - – technical point: caution on parallel updates
  - – solve system of equation: derivatives set to zero
  - – depends on P_{Qi}

$$0 = \mathrm{sum}_{i,Q}\, P_{Qi} \frac{d \log(\mathrm{pr}(Y_i \mid Q, \theta))}{d\theta}$$

4.2.2 M-step for normal phenotype

- $Y = G_Q + e,\ e \sim N(0,\sigma^2)$
- $\text{pr}(Y \mid Q,\theta) = f(Y \mid G_Q,\ \sigma^2)$
- see notes in book for derivative details
- E-step estimates:

$$\hat{G}_Q = \text{sum}_i\ Y_i P_{Qi} / \text{sum}_i\ P_{Qi}$$

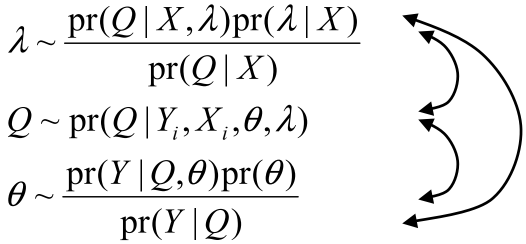
$$\hat{\sigma}^2 = \text{sum}_{i,Q} \left(Y_i - \hat{G}_Q\right)^2 P_{Qi} / n$$

4.2.3 ML via MCMC

- basic idea of simulated annealing
  - start with non-informative priors on (θ,λ)
  - sample from posterior (somehow…)
  - gradually shrink priors toward ML estimate
- slight difficulty
  - need to know (θ,λ) to sample from posterior
  - iteration leads to Markov chain
- point of this section
  - MCMC does not imply a Bayesian perspective!

4.3 Bayesian interval mapping

- sample missing genotypes Q
- decouple effects θ from QTL λ
- but Q depends on (θ,λ) and vice versa
- also need to specify priors



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

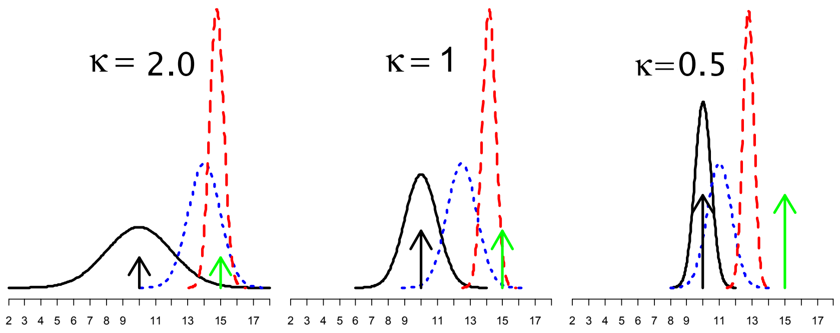
$$Q \sim \mathrm{pr}(Q \mid Y_i, X_i, \theta, \lambda)$$

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

4.3.1 Bayesian priors for QTL

- locus λ may be uniform over genome
  - pr(λ | X) = 1 / length of genome
- missing genotypes Q
  - pr(Q | X, λ)
  - recombination model is formally a prior
- effects $\theta = (G,\sigma^2)$, $G = (G_{QQ}, G_{Qq}, G_{qq})$
  - conjugate priors for normal phenotype
  - $G_Q \sim N(\mu, \kappa\sigma^2)$
  - $\sigma^2 \sim$ inverse-$\chi^2(\nu,\tau^2)$, or $\nu\tau^2 / \sigma^2 \sim \chi^2$

effect of prior variance on posterior



normal prior, posterior for n = 1, posterior for n = 5 , true mean
(solid black) (dotted blue) (dashed red) (green arrow)

details of phenotype priors

- priors depend on "hyper-parameters"
- $G_Q \sim N(\mu, \kappa\sigma^2)$
  - center around phenotype grand mean
  - $\kappa\sigma^2 \approx \sigma_G^2$ = genetic variance
  - $\kappa \approx \sigma_G^2 / \sigma^2 = h^2 / (1 - h^2)$
  - $h^2 = \sigma_G^2 / (\sigma_G^2 + \sigma^2)$ = heritability
- $\sigma^2 \sim$ inverse-$\chi^2(\nu, \tau^2)$, or $\nu\tau^2 / \sigma^2 \sim \chi^2$
  - $\tau^2 \approx s^2$ = total sample variance
  - ν = prior degrees of freedom = small integer

Bayes for normal data

$Y = G + E$	posterior for single individual
environ	$E \sim \mathrm{N}(0,\sigma^2)$, σ^2 known
likelihood	$\mathrm{pr}(Y \mid G,\sigma^2) = \mathrm{N}(Y \mid G,\sigma^2)$
prior	$\mathrm{pr}(G \mid \sigma^2,\mu,\kappa) = \mathrm{N}(G \mid \mu,\kappa\sigma^2)$
posterior	$\mathrm{N}(G \mid \mu+B_1(Y\text{-}\mu), B_1\sigma^2)$
$Y_i = G + E_i$	posterior for sample of n individuals
	shrinkage weights B_n go to 1

$$\mathrm{pr}(G \mid Y,\sigma^2,\mu,\kappa) = \mathrm{N}\left(G \middle| \mu + B_n(\overline{Y}_{\bullet} - \mu), B_n \frac{\sigma^2}{n} \right)$$

