Bayesian Inference for QTLs in Inbred Lines

Brian S Yandell
University of Wisconsin-Madison
`www.stat.wisc.edu/~yandell`
NCSU Statistical Genetics
June 2000

Many Thanks

Michael Newton	Tom Osborn
Daniel Sorensen	David Butruille
Daniel Gianola	Marcio Ferrera
Jaya Satagopan	Josh Udahl
Patrick Gaffney	Pablo Quijada
Fei Zou	Alan Attie
Liang Li	Jonathan Stoehr

USDA Hatch Grants

What is the Goal Today?

- show MCMC ideas
  - Gibbs sampler
  - Metropolis-Hastings
- handle hard problems
  - image analysis
  - genetics
  - large dependent data
- resampling our data
  - permutation tests
  - MCMC
  - other (bootstrap,...)
- Bayesian perspective
  - common in animal model
  - use "prior" information
    - previous experiments
    - related genomes
- inbred lines "easy"
  - can check against *IM
  - ready extension
    - multiple experiments
    - pedigrees
    - non-normal data

Note on Outbred Studies

- Interval Mapping
  - Haley, Knott & Elsen (1994) *Genetics*
  - Thomas & Cortessis (1992) *Hum. Hered.*
  - Hoeschele & vanRanden (1993ab) *Theor. Appl. Genet.* (etc.)
  - Guo & Thompson (1994) *Biometrics*
- Nuances -- faking it
  - experimental outbred crosses
    - collapse markers from 4 to 2 alleles
  - pedigrees
    - polygenic effects not modeled here
    - related individuals are correlated (via coancestry)

Overview

- I: Single QTL
- II: Bayesian Idea
  - Bayes rule
  - posterior & likelihood
- III: MCMC Samples
  - Monte Carlo idea
  - study posterior
- IV: MCMC Details
- V: Multiple QTL
- VI: How many QTL?
  - Reversible Jump
  - analog to regression
- VII: RJ-MCMC Details
- VIII: Bayes Factors
- IX: References
  - Software
  - Articles

Part I: Interval Mapping

- Modelling a `trait` with a QTL
  - linear model for `trait` given genotype
  - recombination near `loci` for genotype
- Likelihoods
- Review Interval Maps & Profile LODs

QTL Components

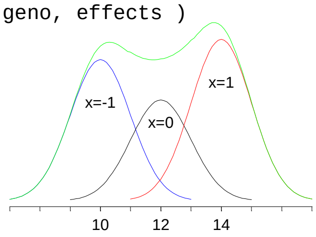
- observed data on individual
  - trait: field or lab measurement
    - log(days to flowering) , yield, …
  - markers: from wet lab (RFLPs, etc.)
    - linkage map of markers assumed known
- unobserved data on individual
  - geno: genotype (QQ=1/Qq=0/qq=-1)
- unknown model parameters
  - effects: mean, difference, variance
  - locus: quantitative trait locus (QTL)

Single QTL trait Model

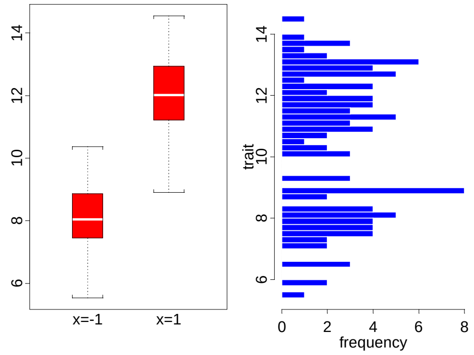
- trait = mean + additive + error
- trait = effect_of_geno + error
- *prob*(trait | geno, effects)

$$y_j=\mu+b^*x_j^*+e_j$$

$$\pi(y_j\mid x_j^*;\mu,b^*,\sigma^2)=\phi\left(\frac{y_j-\mu-b^*x_j^*}{\sigma}\right)$$



Simulated Data with 1 QTL

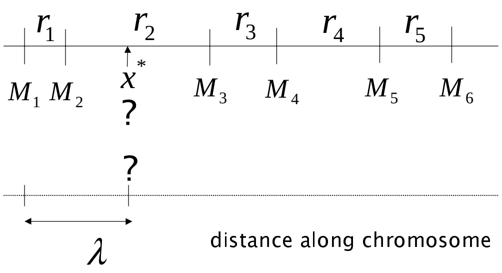


Recombination and Distance

- no interference--easy approximation
  - Haldane map function
  - no interference with recombination
- all computations consistent in approximation
  - rely on given map
    - marker loci assumed known
  - 1-to-1 relation of distance to recombination
  - all map functions are approximate
- assume marker positions along map are known

$$r=\tfrac{1}{2}\left(1-e^{-2\lambda}\right)$$

markers, QTL & recombination rates



distance along chromosome

Interval Mapping of QT genotype

- can express probabilities in terms of distance
  - locus is distance along linkage map
  - flanking markers sufficient if no missing data
  - could consider more complicated relationship

prob(geno | locus, map)
= *prob*(geno | locus, flanking markers)

$$\pi(x_j^*\mid\lambda)=\pi(x_j^*\mid\lambda,M_{j,k},M_{j,k+1})$$

Building trait Likelihood

- likelihood is mixture across possible genotypes
- sum over all possible genotypes at locus

like(effects, locus | trait)
= sum of *prob*(trait, genos | effects, locus)

$$L(\mu,b^*,\sigma^2,\lambda\,|\,y_j)=\pi(y_j\,|\,\mu,b^*,\sigma^2;\lambda)$$

$$=\sum_{x=-1,0,1}\pi(y_j\,|\,x;\mu,b^*,\sigma^2)\pi(x\,|\,\lambda)$$

Likelihood over Individuals

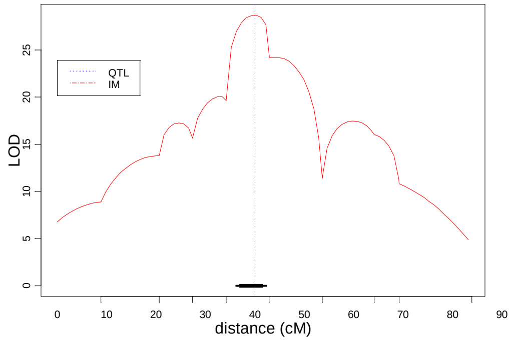
- product of trait probabilities across individuals
  - product of sum across possible genotypes

like(effects, locus | traits, map)
= product of *prob*(trait | effects, locus, map)

$$L(\mu,b^*,\sigma^2;\lambda\,|\,\mathbf{y})=\prod_{j=1}^{n}\pi(y_j\,|\,\mu,b^*,\sigma^2;\lambda)$$

$$=\prod_{j=1}^{n}\sum_{x=-1,0,1}\pi(y_j\,|\,x;\mu,b^*,\sigma^2)\pi(x\,|\,\lambda)$$

Profile LOD for 1 QTL



Interval Mapping for Quantitative Trait Loci

- profile likelihood (LOD) across QTL
  - scan whole genome locus by locus
    - use flanking markers for interval mapping
  - maximize likelihood ratio (LOD) at locus
    - best estimates of effects for each locus
    - EM method (Lander & Botstein 1989)

$$LOD(\lambda)=(\log_{10}e)\sum_{j=1}^{n}\ln\left(\frac{\sum_{x=-1,0,1}\pi(y_j\,|\,x;\hat{\mu},\hat{b}^*,\hat{\sigma}^2)\pi(x\,|\,\lambda)}{\pi(y_j\,|\,\hat{\mu},b^*=0,\hat{\hat{\sigma}}^2)}\right)$$

Interval Mapping Tests

- profile LOD across possible loci in genome
  - maximum likelihood estimates of effects at locus
  - LOD is rescaling of *L*(effects, locus|y)
- test for evidence of QTL at each locus
  - LOD score (*LR* test)
  - adjust (?) for multiple comparisons

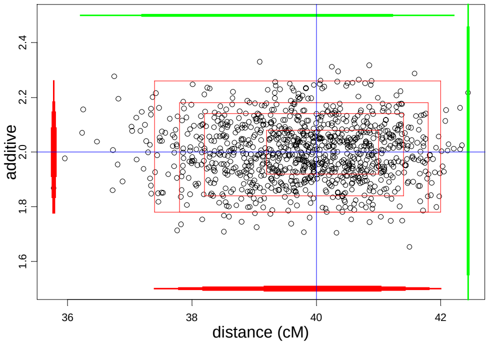
Interval Mapping Estimates

- confidence region for locus
  - based on inverting test of no QTL
  - 2 LODs down gives approximate CI for locus
  - based on chi-square approximation to *LR*
- confidence region for effects
  - approximate CI for effect based on normal
  - point estimate from profile LOD

$$locus\ CI=\left\{\lambda\,|\,LOD(\hat{\lambda})-LOD(\lambda)<2\right\}$$

$$effect\ CI=\hat{b}^*\pm 1.96se(\hat{b}^*)$$

IM Confidence Region



Part II: Bayesian Idea

- joint distribution of known & unknown
  - known: trait, markers, linkage map
  - unknown: locus, genotype, effect, variance
- Use Same Likelihood Components
  - trait given genotype
    - follows linear model
    - depends on size of effect, variance
  - genotype given locus, markers & map
    - depends on recombination near locus
- Inference about unknowns
  - Bayes theorem

Bayes Theorem

- posteriors and priors
  - prior: *prob*(parameters)
  - posterior: *prob*(parameters | data)
- posterior = likelihood * prior / constant
- posterior distribution is proportional to
  - likelihood of parameters given data
  - prior distribution of parameters

$$\pi(param|data)=\frac{\pi(param,data)}{\pi(data)}=\frac{\pi(data|param)\times\pi(param)}{\pi(data)}$$

$$\pi(param|data)=\pi(data|param)\times\pi(param)/C$$

What is Probability?

