Efficient and Robust Model Selection for Quantitative Trait Loci Analysis in Inbred Lines

Brian S. Yandell
University of Wisconsin-Madison
www.stat.wisc.edu/~yandell
with Jaya M. Satagopan, Sloan-Kettering Biostatistics,
Fei Zou, UNC Biostatistics
and Patrick J. Gaffney, Lubrizol

NCSU Statistical Genetics, June 2002

Goals

- model selection with one QTL
  - review interval mapping basics
  - extensions of phenotype model
    - how to map non-normal data?
  - brief digression to multiple crosses
  - Bayesian interval mapping
    - how to sample from the posterior?
- model selection over multiple QTL
  - how many QTL are supported by data?
  - how to sample complicated model space?

Main Topics

away from normality

- fewer assumptions
  - semi-parametric
  - non-parametric
- extended phenotypes
- check robustness
- multiple crosses

how many QTL?

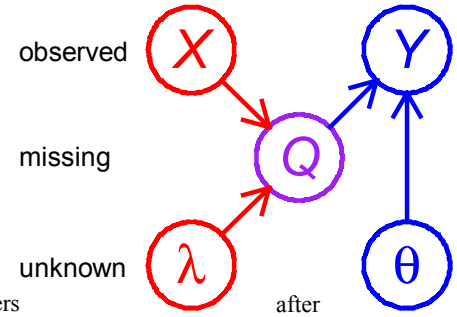
- inferring the number
- sampling all QT loci
- Bayesian analysis
- MCMC methodology
- estimating heritability

Overview

- I: IM basics (review)
- II: Beyond normal data
- III: Bayesian idea
  - Bayes theorem
  - posterior & likelihood
- IV: MCMC samples
  - Monte Carlo idea
  - study posterior
- V: Multiple QTL
- VI: How many QTL?
  - Reversible Jump
  - analog to regression
- VII: RJ-MCMC details
- VII: Model assessment
- IX: References
  - Software
  - Articles
- X: Multiple crosses

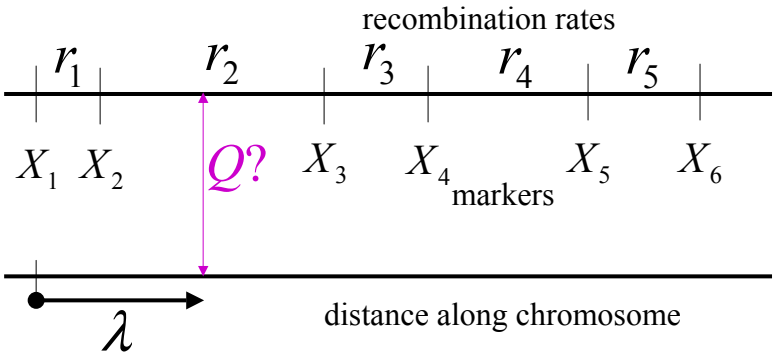
Part I: Interval Mapping Basics

- observed measurements
  - Y = phenotypic trait
  - X = markers & linkage map
    - i = individual index 1,…,n
- missing data
  - missing marker data
  - Q = QT genotypes
    - alleles QQ, Qq, or qq at locus
- unknown quantities
  - λ = QT locus (or loci)
  - θ = phenotype model parameters
- pr($Q|X,\lambda$) recombination model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for Q given X
- pr($Y|Q,\theta$) phenotype model
  - distribution shape (could be assumed normal)
  - unknown parameters θ (could be non-parametric)



after Sen Churchill (2001)

recombination model components

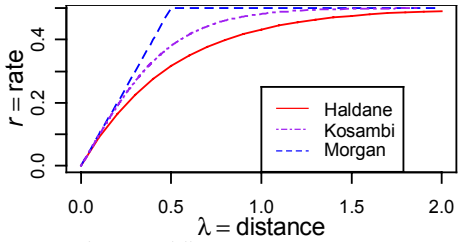


Recombination and Distance

- assume map and marker distances are known
- useful approximation for QTL linkage
  - Haldane map function: no crossover interference
  - independence implies crossover events are Poisson
- all computations consistent in approximation
  - rely on given map with known marker locations
  - 1-to-1 relation of distance to recombination
  - all map functions are approximate anyway