$$\text{with } \overline{Y}_{\bullet} = \text{sum}\frac{Y_i}{n}, B_n = \frac{\kappa n}{\kappa n + 1} \to 1$$

posterior by QT genetic value

$Y = G_Q + E$	genetic	Q = qq,Qq,QQ
	environ	$E \sim \mathrm{N}(0,\sigma^2)$, σ^2 known
	parameters	$\theta = (G,\sigma^2)$

likelihood	$\mathrm{pr}(Y \mid Q,G,\sigma^2) = \mathrm{N}(Y \mid G_Q,\sigma^2)$
prior	$\mathrm{pr}(G_Q \mid \sigma^2,\mu,\kappa) = \mathrm{N}(G_Q \mid \mu,\kappa\sigma^2)$

posterior:

$$\mathrm{pr}(G_Q \mid Y,Q,\sigma^2,\mu,\kappa) = \mathrm{N}\left(G_Q \middle| \mu + B_Q(\overline{Y}_Q - \mu), B_Q \frac{\sigma^2}{n_Q} \right)$$

$$n_Q = \text{count}\{Q_i = Q\}, \overline{Y}_Q = \underset{\{i:Q_i=Q\}}{\text{sum}} \frac{Y_i}{n_Q}, B_Q = \frac{\kappa n_Q}{\kappa n_Q + 1} \to 1$$

Empirical Bayes: choosing hyper-parameters

How do we choose hyper-parameters μ, κ?

Empirical Bayes: marginalize over prior
estimate μ, κ from marginal posterior

likelihood $\text{pr}(Y_i \mid Q_i, G, \sigma^2) = \text{N}(Y_i \mid G(Q_i), \sigma^2)$

prior $\text{pr}(G_Q \mid \sigma^2, \mu, \kappa) = \text{N}(G_Q \mid \mu, \kappa\sigma^2)$

marginal $\text{pr}(Y_i \mid \sigma^2, \mu, \kappa) = \text{N}(Y_i \mid \mu, (\kappa+1)\sigma^2)$

estimates $\hat{\mu} = \overline{Y}_\bullet, s^2 = \text{sum}_i (Y_i - \overline{Y}_\bullet)^2 / n$
$\hat{\sigma}^2 = s^2 / (\kappa + 1) \approx s^2 / (1 - h^2)$

EB posterior $$\text{pr}(G_Q \mid Y) = \text{N}\left(G_Q \middle| \overline{Y}_\bullet + B_Q(\overline{Y}_Q - \overline{Y}_\bullet), B_Q \frac{\hat{\sigma}^2}{n_Q} \right)$$

What if variance σ^2 is unknown?

- recall that sample variance is proportional to chi-square
  - pr($s^2 \mid \sigma^2$) = χ^2 ($ns^2/\sigma^2 \mid n$)
  - or equivalently, $ns^2/\sigma^2 \mid \sigma^2 \sim \chi_n^2$
- conjugate prior is inverse chi-square
  - pr($\sigma^2 \mid \nu, \tau^2$) = inv-$\chi^2$ ($\sigma^2 \mid \nu, \tau^2$)
  - or equivalently, $\nu\tau^2/\sigma^2 \mid \nu, \tau^2 \sim \chi_\nu^2$
  - empirical choice: $\tau^2 = s^2/3$, $\nu=6$
    - E($\sigma^2 \mid \nu,\tau^2$) = $s^2/2$, Var($\sigma^2 \mid \nu,\tau^2$) = $s^4/4$
- posterior given data
  - pr($\sigma^2 \mid Y, \nu, \tau^2$) = inv-$\chi^2$ ($\sigma^2 | \nu+n$, $(\nu\tau^2+ns^2)/(\nu+n)$)
  - weighted average of prior and data

joint effects posterior details

$Y_i = G(Q_i) + E_i$ genetic Q_i = qq,Qq,QQ

environ $E \sim \mathrm{N}(0,\sigma^2)$

parameters $\theta = (G,\sigma^2)$

likelihood $\mathrm{pr}(Y_i \mid Q_i, G, \sigma^2) = \mathrm{N}(Y_i \mid G(Q_i), \sigma^2)$

prior $\mathrm{pr}(G_Q \mid \sigma^2, \mu, \kappa) = \mathrm{N}(G_Q \mid \mu, \sigma^2/\kappa)$

$\mathrm{pr}(\sigma^2 \mid \nu, \tau^2) = \text{inv-}\chi^2\,(\sigma^2 \mid \nu, \tau^2\,)$

posterior:

$$\mathrm{pr}(G_Q \mid Y, Q, \sigma^2, \mu, \kappa) = \mathrm{N}\left(G_Q \middle| \overline{Y}_{\cdot} + B_Q(\overline{Y}_Q - \overline{Y}_{\cdot}), B_Q \frac{\sigma^2}{n_Q}\right)$$

$$\mathrm{pr}(\sigma^2 \mid Y, Q, G_Q, \nu, \tau^2) = \text{inv - }\chi^2\left(\sigma^2 \middle| \nu + n, \frac{\nu\tau^2 + ns_Q^2}{\nu + n}\right)$$

$$\text{with } B_Q = \frac{n_Q}{\kappa + n_Q}, s_Q^2 = \mathrm{sum}_i \left(Y_i - G(Q_i)\right)^2 / n$$