Frequentist Analysis

repeat experiment

- many times
- hypothetical

long term frequency

- Type I error rate
- reject null when true

Bayesian Analysis

uncertainty about true value

prior

- uncertainty before examining data
- incorporate prior knowledge/experience

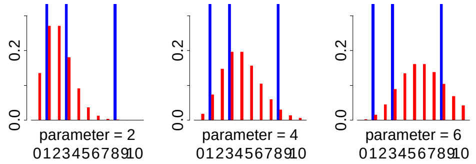
posterior

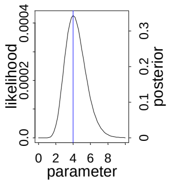
- uncertainty after analyzing current data
- balance prior & current

Bayesian Prior

- “prior” belief used to infer “posterior” estimates
  - higher weight for more probable parameter values
    - based on prior knowledge
  - use previous study to inform current study
    - weather prediction: tomorrow is much like today
    - previous QTL studies on related organisms
  - historical criticism: can get “religious” about priors
- often want negligible effect of prior on posterior
  - pick non-informative priors
    - all parameter values equally likely
    - large variance on priors
  - always check sensitivity to prior

Likelihood & Posterior Example





$$data: y = 1,3,8$$

$$parameter: t = ?$$

$$prob\{y = k \mid t\} = \frac{t^k e^{-t}}{k!}$$

Bayesian Idea for QTLs

- Modelling a `trait` with a QTL
  - linear model for `trait` given genotype
  - recombination near `loci` for genotype
- Bayesian Posterior
- Likelihoods
  - EM & MCMC
  - Frequentists & Bayesians
- Review Interval Maps & Profile LODs
- Case Study: Simulated Single QTL

QTL Effect Posterior

- posterior = likelihood * prior / constant

- posterior distribution is proportional to
  - prior distribution of `effect`
  - likelihood of `traits` given `effect` & genos

$$\pi(b^* \mid \mathbf{y})$$

is proportional to

$$\pi(b^*)\prod_{j=1}^{n}\pi(y_j \mid x_j^*; \mu, b^*, \sigma^2)$$

QTL Full Posterior

- posterior = likelihood * prior / constant
- posterior(paramaters | data)

  prob`( genos, effects, loci | trait, map )`

$$\pi(\mathbf{x}^*; \mu, b^*, \sigma^2; \lambda \mid \mathbf{y})$$

is proportional to

$$\pi(\mu)\pi(b^*)\pi(\sigma^2)\pi(\lambda)\prod_{j=1}^{n}\pi(x_j^* \mid \lambda)$$

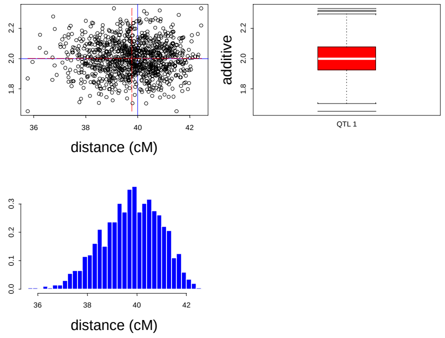
$$\times\prod_{j=1}^{n}\pi(y_j \mid x_j^*; \mu, b^*, \sigma^2)$$

How to Study Posterior?

- exact methods
  - exact if possible
  - can be difficult or impossible to analyze
- approximate methods
  - importance sampling
  - numerical integration
  - Monte Carlo & other

- Monte Carlo methods
  - easy to implement
  - independent samples
- MCMC methods
  - handle hard problems
  - art to efficient use
  - correlated samples

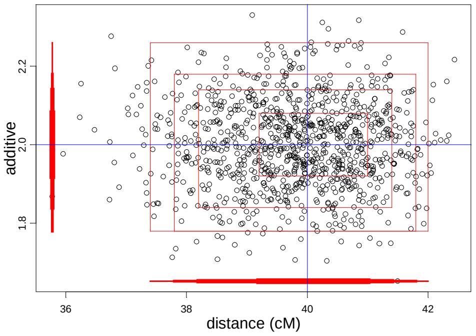
Posterior for `locus` & `effect`



Marginal Posterior Summary

- marginal posterior for locus & effects
- highest probability density (HPD) region
  - smallest region with highest probability
  - credible region for locus & effects
- HPD with 50,80,90,95%
  - range of credible levels can be useful
  - marginal bars and bounding boxes
  - joint regions (harder to draw)

HPD Region for locus & effect



QTL Bayesian Inference

- study posterior distribution of locus & effects
  - sample joint distribution
    - locus, effects & genotypes
  - study marginal distribution of
    - locus
    - effects
      - overall mean, genotype difference, variance
    - locus & effects together
- estimates & confidence regions
  - histograms, boxplots & scatter plots
  - HPD regions

Frequentist or Bayesian?

- Frequentist approach
  - fixed parameters
    - range of values
  - maximize likelihood
    - ML estimates
    - find the peak
  - confidence regions
    - random region
    - invert a test
  - hypothesis testing
    - 2 nested models
- Bayesian approach
  - random parameters
    - distribution
  - posterior distribution
    - posterior mean
    - sample from dist
  - credible sets
    - fixed region given data
    - HPD regions
  - model selection/critique
    - Bayes factors

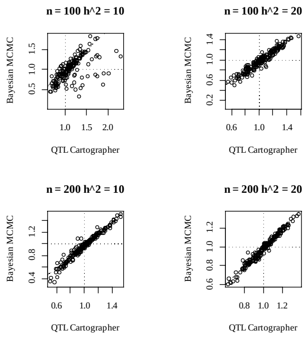
Frequentist or Bayesian?

- Frequentist approach
  - maximize over mixture of QT genotypes
  - locus profile likelihood
    - max over effects
  - HPD region for locus
    - natural for locus
      - 1-2 LOD drop
    - work to get effects
      - approximate shape of likelihood peak
- Bayesian approach
  - joint distribution over QT genotypes
  - sample distribution
    - joint effects & loci
  - HPD regions for
    - joint locus & effects
    - use density estimator

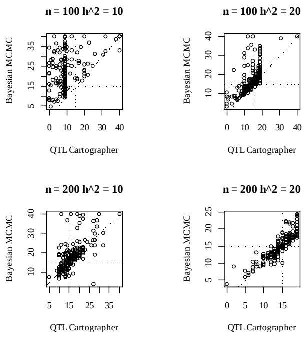
Simulation Study

- 200 simulation runs
- n = 100, 200; h^2 = 10, 20%
- 1 QTL at 15cM
- markers at 0, 10, 20, 40, 60, 80
- effect = 1
- variance depends on h^2

200 Simulations: Effect



200 Simulations: Locus



Basic Idea of Likelihood Use

- build likelihood in steps
  - build from trait & genotypes at locus
  - likelihood for individual *i*
  - log likelihood over individuals
- maximize likelihood (interval mapping)
  - EM method (Lander & Botstein 1989)
  - MCMC method (Guo & Thompson 1994)
- study whole likelihood as posterior (Bayesian)
  - analytical methods (e.g. Carlin & Louis 1998)
  - MCMC method (Satagopan et al 1996)

Studying the Likelihood

- maximize (*IM)
  - find the peak
  - avoid local maxima
  - profile LOD
    - across locus
    - max for effects
- sample (Bayes)
  - get whole curve
  - summarize later
  - posterior
    - locus & effects together
- EM method
  - always go up
  - steepest ascent
- MCMC method
  - jump around
  - go up if you can
  - sometimes go down
  - cool down to find peak
    - simulated annealing
    - simulated tempering

EM-MCMC duality

- EM approach can be redone with MCMC
  - EM estimates & maximizes
  - MCMC draws random samples
  - both can address same problem
- sometimes EM is hard (impossible) to use
- MCMC is tool of “last resort”
  - use exact methods if you can
  - try other approximate methods
  - be clever!
  - very handy for hard problems in genetics

Part III: MCMC Sampling

- Study the Bayesian Posterior
  - use Markov chain to sample
    - Markov chain Monte Carlo
    - Gibbs sampler for effects
    - Metropolis-Hastings for loci
- *Brassica* data on days to flowering

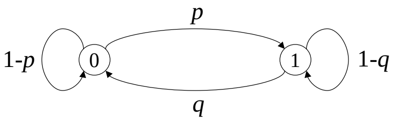
How to Proceed?

- want to study π(parameters|data)
- run Markov chain with stable pattern π()
- study properties of Markov chain to learn about posterior π(parameters|data)
  - Markov chain Monte Carlo
- summarize results in graphical form
- diagnostics

Markov chain idea

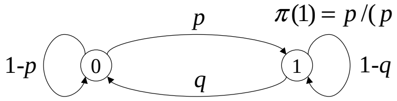
- future given present is independent of past
- update chain based on current value
  - can make chain arbitrarily complicated
  - chain converges to stable pattern π() we wish to study

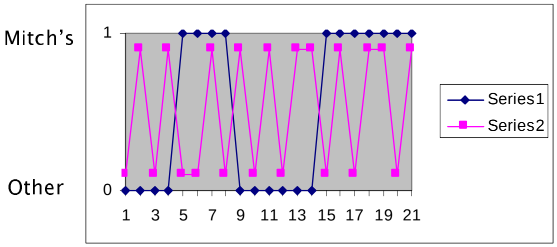
$$\pi(1) = p/(p+q)$$



Markov chain idea

$$\pi(1) = p/(p+q)$$

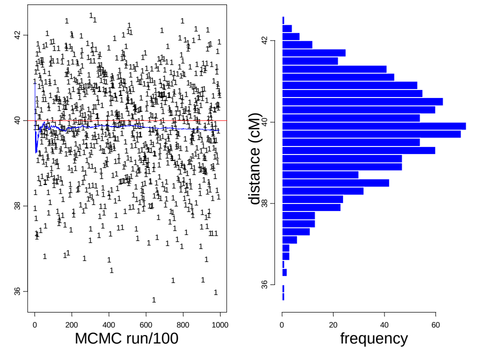




Markov chain Monte Carlo

- 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
  - update some parameters given data and others
  - can fudge on "right" (importance sampling)
  - next step depends only on current estimates
- nice Markov chains have nice properties
  - sample summaries make sense
  - consider almost as random sample from distribution

MCMC Run for 1 locus Data



Why not Ordinary Monte Carlo?

- independent samples of joint distribution
- chaining (or peeling) of effects
- requires numerical integration
  - possible analytically here
  - very messy in general

$$\pi(\mu,b^*,\sigma^2 \mid \mathbf{y},\mathbf{x}^*) =$$
$$\pi(\sigma^2 \mid \mathbf{y},\mathbf{x}^*;\mu,b^*)\times\pi(b^* \mid \mathbf{y},\mathbf{x}^*;\mu)\times\pi(\mu \mid \mathbf{y},\mathbf{x}^*)$$

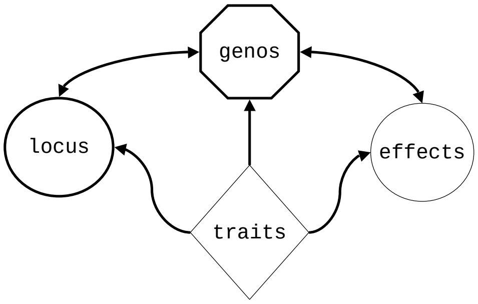
$$\pi(\mu \mid \mathbf{y},\mathbf{x}^*) = E_{(b^*,\sigma^2)}\left(\pi(\mu,b^*,\sigma^2 \mid \mathbf{y},\mathbf{x}^*)\right)$$

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 one (or several) components at a time
  - update effects given genotypes & traits
  - update locus given genotypes & traits
  - update genotypes give locus & effects

$$\theta = (\mathbf{x}^*;\mu,b^*,\sigma^2;\lambda) \sim \pi(\mathbf{x}^*;\mu,b^*,\sigma^2;\lambda \mid \mathbf{y})$$
$$\theta_1 \to \theta_2 \to \cdots \to \theta_N$$