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

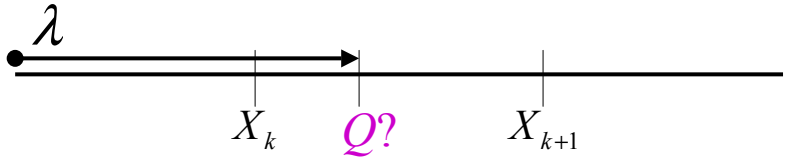
$$\lambda = -\tfrac{1}{2}\log(1-2r)$$



recombination model $\text{pr}(Q|X,\lambda)$

- locus λ is distance along linkage map
  - identifies flanking marker region
- flanking markers provide good approximation
  - map assumed known from earlier study
  - inaccuracy slight using only flanking markers
    - extend to next flanking markers if missing data
  - could consider more complicated relationship
    - but little change in results

$$\text{pr}(Q|X,\lambda) = \text{pr}(\text{geno} \mid \text{map, locus}) \approx \text{pr}(\text{geno} \mid \text{flanking markers, locus})$$



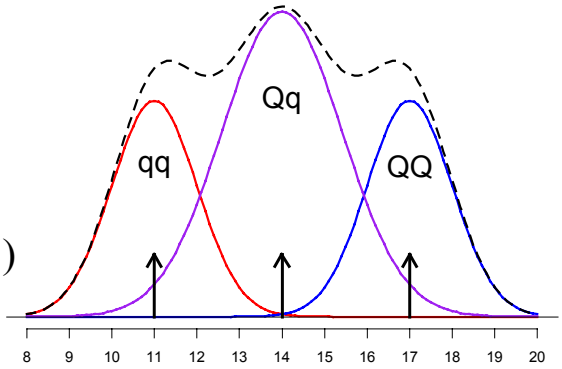
idealized phenotype model

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

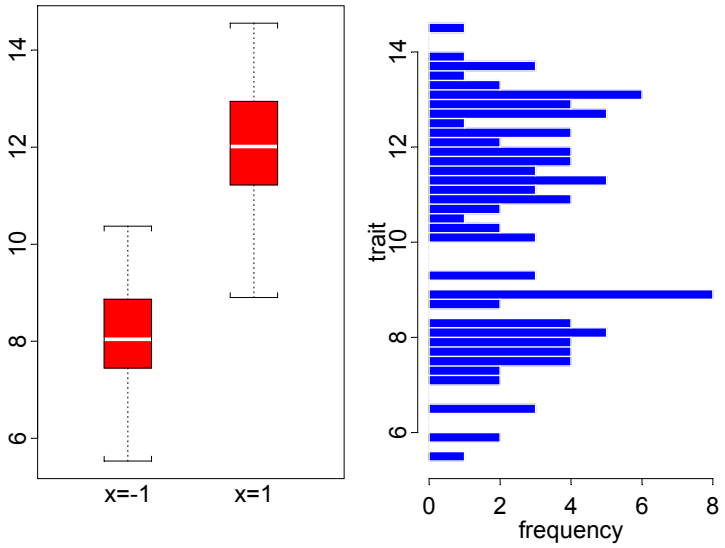
$$Y = G_Q + E$$

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

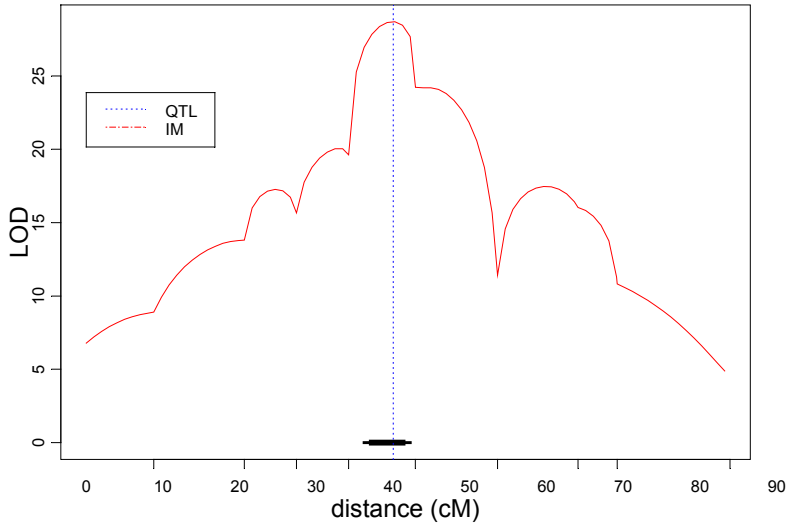
$$\text{normal}(G_Q, \sigma^2)$$



Simulated Data with 1 QTL



Profile LOD for 1 QTL



Interval Mapping Likelihood

- likelihood: basis for scanning the genome
  - product over $i = 1,\ldots,n$ individuals