4.3.2 Bayesian multiple imputation

- basic idea
  - impute multiple copies of missing genotypes Q
    - sample $Q \sim \mathrm{pr}(Q \mid X, \lambda)$
    - weighted to appear as draws from posterior
  - average out gene effects θ
  - study posterior for putative QTL λ
- most effective for multiple QTL
  - use single QTL to introduce idea
  - consider all loci as possible QTL
    - sample on grid Λ of `pseudomarkers' (every 2cM)
    - similar to interval map scan of whole genome

importance sampling idea

- draw samples from one distribution
  - $Q_1, Q_2, Q_3, \ldots, Q_n \sim f(Q)$
- weight them appropriately by $\omega(Q)$
- sample summaries from distribution $g(Q)$
  - $g(Q) = f(Q)\omega(Q)$ / constant
  - mean for f = $\text{sum}_i\ Q_i / n$
  - mean for g = $\text{sum}_i\ Q_i\omega(Q_i)$ / $\text{sum}_i\ \omega(Q_i)$

example: mean copies of Q

genotype	qq	Qq	QQ	sum
Q copies	0	1	2	
true g	0.25	0.5	0.25	1.0
draw f	1/3	1/3	1/3	1.0
weight ω	1	2	1	
$f\times\omega$	1/3	2/3	1/3	4/3
importance sampling				$g = f\times0.75\omega$
sample	30	30	40	100
mean f	0×30	1×30	2×40	110/100=1.1
mean g	0×1×30	1×2×30	2×1×40	140/130=1.08

what are appropriate weights?

- ideally draw genotype from posterior
  - want sample $Q \sim g(Q) = \text{sum}_\theta \, \text{pr}(Q \mid Y, X, \theta, \lambda)\text{pr}(\theta)$
  - but have sample $Q \sim f(Q) = \text{pr}(Q \mid X, \lambda)$
- appropriate weights
  - $\omega(Q,\lambda | Y,X) = \text{pr}(\lambda \mid X) \; \text{sum}_\theta \, \text{pr}(Y \mid Q, \theta)\text{pr}(\theta)$
- estimate marginal posterior for QTL λ
  - draw N samples from prior at each QTL λ

    $Q_1, Q_2, Q_3, \ldots, Q_N \sim \text{pr}(Q \mid X, \lambda)$

    $\text{pr}(\lambda \mid Y,X) \;\; = \text{sum}_Q \; \omega(Q,\lambda|Y,X) \; \text{pr}(Q \mid X,\lambda) \,/\, \text{constant}$

    $\approx \text{sum}_j \; \omega(Q_j,\lambda|Y,X) \,/\, \text{constant}$
  - constant is summed over all λ, but not actually needed

relating weights to posterior

- posterior is simply averaged over θ
- weights comprise terms except $\text{pr}(Q \mid X, \lambda)$
- estimating weights: see Sen & Churchill

$$
\begin{aligned}
\text{pr}(\lambda, Q \mid Y, X) &= \text{sum}_\theta \; \text{pr}(\theta, \lambda, Q \mid Y, X) \\
&= \frac{\text{pr}(Q \mid X, \lambda)\text{pr}(\lambda \mid X) \, \text{sum}_\theta \; \text{pr}(Y \mid Q, \theta)\text{pr}(\theta)}{\text{pr}(Y \mid X)} \\
&= \text{pr}(Q \mid X, \lambda)\omega(Q, \lambda \mid Y, X) / \text{pr}(Y \mid X)
\end{aligned}
$$

estimating effects via imputation

- multiple imputation averages over effects
- difficult to study posterior of effects directly
- can estimate usual summaries

$$
\begin{aligned}
E(\theta \mid Y, X) &= \mathrm{sum}_Q \; E(\theta \mid Y, Q) \; \mathrm{pr}(Q \mid Y, X) \\
&= \mathrm{sum}_{Q,\lambda} \; E(\theta \mid Y, Q) \; \mathrm{pr}(Q \mid X, \lambda) \omega(Q, \lambda \mid Y, X) / \mathrm{pr}(Y \mid X) \\
&\approx \mathrm{sum}_{\lambda, j} \; E(\theta \mid Y, Q_j) \omega(Q_j, \lambda \mid Y, X) / \mathrm{constant}
\end{aligned}
$$

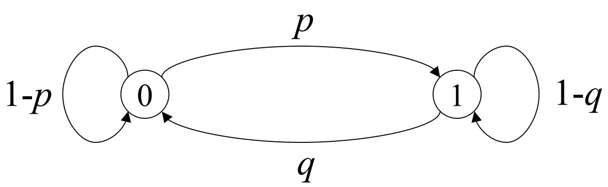
4.3.3 Bayesian MCMC

- Markov chain Monte Carlo
  - – Monte Carlo samples along a Markov chain
- What is a Markov chain?
- What is MCMC?
  - – Sampling from full conditionals
  - – Gibbs sampler, Metropolis-Hastings

What is a Markov chain?