Markov chain updates



Gibbs Sampler for effects

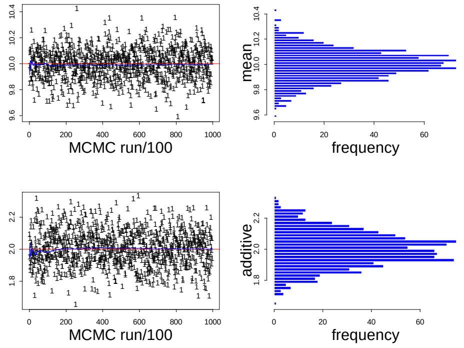
- set up Markov chain around posterior for effects
- sample from posterior by sampling from full conditionals
  - conditional posterior of each parameter given the other
  - update parameter by sampling full conditional

update mean $\pi(\mu \mid \mathbf{y},\mathbf{x}^*;b^*,\sigma^2) = \pi(\mu)\pi(\mathbf{y} \mid \mathbf{x}^*;\mu,b^*,\sigma^2)/c$

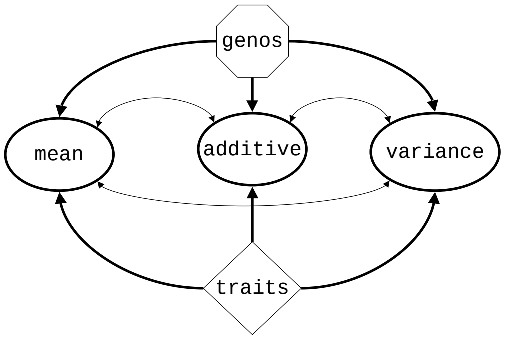
update additive $\pi(b^* \mid \mathbf{y},\mathbf{x}^*;\mu,\sigma^2) = \pi(b^*)\pi(\mathbf{y} \mid \mathbf{x}^*;\mu,b^*,\sigma^2)/c$

update variance $\pi(\sigma^2 \mid \mathbf{y},\mathbf{x}^*;\mu,b^*) = \pi(\sigma^2)\pi(\mathbf{y} \mid \mathbf{x}^*;\mu,b^*,\sigma^2)/c$

MCMC run of mean & effect



Markov chain details



Full Conditional for genos

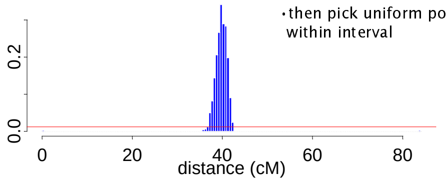
- full conditional for genotype depends on
  - `effects` via `trait` model
  - `locus` via recombination model
- can explicitly decompose by individual *j*
  - binomial (or trinomial) probability

$$x_j^* = -1, 0, or\, 1$$

$$P_j = \pi(x_j^* \mid y_j; \mu, b^*, \sigma^2; \lambda) = \frac{\pi(y_j \mid x_j^*; \mu, b^*, \sigma^2)\pi(x_j^* \mid \lambda)}{\sum_{x=-1,0,1} \pi(y_j \mid x; \mu, b^*, \sigma^2)\pi(x \mid \lambda)}$$

Prior for `locus`

- prior information from other studies
  - concentrate on credible regions
  - use posterior of previous study as new prior



- no prior information on locus
  - uniform prior over genome
  - use framework map
    - choose interval proportional to length
    - then pick uniform position within interval

Full Conditional for `locus`

- cannot easily sample from `locus` full conditional
- cannot explicitly determine full conditional
  - difficult to normalize
  - need to consider all possible genotypes over entire `map`
- Gibbs sampler will not work
  - but can get something proportional ...

$$\pi(\lambda \mid \mathbf{y}, \mathbf{x}^*; \mu, b^*, \sigma^2) = \pi(\lambda \mid \mathbf{x}^*)$$
$$= \pi(\lambda) \prod_{j=1}^{n} \pi(x_j^* \mid \lambda) / c$$

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

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

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”

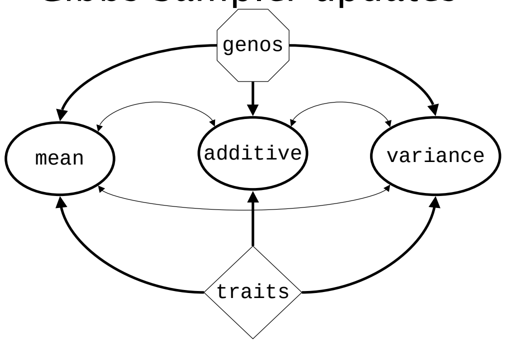
Part IV: MCMC Details

- quick review of trait model
  - single & multiple QTL
  - details of Gibbs sampler full conditionals
  - vector notation

Quick Review of `trait` Model

- single QTL details of Gibbs sampler
  - normal priors & likelihoods
    - mean, additive effects
  - inverse gamma prior for variance
    - or inverse chi-square
  - vague priors lead to usual estimates as posterior means
- multiple QTL `trait` model
  - model with vector notation

Gibbs Sampler updates



Full Conditional for `mean`

- normal prior with large variance τ^2

$$\pi(\mu \mid \mathbf{y},\mathbf{x}^*;b^*,\sigma^2) \propto \phi\left(\frac{\mu-\eta}{\tau}\right)\prod_{j=1}^{n}\phi\left(\frac{y_j-\mu-b^*x_j^*}{\sigma}\right)$$

- leads to normal posterior
- posterior mean

$$E(\mu \mid \mathbf{y},\mathbf{x}^*;b^*,\sigma^2)=\frac{\sum_{j=1}^{n}(y_j-b^*x_j^*)+\eta\frac{\sigma^2}{\tau^2}}{n+\frac{\sigma^2}{\tau^2}}\approx\frac{\sum_{j=1}^{n}(y_j-b^*x_j^*)}{n}$$

- posterior variance

$$V(\mu \mid \mathbf{y},\mathbf{x}^*;b^*,\sigma^2)=\frac{\sigma^2}{n+\frac{\sigma^2}{\tau^2}}\approx\frac{\sigma^2}{n}$$

Full Conditional for `additive` Effect

- normal prior with large variance τ^2

$$\pi(b^* \mid \mathbf{y},\mathbf{x}^*;\mu,\sigma^2) \propto \phi\left(\frac{b^*}{\tau}\right)\prod_{j=1}^{n}\phi\left(\frac{y_j-\mu-b^*x_j^*}{\sigma}\right)$$

- leads to normal posterior
- posterior mean

$$E(b^* \mid \mathbf{y},\mathbf{x}^*;\mu,\sigma^2)=\frac{\sum_{j=1}^{n}x_j^*(y_j-\mu)}{\sum_{j=1}^{n}(x_j^*)^2+\frac{\sigma^2}{\tau^2}}\approx\frac{\sum_{j=1}^{n}x_j^*(y_j-\mu)}{\sum_{j=1}^{n}(x_j^*)^2}$$

- posterior variance

$$V(b^* \mid \mathbf{y},\mathbf{x}^*;\mu,\sigma^2)=\frac{\sigma^2}{\sum_{j=1}^{n}(x_j^*)^2+\frac{\sigma^2}{\tau^2}}\approx\frac{\sigma^2}{\sum_{j=1}^{n}(x_j^*)^2}$$

Full Conditional for `variance`

- inverse gamma prior with large *v/a*

$$\sigma^2 \sim Inv\Gamma(a,v) \propto (\sigma^2)^{-(a+1)}e^{-v/\sigma^2}$$

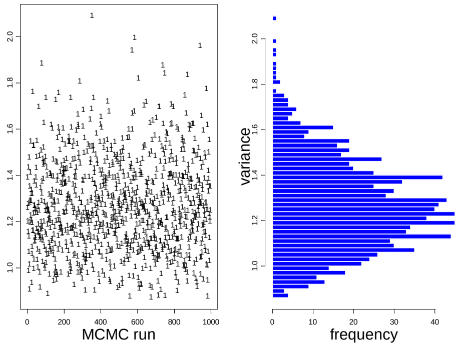
- posterior distribution

$$\sigma^2 \mid \mathbf{y},\mathbf{x}^*;\mu,b^* \sim Inv\Gamma\left(a+\tfrac{n}{2},v+\tfrac{n}{2}\hat{\sigma}^2\right)$$

- posterior mean

$$E(\sigma^2 \mid \mathbf{y},\mathbf{x}^*;\mu,b^*)=\frac{v+\frac{n}{2}\hat{\sigma}^2}{a+\frac{n}{2}-1}\approx\hat{\sigma}^2=\frac{\sum_{j=1}^{n}(y_j-\mu-b^*x_j^*)^2}{n}$$

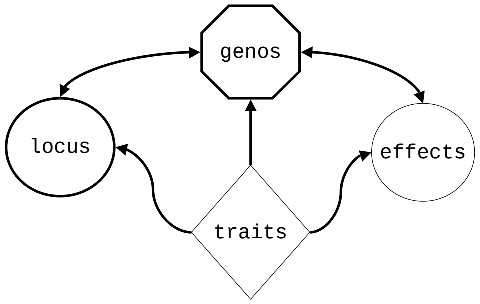
MCMC run for `variance`



Alternative for Variance: use Inverse Chi-square

- inverse chi-square prior with large *d,v* $\sigma^2 \sim Inv\chi^2(d,v) = \frac{vd}{\chi_d^2}, \text{or} \frac{vd}{\sigma^2} \sim \chi_d^2$
- posterior distribution $\sigma^2 \mid \mathbf{y},\mathbf{x}^*;\mu,b^* \sim Inv\chi^2\left(d+n, \frac{vd+\sum_{j=1}^{n}(y_j-\mu-b^*x_j^*)^2}{d+n} \right)$

Markov chain updates



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

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

Full Conditional for `genos`

- full conditional for genotype depends on
  - `effects` via `trait` model
  - `locus` via recombination model
- can explicitly decompose by individual *j*
  - binomial (or trinomial) probability

$$x_j^* = -1, 0, or\, 1$$

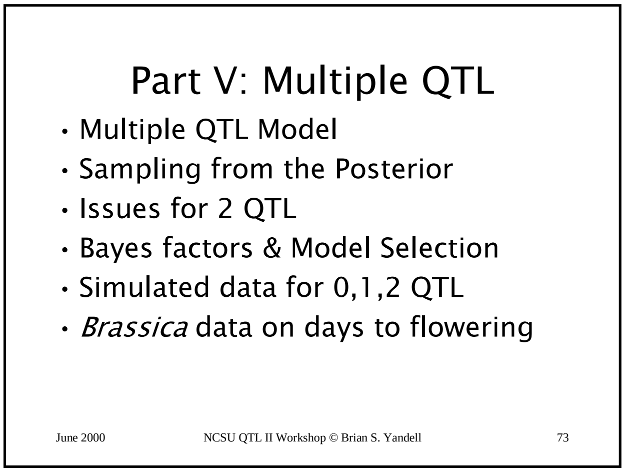
$$\pi(x_j^* \mid y_j;\mu,b^*,\sigma^2;\lambda) = \frac{\pi(y_j \mid x_j^*;\mu,b^*,\sigma^2)\pi(x_j^* \mid \lambda)}{\sum_{x=-1,0,1}\pi(y_j \mid x;\mu,b^*,\sigma^2)\pi(x \mid \lambda)}$$

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 *j*
  - binomial (or trinomial) probability