  $L(\theta,\lambda|Y) = \text{product}_i \text{ pr}(Y_i|X_i,\lambda)$

  $= \text{product}_i \text{ sum}_Q \text{ pr}(Q|X_i,\lambda) \text{ pr}(Y_i|Q,\theta)$
  - complicated procedure to get estimates
- null likelihood: no QTL

  $L(\theta|Y) = \text{product}_i \text{ pr}(Y_i|\theta) = \text{product}_i \text{ N}(Y_i|\mu,\sigma^2)$
  - usual mean and variance estimates

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)

$$LOD(\lambda) = \text{sum}_i \log_{10}\left(\frac{\text{sum}_Q \ \text{pr}(Y_i \mid Q, \hat{\theta}) \ \text{pr}(Q \mid \lambda)}{\text{pr}(Y_i \mid \widetilde{\theta})} \right)$$

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

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

$$\text{genetic effects } CI = \hat{G}_Q \pm 1.96 \times \text{se}(\hat{G}_Q)$$

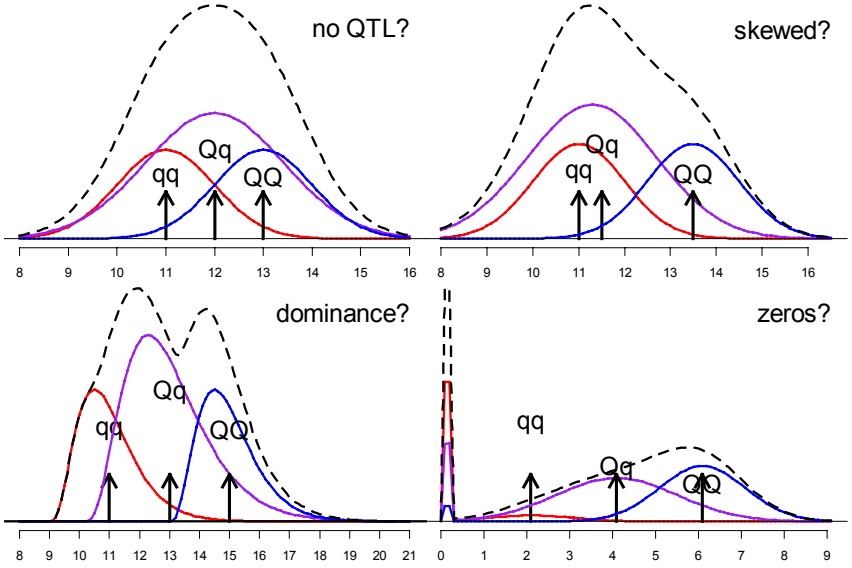
Part II: Extension of Phenotype Model

- limitations of parametric models
- quick fixes (but watch out!)
- semi-parametric approaches
- non-parametric approaches
- bottom line:
  - normal phenotype model works well to pick up loci, but may be poor at estimates of effects

Limitations of Parametric Models

- measurements not normal
  - counts (*e.g.* number of tumors)
  - survival time (*e.g.* days to flowering)
- false positives due to miss-specified model
  - check model assumptions?
- want more robust estimates of effects
  - parametric: only center (mean), spread (SD)
  - shape of distribution may be important

What if data are far away from ideal?



What shape is your histogram?

- histogram conditional on known QT genotype
  - pr(Y|qq, θ) model shape with genotype qq
  - pr(Y|Qq, θ) model shape with genotype Qq
  - pr(Y|QQ, θ) model shape with genotype QQ
- is the QTL at a given locus λ?
  - no QTL pr(Y|qq, θ) = pr(Y|Qq, θ) = pr(Y|QQ, θ)
  - QTL present mixture if genotype unknown
- mixture across possible genotypes
  - sum over Q = qq, Qq, QQ
  - pr($Y|X,\lambda,\theta$) = sum_Q pr($Q|X,\lambda$) pr($Y|Q,\theta$)

Interval Mapping Likelihood

- likelihood: basis for scanning the genome
  - product over $i = 1,\ldots,n$ individuals

$$L(\theta,\lambda|Y) = \text{product}_i \text{ pr}(Y_i|X_i,\lambda)$$
$$= \text{product}_i \text{ sum}_Q \text{ pr}(Q|X_i,\lambda) \text{ pr}(Y_i|Q,\theta)$$

- problem: unknown phenotype model
  - parametric pr($Y|Q,\theta$) = normal(G_Q,σ^2)
  - semi-parametric pr($Y|Q,\theta$) = $f(Y)\exp(Y\beta_Q)$
  - non-parametric pr($Y|Q,\theta$) = $F_Q(Y)$