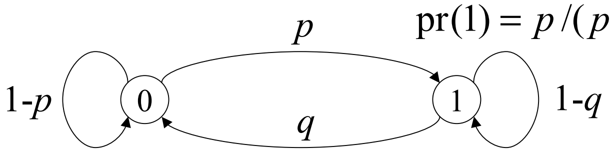
- future given present is independent of past
- update chain based on current value
  - can make chain arbitrarily complicated
  - chain converges to stable pattern π() we wish to study

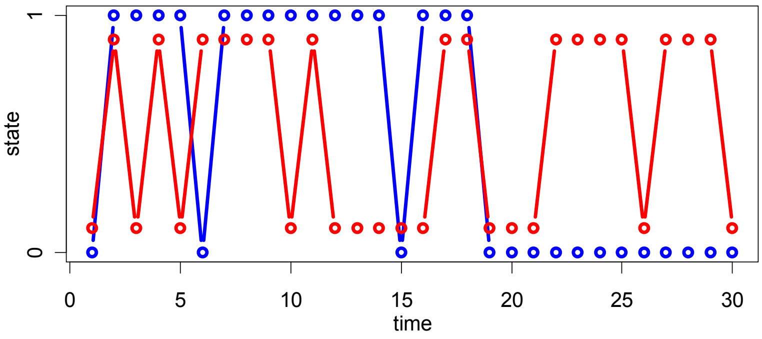
$$\mathrm{pr}(1) = p/(p+q)$$



Markov chain idea

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





Markov chain Monte Carlo

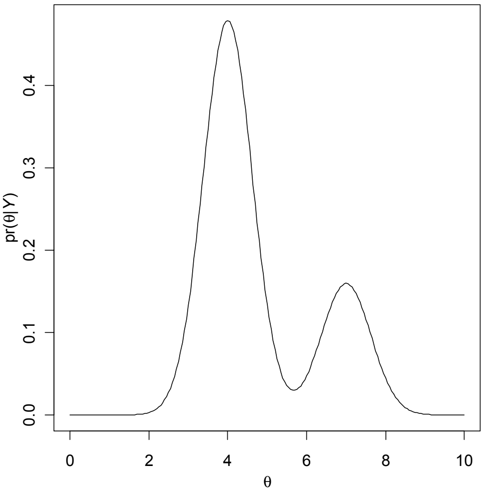
- can study arbitrarily complex models
  - need only specify how parameters affect each other
  - can reduce to specifying full conditionals
- construct Markov chain with "right" model
  - joint posterior of unknowns as limiting "stable" distribution
  - update unknowns given data and all other unknowns
    - sample from full conditionals
    - cycle at random through all parameters
  - next step depends only on current values
- nice Markov chains have nice properties
  - sample summaries make sense
  - consider almost as random sample from distribution
  - ergodic theorem and all that stuff

Markov chain Monte Carlo idea

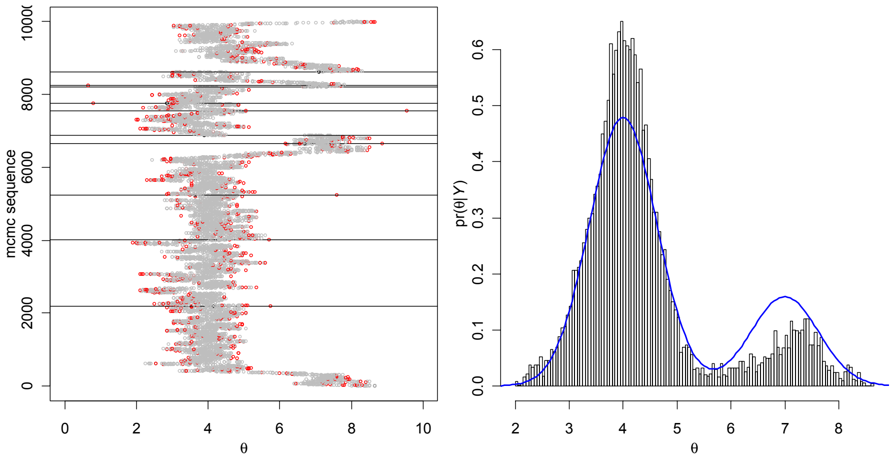
have posterior pr(θ|Y)
want to draw samples

propose θ ~ pr(θ|Y)
(ideal: Gibbs sample)

propose new θ "nearby"
accept if more probable
toss coin if less probable
based on relative heights
(Metropolis-Hastings)



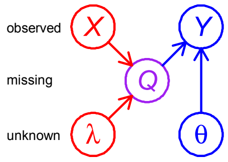
MCMC realization



added twist: occasionally propose from whole domain

marginal posteriors

- joint posterior
  - pr(λ,Q,θ|Y,X) = pr(θ)pr(λ)pr(Q|X,λ)pr(Y|Q,θ) /constant
- genetic effects
  - pr(θ|Y,X) = sum$_Q$ pr(θ|Y,Q) pr(Q|Y,X)
- QTL locus
  - pr(λ|Y,X) = sum$_Q$ pr(λ|X,Q) pr(Q|Y,X)


observed
missing
unknown

- QTL genotypes more complicated
  - pr(Q|Y,X) = sum$_{\lambda,\theta}$ pr(Q|Y,X,λ,θ) pr(λ,θ|Y,X)
  - impossible to separate λ and θ in sum

Why not Ordinary Monte Carlo?