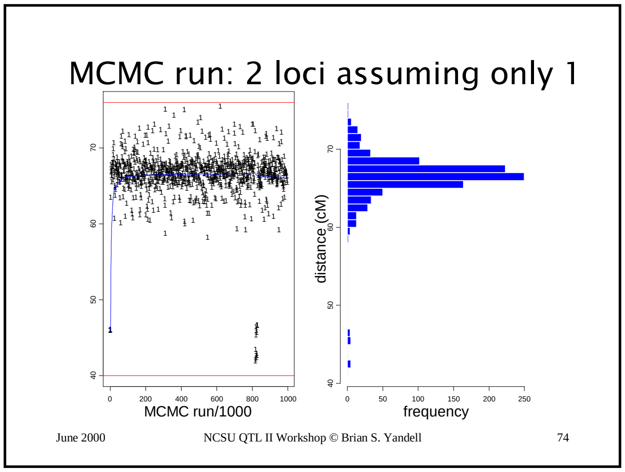
$$M_{kj} = -1, 0, or\, 1$$

$$\pi(M_{kj} \mid x_j^*, y_j;\mu,b^*,\sigma^2;\lambda;\mathbf{M}_j) = \pi(M_{kj} \mid x_j^*;\mathbf{M}_j)$$

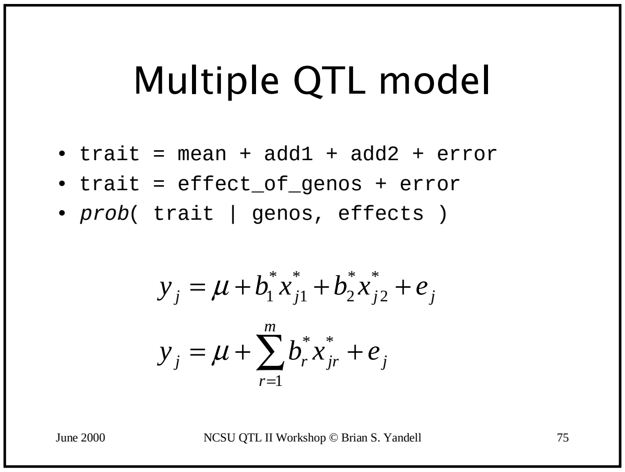


Part V: Multiple QTL

- Multiple QTL Model
- Sampling from the Posterior
- Issues for 2 QTL
- Bayes factors & Model Selection
- Simulated data for 0,1,2 QTL
- *Brassica* data on days to flowering



MCMC run: 2 loci assuming only 1

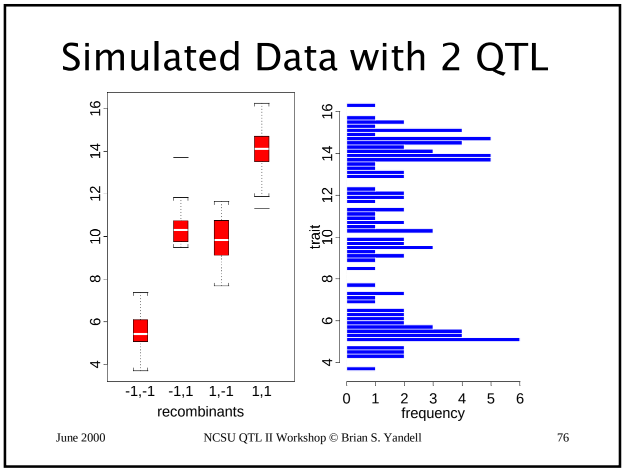


Multiple QTL model

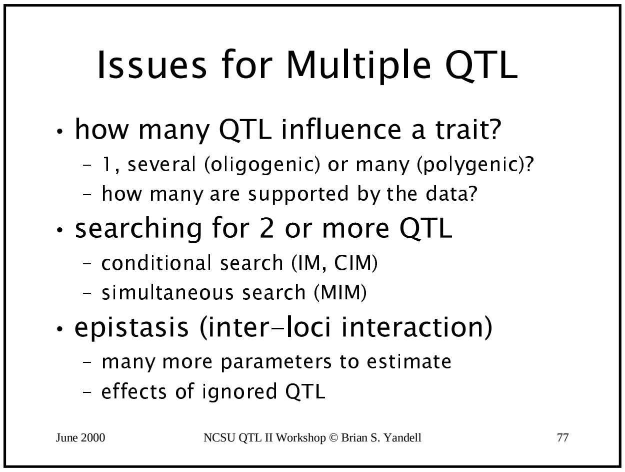
- trait = mean + add1 + add2 + error
- trait = effect_of_genos + error
- *prob*(trait | genos, effects)

$$y_j = \mu + b_1^* x_{j1}^* + b_2^* x_{j2}^* + e_j$$

$$y_j = \mu + \sum_{r=1}^{m} b_r^* x_{jr}^* + e_j$$

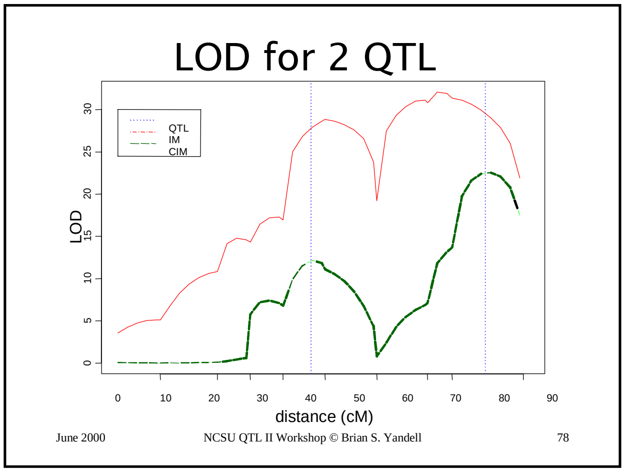


Simulated Data with 2 QTL



Issues for Multiple QTL

- how many QTL influence a trait?
  - 1, several (oligogenic) or many (polygenic)?
  - how many are supported by the data?
- searching for 2 or more QTL
  - conditional search (IM, CIM)
  - simultaneous search (MIM)
- epistasis (inter-loci interaction)
  - many more parameters to estimate
  - effects of ignored QTL



LOD for 2 QTL

Interval Mapping Approach

- interval mapping (IM)
  - scan genome for 1 QTL
- composite interval mapping (CIM)
  - scan for 1 QTL while adjusting for others
  - use markers as surrogates for other QTL
- multiple interval mapping (MIM)
  - search for multiple QTL

Multiple QTL model

- `trait = mean + add1 + add2 + error`
- `trait = effect_of_genos + error`
- *`prob`*`( trait | genos, effects )`

$$y_j = \mu + b_1^* x_{j1}^* + b_2^* x_{j2}^* + e_j$$

$$y_j = \mu + \sum_{r=1}^{m} b_r^* x_{jr}^* + e_j$$

Vector Notation for QTLs

- inner product for sum
- condense notation

$$\sum_{r=1}^{m} b_r^* x_{jr}^* = < \mathbf{b}^*, \mathbf{x}_j^* >$$

$$\mathbf{b}^* = \begin{pmatrix} b_1^* \\ \vdots \\ b_m^* \end{pmatrix}, \mathbf{x}_j^* = \begin{pmatrix} x_{j1}^* \\ \vdots \\ x_{jm}^* \end{pmatrix}, \mathbf{X}^* = \left(\mathbf{x}_1^*, \cdots, \mathbf{x}_n^*\right)$$

Multiple `loci`

- vector of `loci` across linkage map
- careful bookkeeping during update
  - identifiability & bump hunting
  - possibility of two loci in one marker interval
- ordered `loci` are sufficient

$$\pi(\Lambda \mid \mathbf{X}^*) = \prod_{r=1}^{m} \pi(\lambda_r \mid \mathbf{X}^*), \Lambda = (\lambda_1, \cdots, \lambda_m)$$

$$\pi(\lambda_r \mid \mathbf{X}^*) \propto \pi(\lambda_r) \prod_{j=1}^{n} \pi(x_{jr}^* \mid \lambda_r)$$

Posterior: Multiple QTLs

- posterior = likelihood * prior / constant
- posterior(paramaters | data)

  `prob``( genos, effects, loci | traits, map )`

$$\pi(\mathbf{X}^*; \mu, b^*, \sigma^2; \Lambda \mid \mathbf{y})$$

is proportional to

$$\pi(\mu)\pi(\sigma^2) \prod_{r=1}^{m} \left(\pi(b_r^*) \pi(\lambda_r) \prod_{j=1}^{n} \pi(x_{jr}^* \mid \lambda_r) \right)$$

$$\times \prod_{j=1}^{n} \pi(y_j \mid \mathbf{x}_j^*; \mu, \mathbf{b}^*, \sigma^2)$$

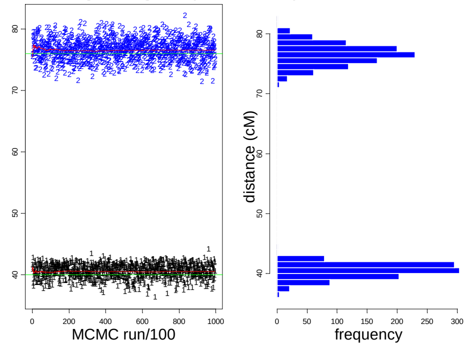
MCMC for Multiple QTLs

- construct Markov chain around posterior
- update one (or several) components at a time
  - update `effects` given genotypes & `traits`
  - update `loci` given genotypes & `traits`
  - update genotypes give `loci` & `effects`
- update all terms for each `locus` at one time?
  - open questions of efficient mixing

$$\theta = (\mathbf{X}^*; \mu, \mathbf{b}^*, \sigma^2; \Lambda) \sim \pi(\mathbf{X}^*; \mu, \mathbf{b}^*, \sigma^2; \Lambda \mid \mathbf{y})$$

$$\theta_1 \to \theta_2 \to \cdots \to \theta_N$$

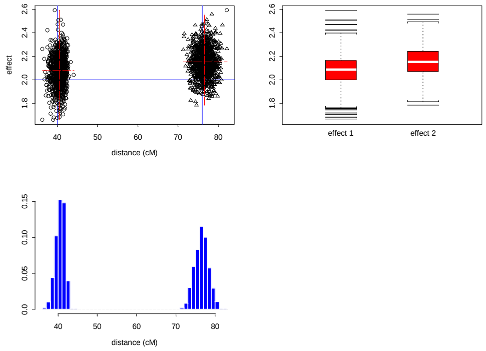
MCMC run with 2 loci



Bayesian Approach

- simultaneous search for multiple QTL
- use Bayesian paradigm
  - easy to consider joint distributions
  - easy to modify later for other types of data
    - counts, proportions, etc.
  - employ MCMC to estimate posterior dist
- study estimates of loci & effects
- use Bayes factors for model selection
  - number of QTL