Useful Models & Transformations

- binary trait (yes/no, hi/lo, …)
  - map directly as another marker
  - categorical: break into binary traits?
  - mixed binary/continuous: condition on $Y > 0$?
- known model for biological mechanism
  - counts Poisson
  - fractions binomial
  - clustered negative binomial
- transform to stabilize variance
  - counts $\sqrt{Y}$ = sqrt(Y)
  - concentration log(Y) or log($Y+c$)
  - fractions arcsin($\sqrt{Y}$)
- transform to symmetry (approx. normal)
  - fraction log(Y/(1-Y)) or log(($Y+c$)/(1+c-Y))
- empirical transform based on histogram
  - watch out: hard to do well even without mixture
  - probably better to map untransformed, then examine residuals

Semi-parametric QTL

- phenotype model $\text{pr}(Y|Q,\theta) = f(Y)\exp(Y\beta_Q)$
  - unknown parameters $\theta = (f, \beta)$
    - $f(Y)$ is a (unknown) density if there is no QTL
    - $\beta = (\beta_{qq}, \beta_{Qq}, \beta_{QQ})$
    - $\exp(Y\beta_Q)$ `tilts' f based on genotype Q and phenotype Y
- test for QTL at locus λ
  - $\beta_Q = 0$ for all Q, or $\text{pr}(Y|Q,\theta) = f(Y)$
- includes many standard phenotype models

  normal $\text{pr}(Y|Q,\theta) = N(G_Q,\sigma^2)$

  Poisson $\text{pr}(Y|Q,\theta) = \text{Poisson}(G_Q)$

  exponential, binomial, …, but not negative binomial

Semi-parametric Empirical Likelihood

- phenotype model $\text{pr}(Y|Q,\theta) = f(Y)\exp(Y\beta_Q)$
  - "point mass" at each measured phenotype Y_i
  - subject to distribution constraints for each Q:
    $1 = \text{sum}_i f(Y_i)\exp(Y_i\beta_Q)$
- non-parametric empirical likelihood (Owen 1988)

$$\begin{aligned} L(\theta,\lambda|Y,X) &= \text{product}_i\,[\text{sum}_Q\,\text{pr}(Q|X_i,\lambda)\,f(Y_i)\exp(Y_i\beta_Q)] \\ &= \text{product}_i\,f(Y_i)\,[\text{sum}_Q\,\text{pr}(Q|X_i,\lambda)\exp(Y_i\beta_Q)] \\ &= \text{product}_i\,f(Y_i)\,w_i \end{aligned}$$

  - weights $w_i = w(Y_i|X_i,\beta,\lambda)$ rely only on flanking markers
    - 4 possible values for BC, 9 for F2, etc.
- profile likelihood: $L(\lambda|Y,X) = \max_\theta L(\theta,\lambda|Y,X)$

Semi-parametric Formal Tests

- clever trick: use partial empirical LOD
  - Zou, Fine, Yandell (2002 *Biometrika*)
  - $\text{LOD}(\lambda) \approx \log_{10} L(\lambda|Y,X)$
- has same formal behavior as parametric LOD
  - single locus test: approximately χ^2 with 1 d.f.
  - genome-wide scan: can use same critical values
  - permutation test: possible with some work
- can estimate cumulative distributions
  - nice properties (converge to Gaussian processes)

log empirical likelihood details

$\log(L(\theta,\lambda|Y,X)) = \text{sum}_i \log(f(Y_i)) + \log(w_i)$

now profile with respect to β,λ

$\log(L(\beta,\lambda|Y,X)) = \text{sum}_i \log(f_i) + \log(w_i)$

$+ \text{sum}_Q \alpha_Q(1-\text{sum}_i f_i \exp(Y_i\beta_Q))$

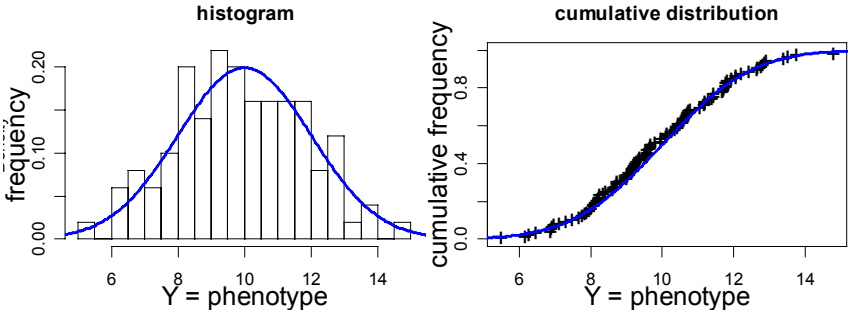
partial likelihood: set Lagrange multipliers α_Q to 0