- independent samples of joint distribution
- chaining (or peeling) of effects

$$\mathrm{pr}(\theta|Y,Q)=\mathrm{pr}(G_Q \mid Y,Q,\sigma^2)\ \mathrm{pr}(\sigma^2 \mid Y,Q)$$

- possible analytically here given genotypes Q
- Monte Carlo: draw N samples from posterior
  - sample variance σ^2
  - sample genetic values G_Q given variance σ^2
- but we know markers X, not genotypes Q!
  - would have messy average over possible Q
  - $\mathrm{pr}(\theta|Y,X) = \mathrm{sum}_Q\ \mathrm{pr}(\theta|Y,Q)\ \mathrm{pr}(Q|Y,X)$

MCMC Idea for QTLs

- 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 components from full conditionals
  - update effects θ given genotypes & traits
  - update locus λ given genotypes & marker map
  - update genotypes Q given traits, marker map, locus & effects

$$(\lambda,Q,\theta) \sim \mathrm{pr}(\lambda,Q,\theta \mid Y,X)$$
$$(\lambda,Q,\theta)_1 \rightarrow (\lambda,Q,\theta)_2 \rightarrow \cdots \rightarrow (\lambda,Q,\theta)_N$$

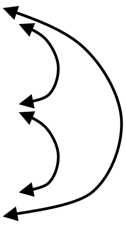
sample from full conditionals

- hard to sample from joint posterior
- update each unknown given all others
- examine posterior: keep terms with unknown
- normalizing denominator make a distribution

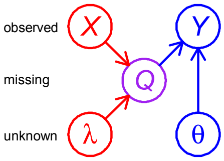
$$\lambda \sim \frac{\text{pr}(Q \mid X, \lambda)\text{pr}(\lambda \mid X)}{\text{pr}(Q \mid X)}$$

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

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



sample from full conditionals for model with *m* QTL



- hard to sample from joint posterior
  - $\text{pr}(\lambda, Q, \theta \mid Y, X) = \text{pr}(\theta)\text{pr}(\lambda)\text{pr}(Q|X, \lambda)\text{pr}(Y|Q, \theta)$ /constant
- easy to sample parameters from full conditionals
  - full conditional for genetic effects
    - $\text{pr}(\theta \mid Y, X, \lambda, Q) = \text{pr}(\theta \mid Y, Q) = \text{pr}(\theta)\ \text{pr}(Y|Q, \theta)$ /constant
  - full conditional for QTL locus
    - $\text{pr}(\lambda \mid Y, X, \theta, Q) = \text{pr}(\lambda \mid X, Q) = \text{pr}(\lambda)\ \text{pr}(Q|X, \lambda)$ /constant
  - full conditional for QTL genotypes
    - $\text{pr}(Q|Y, X, \lambda, \theta) = \text{pr}(Q|X, \lambda)\ \text{pr}(Y|Q, \theta)$ /constant

Gibbs sampler idea

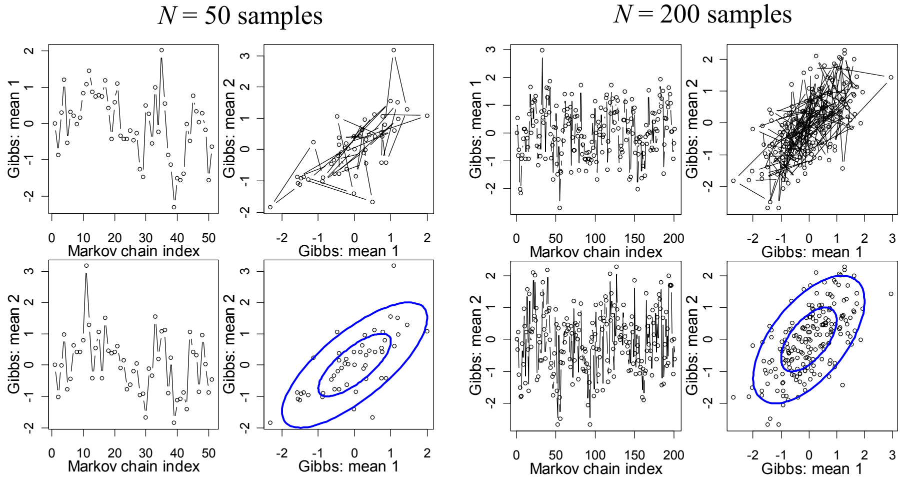
- want to study two correlated normals
- could sample directly from bivariate normal
- Gibbs sampler:
  - sample each from its full conditional
  - pick order of sampling at random
  - repeat N times