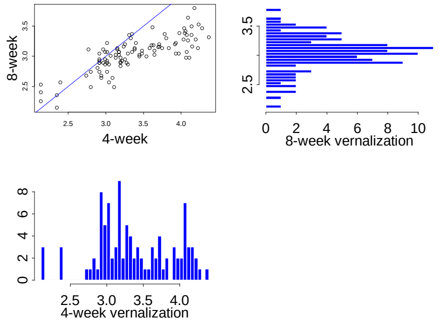
Effects for 2 Simulated QTL



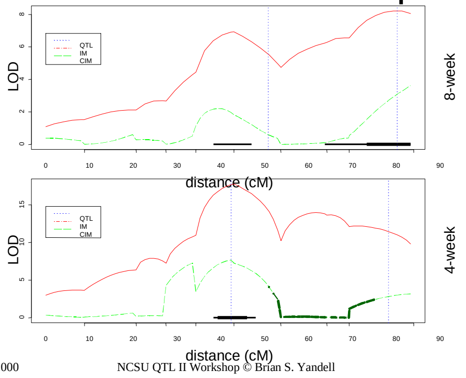
Brassica napus Data

- 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



Brassica Data LOD Maps



4-week vs 8-week vernalization

4-week vernalization

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

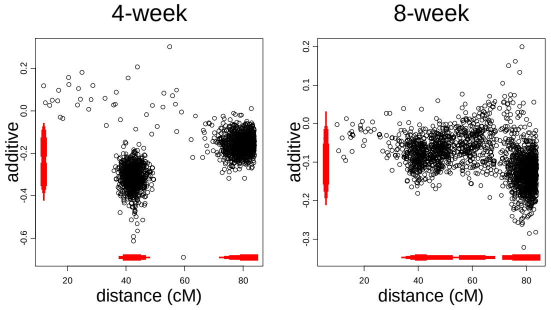
cM	add
40	.30
80	.16

8-week vernalization

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

cM	add
40	.06
80	.13

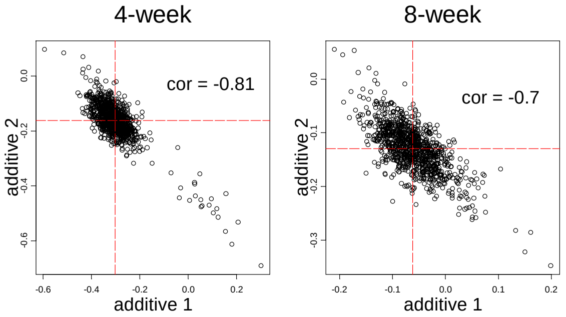
Brassica Credible Regions



Collinearity of QTLs

- multiple QT genotypes are correlated
  - QTL linked on same chromosome
  - difficult to distinguish if close
- estimates of QT `effects` are correlated
  - poor identifiability of `effects` parameters
  - correlations give clue of how much to trust
- which QTL to go after in breeding?
  - largest `effect`?
  - may be biased by nearby QTL

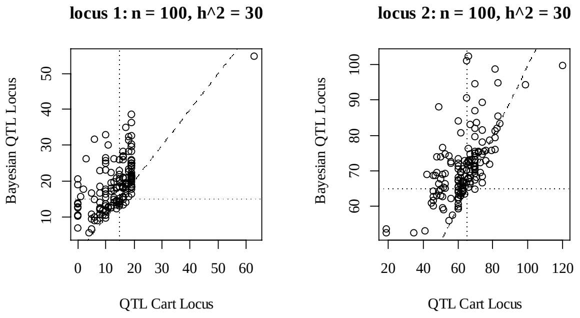
Brassica `effect` Correlations



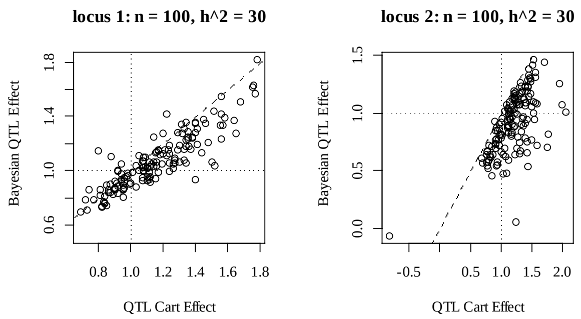
Simulation Study

- 2 linked QTL
- QTL Cart vs. Bayesian QTL estimates
  - locus: 15, 65cM
  - effect: 1, 1
- n = 100, h^2 = 30
- also considered
  - n = 200, h^2 = 25, 30, 40

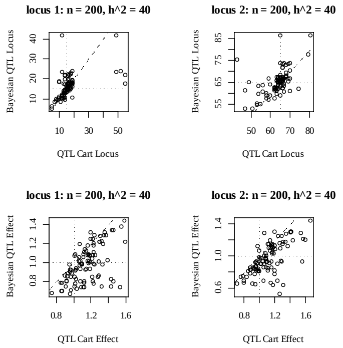
2 QTL: Loci Estimates



2 QTL: Effect Estimates



2 QTL: Loci & Effects



Bayes Factors

Which model (1 or 2 or 3 QTLs?) has higher probability of supporting the data?

- ratio of posterior odds to prior odds
- ratio of model likelihoods

$$B_{12} = \frac{\pi(\text{model}_1 \mid \mathbf{y})/\pi(\text{model}_2 \mid \mathbf{y})}{\pi(\text{model}_1)/\pi(\text{model}_2)} = \frac{\pi(\mathbf{y} \mid \text{model}_1)}{\pi(\mathbf{y} \mid \text{model}_2)}$$

BF(1:2)	2log(BF)	evidence for 1st
< 1	< 0	negative
1 to 3	0 to 2	negligible
3 to 12	2 to 5	positive
12 to 150	5 to 10	strong
> 150	> 10	very strong

Bayes Factors & *LR*

- equivalent to *LR* statistic when
  - comparing two nested models
  - simple hypotheses (e.g. 1 vs 2 QTL)
- Bayes Information Criteria (BIC) in general
  - Schwartz introduced for model selection
  - penalty for different number of parameters *p*

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

Model Determination using Bayes Factors

- pick most plausible model
  - histogram for range of models
  - posterior distribution of models
  - use Bayes theorem
  - often assume flat prior across models
- posterior distribution of number of QTLs

Brassica Bayes Factors

- compare models for 1, 2, 3 QTL
- Bayes factor and $-2\log(LR)$
- large value favors first model
- 8-week vernalization only here

i vs. *j*	Bayes Factor	-log(*LR)*
2 vs. 1	2.49	7.82
3 vs. 1	.005	7.41
3 vs. 2	.002	4.17

Computing Bayes Factors

- arithmetic mean
  - using samples from prior
  - mean across Monte Carlo or MCMC runs
  - can be inefficient if prior differs from posterior

$$\pi(\mathbf{y}\,|\,\text{model}_k) = \int \pi(\mathbf{y}\,|\,\theta_k;\text{model}_k)\pi(\theta_k\,|\,\text{model}_k)d\theta_k$$

- harmonic mean
  - using samples from posterior
  - more efficient but less stable
  - careful choice of weight $h()$ close to posterior

$$\hat{\pi}(\mathbf{y}\,|\,\text{model}_k) = G\left[\sum_{g=1}^{G}\frac{h(\theta_k)}{\pi(\mathbf{y}\,|\,\theta_k;\text{model}_k)\pi(\theta_k\,|\,\text{model}_k)}\right]^{-1}$$

Part VI: How many QTLs?

- Reversible Jump MCMC
  - basic idea of Green(1995)
  - model selection in regression
- how many QTLs?
  - number of QTL is random
  - estimate the number *m*
- RJ-MCMC vs. Bayes factors
- other similar ideas

Jumping the Number of QTL

- model changes with number of QTL
  - almost analogous to stepwise regression
  - use reversible jump MCMC to change number
    - book keeping helps in comparing models
    - change of variables between models
- prior on number of QTL
  - uniform over some range
  - Poisson with prior mean

$$\pi(m\,|\,\ell) = \frac{\ell^m e^{-\ell}}{m!}$$

Posterior: Number of QTL

- posterior = likelihood * prior / constant
- posterior(paramaters | data)

```
*prob*( genos, effects, loci, *m* | traits, map )
```

$$\pi(\mathbf{X}^*;\mu,b^*,\sigma^2;\Lambda,m\,|\,\mathbf{y})$$

is proportional to

$$\prod_{j=1}^{n}\pi(y_j\,|\,\mathbf{x}_j^*;\mu,\mathbf{b}^*,\sigma^2;m)\times$$

$$\pi(m)\pi(\mu)\pi(\sigma^2)\prod_{r=1}^{m}\left(\pi(b_r^*)\pi(\lambda_r)\prod_{j=1}^{n}\pi(x_{jr}^*\,|\,\lambda_r)\right)$$

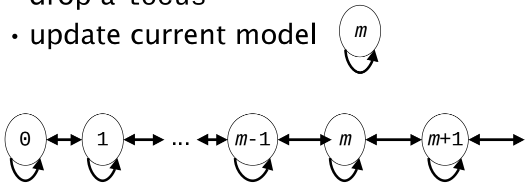
Reversible Jump Choices

action step: draw one of three choices

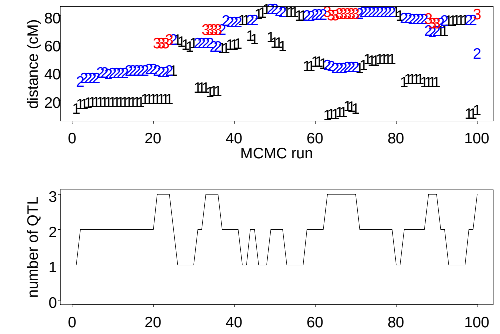
- update step with probability $1\text{-}b(m+1)\text{-}d(m)$
  - update current model
  - `loci`, `effects`, genotypes as before
- add a locus with probability $b(m+1)$
  - propose a new `locus`
  - innovate `effect` and genotypes at new `locus`
  - decide whether to accept the “birth” of new `locus`
- drop a locus with probability $d(m)$
  - pick one of existing `loci` to drop
  - decide whether to accept the “death” of `locus`

Markov chain for number *m*

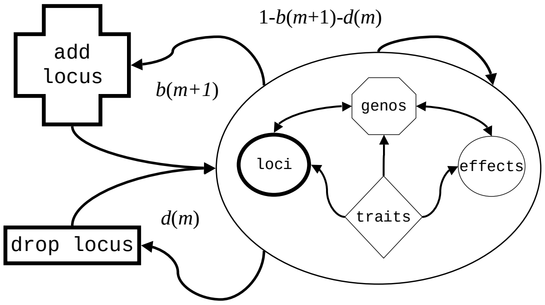
- add a new locus →
- drop a locus ←
- update current model



Jumping QTL number & loci



RJ-MCMC Updates



Propose to Add a locus

- propose a new locus
  - similar proposal to ordinary update $q_b(\lambda) = 1/D$
    - uniform chance over genome
    - easier to avoid interval with another QTL
  - need genotypes at locus & model effect
- innovate effect & genotypes at new locus
  - draw genotypes based on recombination (prior)
    - no dependence on trait model yet
  - draw effect as in Green's reversible jump
    - adjust for collinearity
    - modify other parameters accordingly
- check acceptance …

Propose to Drop a locus

- choose an existing locus $q_d(r;m) = 1/m$
  - equal weight for all loci ?
  - more weight to loci with small effects?
- "drop" effect & genotypes at old locus
  - adjust effects at other loci for collinearity
  - this is reverse jump of Green (1995)
- check acceptance …
  - do not drop locus, effects & genotypes
  - until move is accepted