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

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

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

Gibbs sampler samples: $\rho = 0.6$



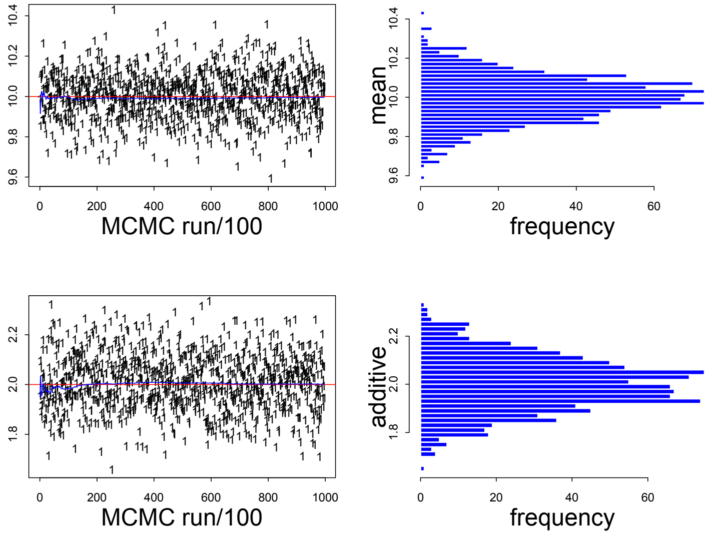
Gibbs Sampler: effects & genotypes

- for given locus λ, can sample effects θ and genotypes Q
  - effects parameter vector $\theta = (G, \sigma^2)$ with $G=(G_{\mathrm{qq}}, G_{\mathrm{Qq}}, G_{\mathrm{QQ}})$
  - missing genotype vector $Q = (Q_1, Q_2, \ldots, Q_n)$
- Gibbs sampler: update one at a time via full conditionals
  - randomly select order of unknowns
  - update each given current values of all others, locus λ and data (Y,X)
    - sample variance σ^2 given Y,Q and genetic values G
    - sample genotype Q_i given markers X_i and locus λ
  - can do block updates if more efficient
    - sample all genetic values G given Y,Q and variance σ^2

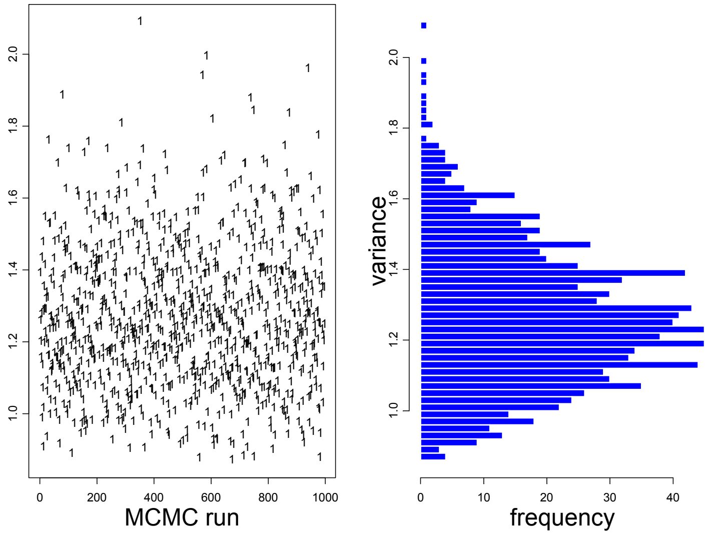
phenotype model: alternate form

- genetic value G_Q in "cell means" form easy
- but often useful to model effects directly
  - sort out additive and dominance effects
  - useful for reduced models with multiple QTL
    - QTL main effects and interactions (pairwise, 3-way, etc.)
- we only consider additive effects here
  - $G_{\mathrm{qq}} = \mu - a$, $G_{\mathrm{Qq}} = \mu$, $G_{\mathrm{QQ}} = \mu + a$
- recoding for regression model
  - $Q_i = -1$ for genotype qq
  - $Q_i = 0$ for genotype Qq
  - $Q_i = 1$ for genotype QQ
  - $G(Q_i) = \mu + aQ_i$

MCMC run of mean & additive



MCMC run for variance



missing marker data

- sample missing marker data a la QT genotypes
- full conditional for missing markers depends on
  - flanking markers
  - possible flanking QTL
- can explicitly decompose by individual *i*
  - binomial (or trinomial) probability

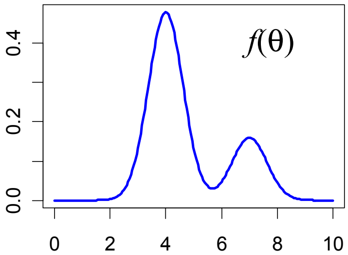
$$X_{ik} = \text{aa, Aa or AA}$$

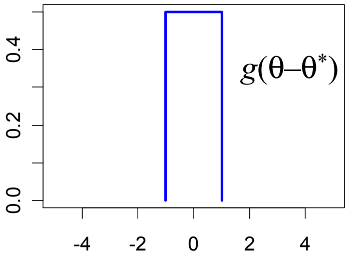
$$\text{pr}(X_{ik} \mid Y_i, X_i, Q_i, \theta, \lambda) = \text{pr}(X_{ik} \mid X_i, Q_i, \lambda)$$

Metropolis-Hastings idea

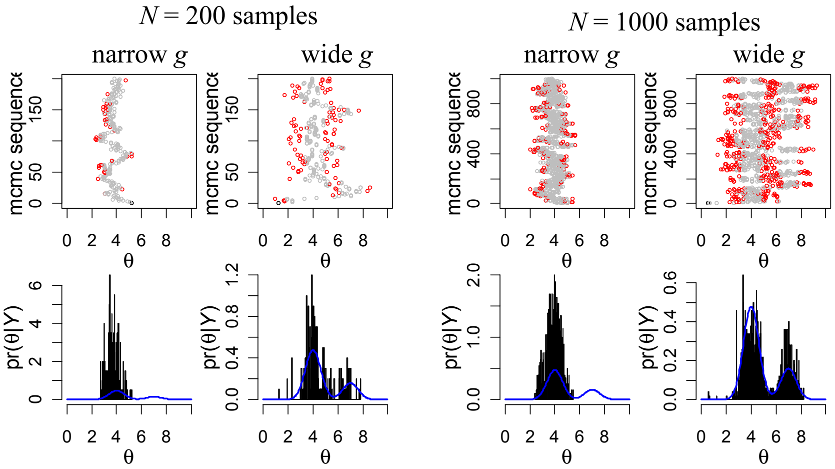
- want to study distribution *f*(θ)
- take Monte Carlo samples
  - unless too complicated
- Metropolis-Hastings samples:
  - current sample value θ
  - propose new value θ^*
    - from some distribution $g(\theta,\theta^*)$
    - Gibbs sampler: $g(\theta,\theta^*) = f(\theta^*)$
  - accept new value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

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





Metropolis-Hastings samples



full conditional for locus

- cannot easily sample from locus full conditional

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

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

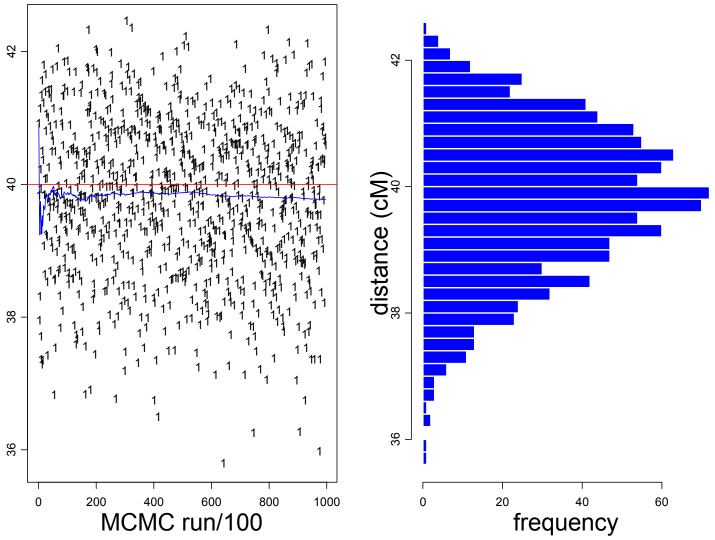
- cannot explicitly determine full conditional
  - difficult to normalize
  - need to average over all possible genotypes over entire map
- Gibbs sampler will not work
  - but can use method based on ratios of probabilities…

Metropolis-Hastings Step

- pick new locus based upon current locus
  - propose new locus from distribution *q*( )
    - pick value near current one?
    - pick uniformly across genome?
  - accept new locus with probability *a*()
- Gibbs sampler is special case of M-H
  - always accept new proposal
- acceptance insures right stable distribution
  - accept new proposal with probability *A*
  - otherwise stick with current value

$$A(\lambda_{old},\lambda_{new}) = \min\left(1, \frac{\pi(\lambda_{new} \mid \mathbf{x}^*)q(\lambda_{new},\lambda_{old})}{\pi(\lambda_{old} \mid \mathbf{x}^*)q(\lambda_{old},\lambda_{new})}\right)$$

MCMC Run for 1 locus at 40cM



Care & Use of MCMC

- sample chain for long run (100,000-1,000,000)
  - longer for more complicated likelihoods
  - use diagnostic plots to assess "mixing"
- standard error of estimates
  - use histogram of posterior
  - compute variance of posterior--just another summary
- studying the Markov chain
  - Monte Carlo error of series (Geyer 1992)
    - time series estimate based on lagged auto-covariances
  - convergence diagnostics for "proper mixing"

4.4 bootstrapped variance estimates

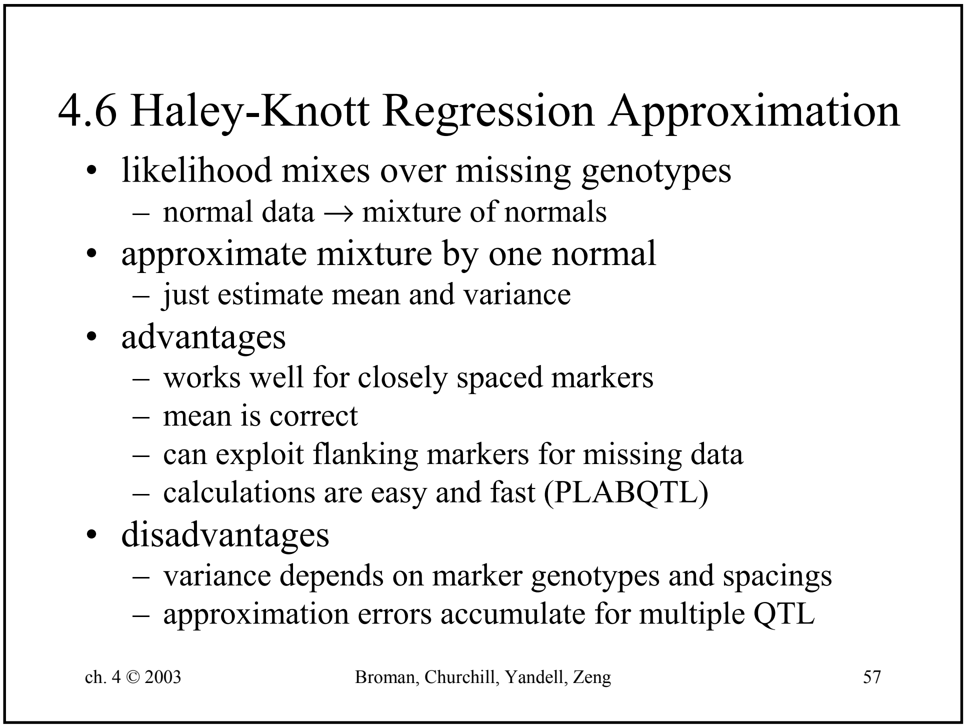
- (re)sample (Y_i,X_i) with replacement
  - create bootstrap sample "new" data of size n
  - estimate loci λ and effects θ
- repeat this N times
- construct summaries of these
  - mean, variance, median, percentile
- construct 95% confidence intervals for λ and θ
  - (2.5%ile, 97.5%ile)
  - order estimates, pick number $.025N$ and $.975N$