Acceptance of Reversible Jump

- accept birth of new locus with probability

  $\min(1,A)$
- accept death of old locus with probability

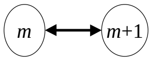
  $\min(1,1/A)$

$$A = \frac{\pi(\theta_{m+1}, m+1 \mid \mathbf{y})}{\pi(\theta_m, m \mid \mathbf{y})} \times \frac{d(m+1)}{b(m)} \frac{q_b(\lambda_{m+1})}{q_d(r; m+1)} \frac{1}{J}$$

$$\theta_m = \left(\mathbf{X}^*; \mu, \mathbf{b}^*, \sigma^2; \Lambda, m\right)$$

Acceptance of Reversible Jump

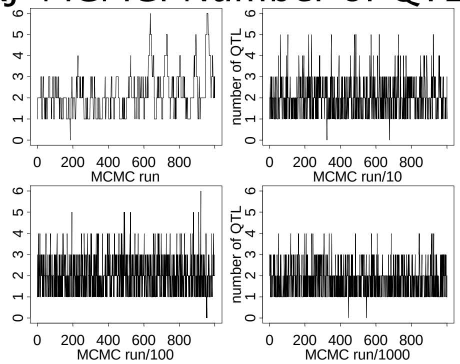
- move probabilities $\dfrac{d(m+1)}{b(m)}$



- birth & death proposals $\dfrac{q_b(\lambda_{m+1})}{q_d(r;m+1)}$

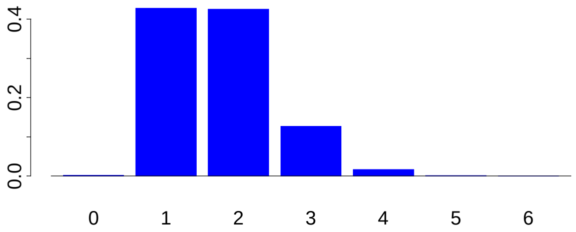
- Jacobian between models $J = \dfrac{\sigma}{s_{r|others}\sqrt{n}}$
  - –fudge factor
  - –see stepwise regression example

RJ-MCMC: Number of QTL



Posterior # QTL for 8-week Data

98% credible region for m: (1,3)
based on 1 million steps
with prior mean of 3



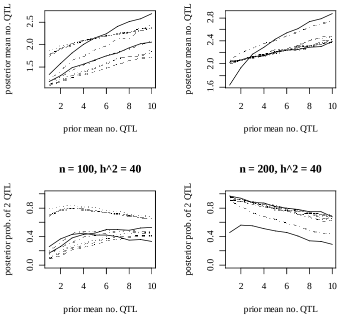
How Good is RJ-MCMC?

- simulations with 0, 1 or 2 QTL
  - strong effects (additive = 2, variance = 1)
  - linked loci 36cM apart
- differences with number of QTL
  - clear differences by actual number
  - works well with 100,000, better with 1M
- effect of Poisson prior mean
  - larger prior mean shifts posterior up
  - but prior does not take over

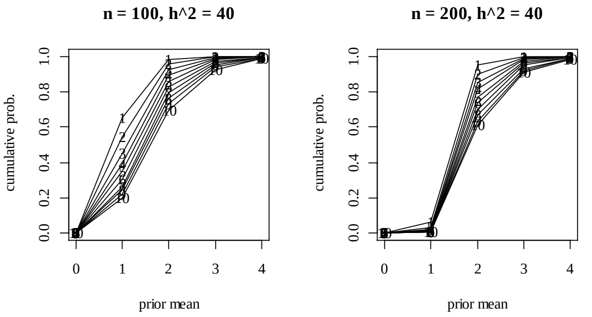
Simulation Study: Prior

- 2 QTL at 15, 65cM
- n = 100, 200; h^2 = 40%
- vary prior mean from 1 to 10 QTL
  - Poisson prior
- 10 independent simulations
- examine posterior mean, probability

Prior on Number of QTL



Prior on Number of QTL



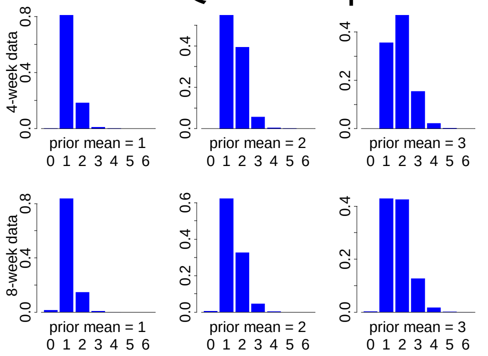
QTL in *Brassica* Data

- 4-week & 8-week vernalization
  - log(days to flower)
  - 105 lines, 10 markers
  - modest effects
  - evidence of 1 or 2 QTL using Bayes factors
- histograms of posterior number of QTL
  - depends somewhat on prior
  - mode is 1 or 2 QTL
- 90% credible sets
  - all include 2 QTL
  - include 1 QTL if prior not huge

Brassica #QTL 90% Credible Sets

	8-week			4-week		
prior	lo	hi	level	lo	hi	level
1	1	2	0.98	1	2	0.99
2	1	2	0.95	1	2	0.94
3	1	3	0.98	1	3	0.98
4	1	3	0.95	1	3	0.93
6	1	4	0.96	1	4	0.94
10	2	5	0.90	2	6	0.97

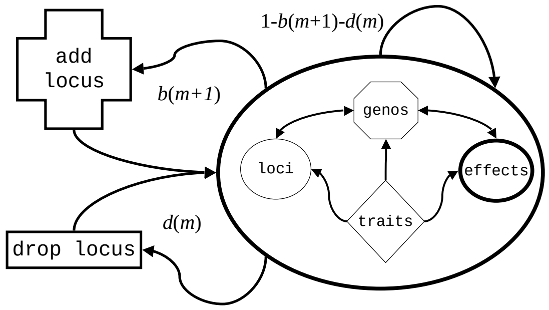
Brassica #QTL Comparison



VII: Reversible Jump Details

- reversible jump MCMC details
  - can update model with *m* QTL
  - have basic idea of jumping models
  - now: careful bookkeeping between models
- RJ-MCMC & Bayes factors
  - Bayes factors from RJ-MCMC chain
  - components of Bayes factors

RJ-MCMC Updates



Reversible Jump Idea

- expand idea of MCMC to compare models
- adjust for parameters in different models
  - augment smaller model with innovations
  - constraints on larger model
- calculus "change of variables" is key
  - add or drop parameter(s)
  - carefully compute the Jacobian
- consider stepwise regression
  - Mallick (1995) & Green (1995)
  - efficient calculation with Hausholder decomposition

Model Selection in Regression

- known regressors (e.g. markers)
  - models with 1 or 2 regressors
- jump between models
  - centering regressors simplifies calculations

$$m = 1 : y_j = \mu + b(x_{j1} - \bar{x}_1) + e_j$$

$$m = 2 : y_j = \mu + b_1(x_{j1} - \bar{x}_1) + b_2(x_{j2} - \bar{x}_2) + e_j$$

Slope Estimate for 1 Regressor

recall least squares estimate of slope
note relation of slope to correlation

$$\hat{b} = \frac{r_{1y}\, s_y}{s_1}, \quad r_{1y} = \frac{\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)(y_j - \bar{y})/n}{s_1\, s_y}$$

$$s_1^2 = \sum_{j=1}^{n}(x_{j1} - \bar{x}_1)^2/n, s_y^2 = \sum_{j=1}^{n}(y_j - \bar{y})^2/n$$

2 Correlated Regressors

slopes adjusted for other regressors

$$\hat{b}_1 = \frac{(r_{1y} - r_{12}r_{2y})s_y}{s_1} = \hat{b} - \frac{r_{2y}s_y}{s_2}c_{21}, \quad c_{21} = \frac{r_{12}s_2}{s_1}$$

$$\hat{b}_2 = \frac{(r_{2y} - r_{12}r_{1y})s_y}{s_2}$$

Gibbs Sampler for Model 1

- mean $\mu \sim \phi\left(\frac{n\bar{y} + \eta\frac{\sigma^2}{\tau^2}}{n + \frac{\sigma^2}{\tau^2}}, \frac{\sigma^2}{n + \frac{\sigma^2}{\tau^2}}\right)$
- slope $b \sim \phi\left(\frac{\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)(y_j - \mu)}{ns_1^2 + \frac{\sigma^2}{\tau^2}}, \frac{\sigma^2}{ns_1^2 + \frac{\sigma^2}{\tau^2}}\right)$
- variance $\sigma^2 \sim Inv\Gamma\left(a + \frac{n}{2}, v + \frac{1}{2}\sum_{j=1}^{n}\left(y_j - \mu - b(x_{j1} - \bar{x}_1)\right)^2\right)$

Gibbs Sampler for Model 2

- mean $\mu \sim \phi\left(\frac{n\bar{y} + \eta\frac{\sigma^2}{\tau^2}}{n + \frac{\sigma^2}{\tau^2}}, \frac{\sigma^2}{n + \frac{\sigma^2}{\tau^2}}\right)$
- slopes $b_2 \sim \phi\left(\frac{\sum_{j=1}^{n}(x_{j2} - \bar{x}_2)(y_j - \mu - b_1(x_{j1} - \bar{x}_1))}{ns_{2|1}^2 + \frac{\sigma^2}{\tau^2}}, \frac{\sigma^2}{ns_{2|1}^2 + \frac{\sigma^2}{\tau^2}}\right)$

$$s_{2|1}^2 = \sum_{j=1}^{n}(x_{j2} - \bar{x}_2 - c_{21}(x_{j1} - \bar{x}_1))^2/n$$

- variance $\sigma^2 \sim Inv\Gamma\left(a + \frac{n}{2}, v + \frac{1}{2}\sum_{j=1}^{n}\left(y_j - \mu - \sum_{k=1}^{2}b_k(x_{jk} - \bar{x}_k)\right)^2\right)$

Updates from 2->1

- drop 2nd regressor
- adjust other regressor

$$b \to b_1 + b_2 c_{21}$$

$$b_2 \to 0$$

Updates from 1->2

- add 2nd slope, adjusting for collinearity
- adjust other slope & variance

$$z \sim \phi(0,1), \qquad J = \frac{\sigma}{s_{2\cdot1}\sqrt{n}}$$

$$b_2 \to \hat{b}_2 + z \times J, \quad \hat{b}_2 = \frac{\sum_{j=1}^{n} (x_{j2} - \bar{x}_2)\left(y_j - \hat{\mu} - \hat{b}_1 (x_{j1} - \bar{x}_1)\right)}{n s_{2\cdot1}^2}$$

$$b_1 \to b - b_2 c_{21} = b - z \times c_{21} J - \hat{b}_2 c_{21}$$

Model Selection in Regression

- known regressors (e.g. markers)
  - models with 1 or 2 regressors
- jump between models
  - augment with new innovation z

m	*parameters*	*innovations*	*transformations*
$1 \to 2$	$(\mu, b, \sigma^2; z)$	$z \sim \phi(0,1)$	$\begin{Bmatrix} b_2 \to \hat{b}_2 + z \times J \\ b_1 \to b - b_2 c_{21} \end{Bmatrix}$
$2 \to 1$	$(\mu, b_1, b_2, \sigma^2)$		$\begin{Bmatrix} b \to b_1 + b_2 c_{21} \\ z \to 0 \end{Bmatrix}$

Change of Variables

- change variables from model 1 to model 2
- calculus issues for integration
  - need to formally account for change of variables
  - infinitessimal steps in integration (*db*)
  - involves partial derivatives (next page)

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{bmatrix} 1 & -c_{21}J & -c_{21} \\ 0 & J & 1 \end{bmatrix} \times \begin{pmatrix} b \\ z \\ \hat{b}_2 \end{pmatrix} = g(b; z \mid \mathbf{y}, \mathbf{x}_1, \mathbf{x}_2)$$

$$\int \pi(b_1, b_2 \mid \mathbf{y}, \mathbf{x}_1, \mathbf{x}_2) db_1 db_2 = \int \pi(b; z \mid \mathbf{y}, \mathbf{x}_1, \mathbf{x}_2) J db dz$$

Jacobian & the Calculus

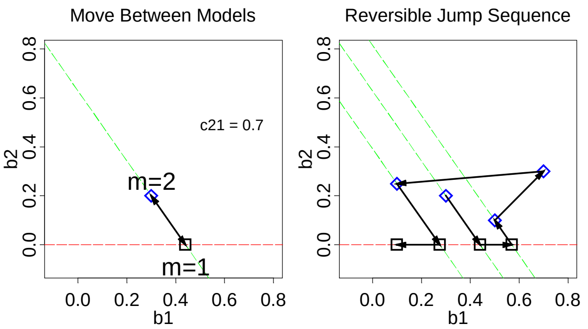
- Jacobian sorts out change of variables
  - <u>careful</u>: easy to mess up here!

$$g(b; z) = (b_1, b_2), \ \frac{\partial g(b; z)}{\partial b \partial z} = \begin{bmatrix} 1 & -c_{21}J \\ 0 & J \end{bmatrix}$$

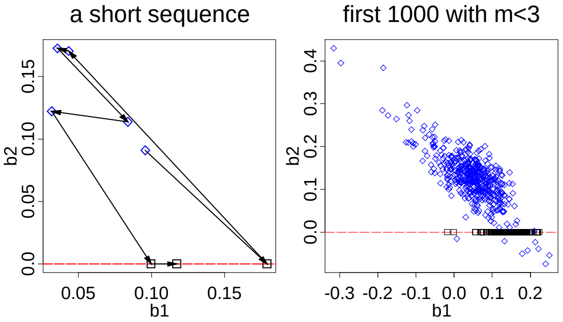
$$\left|\det\left(\begin{bmatrix} 1 & -c_{21}J \\ 0 & J \end{bmatrix}\right)\right| = \left|1 \times J - 0 \times (-c_{21}J)\right| = J$$

$$db_1 db_2 = \left|\det\left(\frac{\partial g(\mu, b, \sigma^2; z)}{\partial b \partial z}\right)\right| db_1 db_2 = J db dz$$

Geometry of Reversible Jump



QT additive Reversible Jump

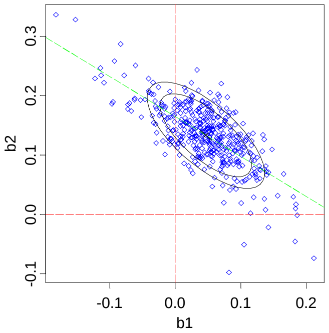

a short sequence

first 1000 with m<3

Credible Set for additive

90% & 95% sets
based on normal

regression line
corresponds to
slope of updates



Efficient Updating of additive

- more computations when $m > 2$
- want to avoid matrix inverses
  - decompose matrix instead
  - solve linear system of equations
- use linear algebra
  - Hausholder (QR) decomposition
  - *LAPACK User's Guide* (1995, 2nd ed) Anderson et al., SIAM.

Hausholder (QR) Decomposition

- decomposition
  - G is upper triangular
  - F is orthogonal

$$\mathbf{X} = \mathbf{F}\mathbf{G} = [\mathbf{F}_1 : \mathbf{F}_2]\begin{bmatrix}\mathbf{G}_1 \\ \mathbf{0}\end{bmatrix} = \mathbf{F}_1\mathbf{G}_1$$

- orthogonality

$$\mathbf{F}^T\mathbf{F} = \begin{bmatrix}\mathbf{F}_1^T\mathbf{F}_1 & \mathbf{F}_1^T\mathbf{F}_2 \\ \mathbf{F}_2^T\mathbf{F}_1 & \mathbf{F}_2^T\mathbf{F}_2\end{bmatrix} = \mathbf{I}_n$$

$$\mathbf{F}_1^T\mathbf{F}_1 = \mathbf{I}_m, \mathbf{F}_2^T\mathbf{F}_2 = \mathbf{I}_{n-m}$$

$$\mathbf{F}_1^T\mathbf{F}_2 = \mathbf{0}, \mathbf{F}_2^T\mathbf{F}_1 = \mathbf{0}$$

- design matrix

$$\mathbf{F}_2^T\mathbf{X} = \mathbf{0}$$

QR & Regression

- model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

- error piece

$$\mathbf{F}_2^T\mathbf{y} = \mathbf{F}_2^T\mathbf{X}\mathbf{b} + \mathbf{F}_2^T\mathbf{e} = \mathbf{F}_2^T\mathbf{e}$$

$$\mathbf{y}^T\mathbf{F}_2\mathbf{F}_2^T\mathbf{y} = SS_{ERROR}$$

- model piece

$$\mathbf{F}_1^T\mathbf{y} = \mathbf{F}_1^T\mathbf{X}\mathbf{b} + \mathbf{F}_1^T\mathbf{e} = \mathbf{G}_1\mathbf{b} + \mathbf{F}_1^T\mathbf{e}$$

$$\mathbf{y}^T\mathbf{F}_1\mathbf{F}_1^T\mathbf{y} = SS_{MODEL}$$

- estimators

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{G}_1^{-1}\mathbf{F}_1^T\mathbf{y} = \frac{r_{1y}s_y}{r_1}$$

Absorbing Old Model

- *old* model
  - *m* regressors
  - QR decomposition

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{old} + \mathbf{e}$$

$$\mathbf{X} = \mathbf{F}\mathbf{G} = \mathbf{F}_1\mathbf{G}_1$$

- *new* model
  - *m*+1 regressor
  - use QR to absorb old model

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{old} + \mathbf{x}_{m+1}b_{m+1} + \mathbf{e}$$

$$\mathbf{F}_2^T\mathbf{y} = \mathbf{F}_2^T\mathbf{x}_{m+1}b_{m+1} + \mathbf{F}_2^T\mathbf{e}$$

Adjusted Slope Estimators

- *old* slopes
  - note m=1 case

$$\hat{\mathbf{b}}_{old} = \mathbf{G}_1^{-1}\mathbf{F}_1^T\mathbf{y} = \frac{r_{1y}s_y}{s_1}$$

- added slope
  - note sum of squares

$$\hat{b}_{m+1} = V^{-1}\mathbf{x}_{m+1}^T\mathbf{F}_2\mathbf{F}_2^T\mathbf{y} = \frac{(r_{2y} - r_{12}r_{1y})s_y}{s_2}$$

$$V = \mathbf{x}_{m+1}^T\mathbf{F}_2\mathbf{F}_2^T\mathbf{x}_{m+1} = ns_{2|1}^2$$

- variance
  - note Jacobian

$$V(\hat{b}_{m+1}) = \sigma^2 / V = J^2$$

- *new* slopes

$$\hat{\mathbf{b}}_{new} = \mathbf{G}_1^{-1}\mathbf{F}_1^T(\mathbf{y} - \mathbf{x}_{m+1}\hat{b}_{m+1})$$

$$\hat{\mathbf{b}}_{new} = \hat{\mathbf{b}}_{old} - \mathbf{G}_1^{-1}\mathbf{F}_1^T\mathbf{x}_{m+1}\hat{b}_{m+1}$$

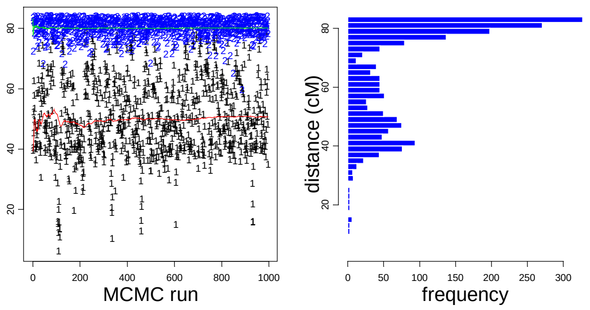
VIII: RJMCMC & Bayes Factors

- RJ-MCMC & Bayes factors
  - Bayes factors from RJ-MCMC chain
  - components of Bayes factors

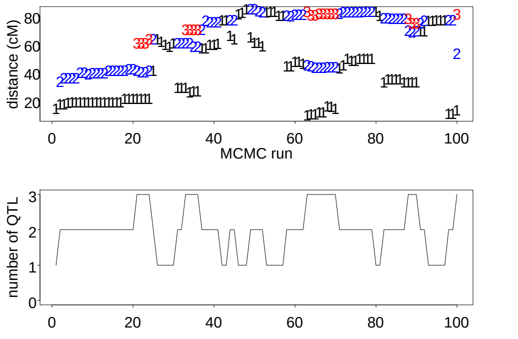
How To Infer `loci`?