4.5 advantages & shortcomings of IM

- advantages over single marker analysis
  - can infer position and effect of QTL
  - estimated locations & effects almost unbiased
    - if only one segregating QTL per chromosome
  - requires fewer individuals for detection of QTL

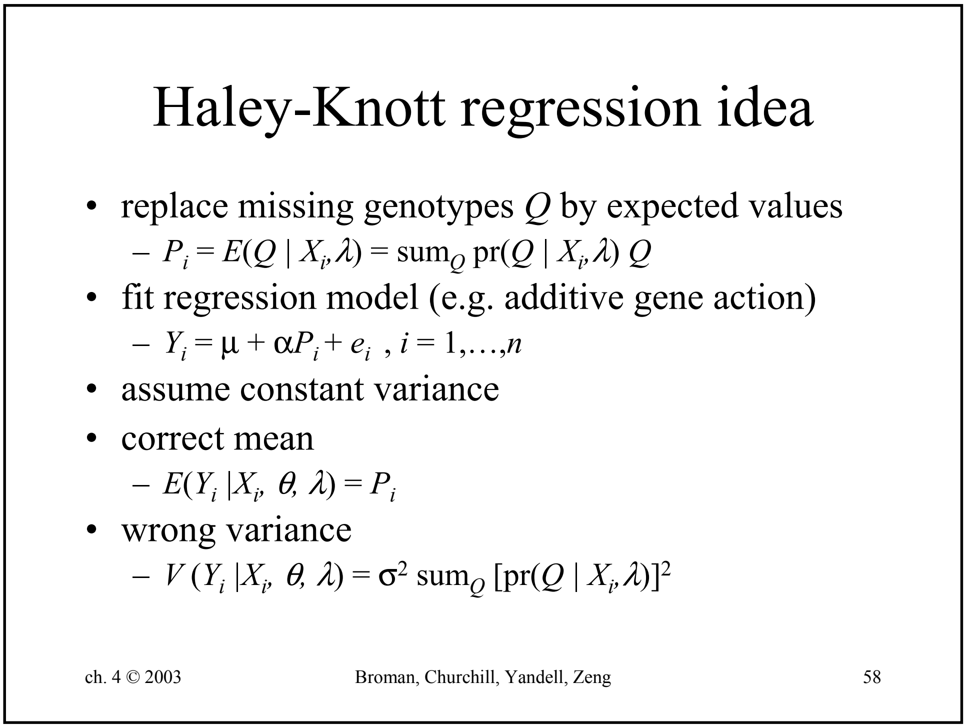
shortcomings of IM

- not an interval test
  - cannot say whether or not QTL is in an interval
  - not independent of effects of QTL outside interval
- can give false positives due to linkage
  - high LOD score due to nearby QTL
  - less of a problem for unlinked QTL
- can detect "ghost QTL"
  - higher peak between two linked QTL
  - estimated position and effect are biased



4.6 Haley-Knott Regression Approximation

- likelihood mixes over missing genotypes
  - normal data → mixture of normals
- approximate mixture by one normal
  - just estimate mean and variance
- advantages
  - works well for closely spaced markers
  - mean is correct
  - can exploit flanking markers for missing data
  - calculations are easy and fast (PLABQTL)
- disadvantages
  - variance depends on marker genotypes and spacings
  - approximation errors accumulate for multiple QTL



Haley-Knott regression idea

- replace missing genotypes Q by expected values
  - $P_i = E(Q \mid X_i, \lambda) = \text{sum}_Q \text{ pr}(Q \mid X_i, \lambda)\, Q$
- fit regression model (e.g. additive gene action)
  - $Y_i = \mu + \alpha P_i + e_i$, $i = 1,\ldots,n$
- assume constant variance
- correct mean
  - $E(Y_i \mid X_i, \theta, \lambda) = P_i$
- wrong variance
  - $V(Y_i \mid X_i, \theta, \lambda) = \sigma^2 \text{ sum}_Q [\text{pr}(Q \mid X_i, \lambda)]^2$

Haley-Knott and EM

- both use expected value of genotypes Q
  - HK: $P_i = E(Q \mid X_i, \lambda)$ = prior expectation
  - EM: $P_i = E(Q \mid Y_i, X_i, \theta, \lambda)$ = prior expectation
- both solve regression problems for effects
- difference is in iteration
  - HK is first step iteration
  - EM iterates E-step and M-steps to convergence