- if m is known, use fixed MCMC
  - histogram of loci
  - issue of bump hunting
- combining loci estimates in RJ-MCMC
  - some steps are from wrong model
    - too few loci (bias)
    - too many loci (variance/identifiability)
  - condition on number of loci
    - subsets of Markov chain

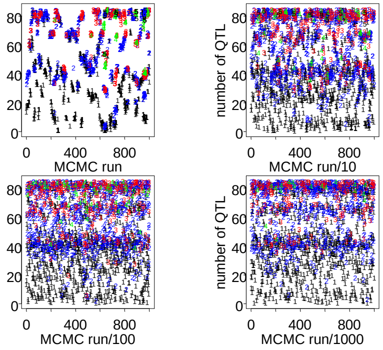
Brassica 8-week Data `locus` MCMC with m=2



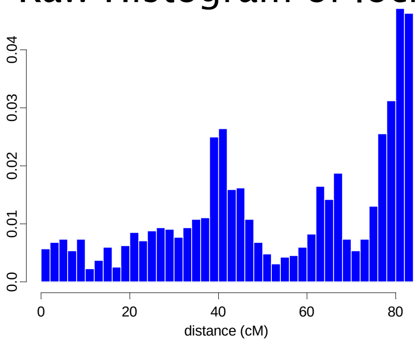
Jumping QTL number & `loci`



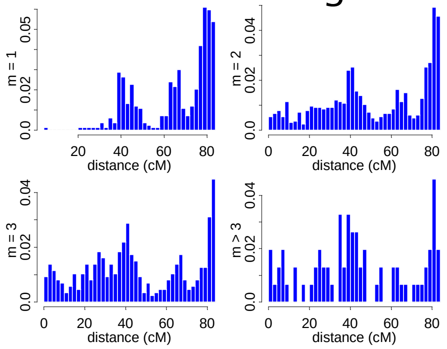
RJ-MCMC `loci` chain



Raw Histogram of loci



Conditional Histograms



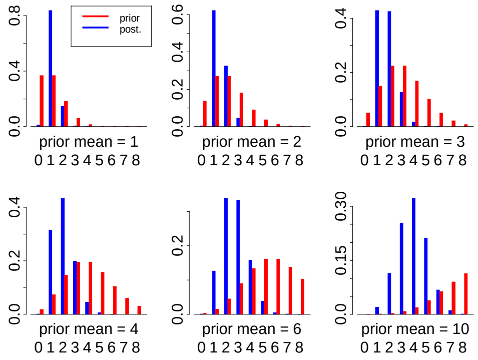
Bayes Factors

- ratio of posterior odds to prior odds
  - RJ-MCMC gives posterior on number of QTL
  - prior is Poisson

$$B_{12} = \frac{\pi(\text{model}_1 \,|\, \mathbf{y}) / \pi(\text{model}_2 \,|\, \mathbf{y})}{\pi(\text{model}_1) / \pi(\text{model}_2)} = \frac{\pi(\mathbf{y} \,|\, \text{model}_1)}{\pi(\mathbf{y} \,|\, \text{model}_2)}$$

BF(1:2)	2log(BF)	evidence for 1st
< 1	< 0	negative
1 to 3	0 to 2	negligible
3 to 12	2 to 5	positive
12 to 150	5 to 10	strong
> 150	> 10	very strong

#QTL for *Brassica* 8-week



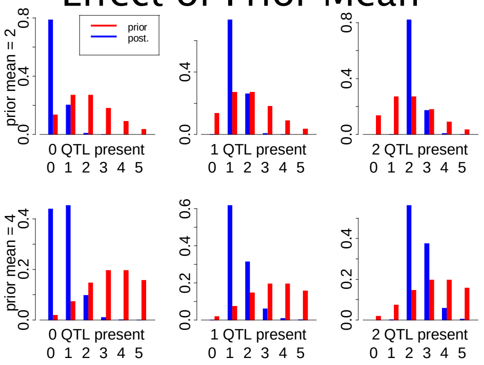
RJ-Bayes Factors (8-week Brassica data)

prior mean	1	2	3	4	6	10
ratio						
1:2	2.87	1.91	1.51	1.45	1.12	0.85
1:3	27.62	9.10	5.06	4.22	2.28	1.28
1:4	1743.29	81.30	28.85	18.51	7.17	2.51
2:3	9.63	4.76	3.35	2.91	2.04	1.5
2:4	608.00	42.51	19.09	12.75	6.41	2.95
3:4	63.13	8.93	5.70	4.39	3.15	1.96

Simulation Study of Prior Effect

- how dramatic is the effect of prior?
- simulations of 0, 1 or 2 QTL
  - QTL have large effect
    - additive = 2, variance = 1
  - 2 QTL spaced 36cM apart
  - sample sized of 105
- RJ-MCMC runs of 100,000

Effect of Prior Mean



Bayes Factor

prior of 2

m ratio	0	1	2
0:1	3.85	0	0
0:2	50.93	0	0
0:3	569.11	0.03	0
1:2	13.22	1.87	0
1:3	147.75	30.09	0
2:3	11.17	16.05	2.38

prior of 4

m ratio	0	1	2
0:1	0.97	0	0
0:2	3.02	0	0
0:3	15.07	0	0
1:2	3.12	1.32	0
1:3	15.54	3.04	0
2:3	4.99	2.58	0.75

Bayes Factors & LODs

- others have tried arithmetic & harmonic mean
- why not geometric mean?
- terms that are averaged are log likelihoods...

$$\hat{\pi}(\mathbf{y}\,|\,\text{model}_k) = \exp\left(\frac{\sum_{g=1}^{G} \log \pi(\mathbf{y}\,|\,\theta_k;\text{model}_k)}{G}\right)$$

$$g = 1,\cdots,G \quad \textit{MCMC runs}$$

Bayesian LOD

- Bayesian “LOD” computed at each step
  - based on *LR* given sampled genotypes and effects
  - can be larger or smaller than profile LOD
  - informal diagnostic of fit
  - combine to for geometric estimates of Bayes factors

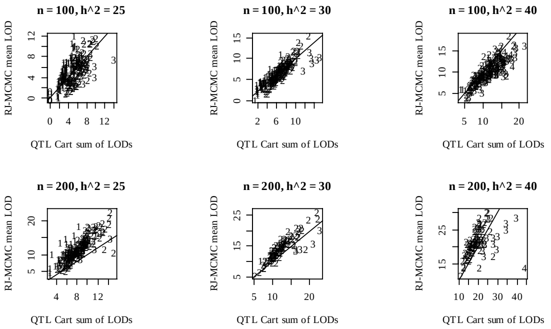
$$LOD(\lambda) = (\log_{10} e)\sum_{j=1}^{n} \ln\left(\frac{\sum_{x=-1,0,1} \pi(y_j\,|\,x;\hat{\mu},\hat{b}^*,\hat{\sigma}^2)\pi(x\,|\,\lambda)}{\pi(y_j\,|\,\hat{\mu},b^*=0,\hat{\hat{\sigma}}^2)}\right)$$

$$BLOD = (\log_{10} e)\sum_{j=1}^{n} \ln\left(\frac{\pi(y_j\,|\,x_j^*;\mu,b^*,\sigma^2)}{\pi(y_j\,|\,\mu,b^*=0,\sigma^2)}\right)$$

Compare LODs

- 200 simulations (only 100 for some)
- n = 100, 200; h^2 = 25, 30, 40%
- 2 QTL at 15, 65cM
- Bayesian vs. CIM-based LODs
  - Bayesian for simultaneous fit
  - classical for sum of CIM LODs at peaks
- plot symbol is number of CIM peaks

Comparing LODs



IX: RJ-MCMC Software

- General MCMC software
  - U Bristol links
    - http://www.stats.bris.ac.uk/MCMC/pages/links.html
  - BUGS (Bayesian inference Using Gibbs Sampling)
    - http://www.mrc-bsu.cam.ac.uk/bugs/
- Our MCMC software for QTLs
  - C code using LAPACK
    - ftp://ftp.stat.wisc.edu/pub/yandell/revjump.tar.gz
  - coming soon:
    - perl preprocessing (to/from QtlCart format)
    - Splus post processing
    - Bayes factor computation

The Art of MCMC

- convergence issues
  - burn-in period & when to stop
- proper mixing of the chain
  - smart proposals & smart updates
- frequentist approach
  - simulated annealing: reaching the peak
  - simulated tempering: heating & cooling the chain
- Bayesian approach
  - influence of priors on posterior
  - Rao-Blackwell smoothing
- bump-hunting for mixtures (e.g. QTL)

Bayes Factor References

- MA Newton & AE Raftery (1994) “Approximate Bayesian inference with the weighted likelihood bootstrap”, *J Royal Statist Soc B* 56: 3-48.
- RE Kass & AE Raftery (1995) “Bayes factors”, *J Amer Statist Assoc* 90: 773-795.
- JM Satagopan, MA Newton & AE Rafter (1999) “On the harmonic mean estimator of marginal probability”, ms in prep, mailto:satago@biosta.mskcc.org.

Reversible Jump MCMC References

- PJ Green (1995) “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination”, *Biometrika* 82: 711-732.
- S Richardson & PJ Green (1997) “On Bayesian analysis of mixture with an unknown of components”, *J Royal Statist Soc B* 59: 731-792.
- BK Mallick (1995) “Bayesian curve estimation by polynomials of random order”, TR 95-19, Math Dept, Imperial College London.
- L Kuo & B Mallick (1996) “Bayesian variable selection for regression models”, *ASA Proc Section on Bayesian Statistical Science*, 170-175.

QTL Reversible Jump MCMC: Inbred Lines

- JM Satagopan & BS Yandell (1996) “Estimating the number of quantitative trait loci via Bayesian model determination”, *Proc JSM Biometrics Section*.
- DA Stephens & RD Fisch (1998) “Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo”, *Biometrics* 54: 1334-1347.
- MJ Sillanpaa & E Arjas (1998) “Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data”, *Genetics* 148: 1373-1388.
- R Waagepetersen & D Sorensen (1999) “Understanding reversible jump MCMC”, mailto:sorensen@inet.uni2.dk.

QTL Reversible Jump MCMC: Pedigrees

- S Heath (1997) “Markov chain Monte Carlo segregation and linkage analysis for oligenic models”, *Am J Hum Genet* 61: 748-760.
- I Hoeschele, P Uimari , FE Grignola, Q Zhang & KM Gage (1997) “Advances in statistical methods to map quantitative trait loci in outbred populations”, *Genetics* 147:1445-1457.
- P Uimari and I Hoeschele (1997) “Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms”, *Genetics* 146: 735-743.
- MJ Sillanpaa & E Arjas (1999) “Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data”, *Genetics* 151, 1605-1619.

Bayes & MCMC References

- CJ Geyer (1992) “Practical Markov chain Monte Carlo”, *Statistical Science* 7: 473-511
- L Tierney (1994) “Markov Chains for exploring posterior distributions”, *The Annals of Statistics* 22: 1701-1728 (with disc:1728-1762).
- A Gelman, J Carlin, H Stern & D Rubin (1995) *Bayesian Data Analysis*, CRC/Chapman & Hall.
- BP Carlin & TA Louis (1996) *Bayes and Empirical Bayes Methods for Data Analysis*, CRC/Chapman & Hall.
- WR Gilks, S Richardson, & DJ Spiegelhalter (Ed 1996) *Markov Chain Monte Carlo in Practice*, CRC/Chapman & Hall.

MCMC Software

- General MCMC software
  - U Bristol links
    - http://www.stats.bris.ac.uk/MCMC/pages/links.html
  - BUGS (Bayesian inference Using Gibbs Sampling)
    - http://www.mrc-bsu.cam.ac.uk/bugs/
- Our MCMC software for QTLs
  - C code using LAPACK
    - ftp://ftp.stat.wisc.edu/pub/yandell/revjump.tar.gz
  - coming soon:
    - perl preprocessing (to/from QtlCart format)
    - Splus post processing
    - Bayes factor computation within QtlCart

QTL References

- D Thomas & V Cortessis (1992) “A Gibbs sampling approach to linkage analysis”, *Hum. Hered.* 42: 63-76.
- I Hoeschele & P vanRanden (1993) “Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge”, *Theor. Appl. Genet.* 85:953-960.
- I Hoeschele & P vanRanden (1993) “Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence”, *Theor. Appl. Genet.* 85:946-952.
- SW Guo & EA Thompson (1994) “Monte Carlo estimation of mixed models for large complex pedigrees”, *Biometrics* 50: 417-432.
- JM Satagopan, BS Yandell, MA Newton & TC Osborn (1996) “A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo”, *Genetics* 144: 805-816.