point mass density estimates

$$f_i = \left[\text{sum}_Q \exp(Y_i\beta_Q)\, p(Q \mid X, \lambda)\right]^{-1}$$

with $p(Q \mid X, \lambda) = \text{sum}_i \text{pr}(Q \mid X_i, \lambda)$

Histograms and CDFs

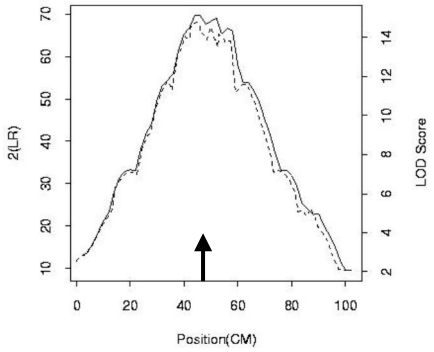


histograms capture shape but are not very accurate

CDFs are more accurate but not always intuitive

Rat study of breast cancer
Lan *et al.* (2001 *Genetics*)

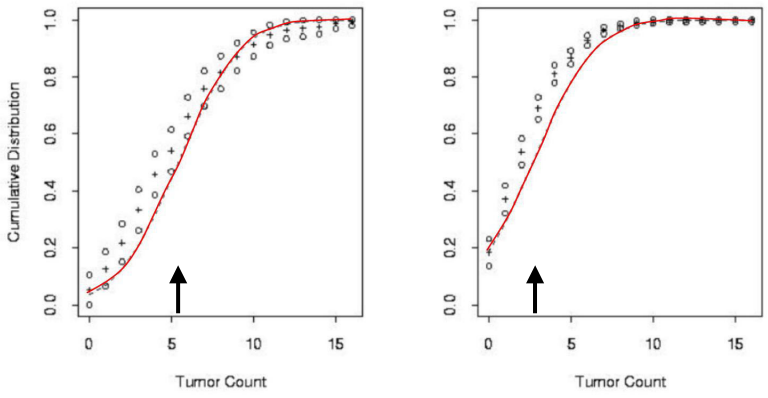
- rat backcross
  - two inbred strains
    - Wistar-Furth susceptible
    - Wistar-Kyoto resistant
  - backcross to WF
  - 383 females
  - chromosome 5, 58 markers
- search for resistance genes
- Y = # mammary carcinomas
- where is the QTL?



dash = normal
solid = semi-parametric

What shape histograms by genotype?

WF/WF WKy/WF



line = normal, + = semi-parametric, o = confidence interval

Non-parametric Methods

- phenotype model $\text{pr}(Y|Q,\theta) = F_Q(Y)$
  - $\theta = F = (F_{qq},F_{Qq},F_{QQ})$ arbitrary distribution functions
- Kruglyak Lander (1995)
  - interval mapping rank-sum test, replacing Y by rank(Y)
  - claimed no estimator of QTL effects
- estimators are indeed possible
  - semi-parametric shift (Hodges-Lehmann)
    - Zou (2001) thesis, Zou, Yandell, Fine (2002 in review)
  - non-parametric cumulative distribution
    - Fine, Zou, Yandell (2001 in review)

Rank-Sum QTL Methods

- phenotype model $\text{pr}(Y|Q,\theta) = F_Q(Y)$
- replace Y by rank(Y) and perform IM
  - extension of Wilcoxon rank-sum test
  - fully non-parametric
- Hodges-Lehmann estimator of shift β
  - most efficient if $\text{pr}(Y|Q,\theta) = F(Y+Q\beta)$
  - find β that matches medians
    - problem: genotypes Q unknown
    - resolution: Haley-Knott (1992) regression scan
  - works well in practice, but theory is elusive
    - Zou, Yandell Fine (*Genetics*, in review)

Non-Parametric QTL CDFs

- estimate non-parametric phenotype model
  - cumulative distributions $F_Q(y) = \text{pr}(Y \le y \,|Q)$
  - can use to check parametric model validity
- basic idea:

  $\text{pr}(Y \le y \,|X,\lambda) = \text{sum}_Q \text{pr}(Q|X,\lambda)F_Q(y)$
  - depends on X only through flanking markers
  - few possible flanking marker genotypes
    - 4 for BC, 9 for F2, etc.

Finding NP QTL CDFs

- cumulative distribution $F_Q(y) = \text{pr}(Y \le y \,|Q)$
- $F = \{F_Q$, all possible QT genotypes $Q\}$
  - BC with 1 QTL: $F = \{F_{QQ}, F_{Qq}\}$
- find F to minimize over all phenotypes y

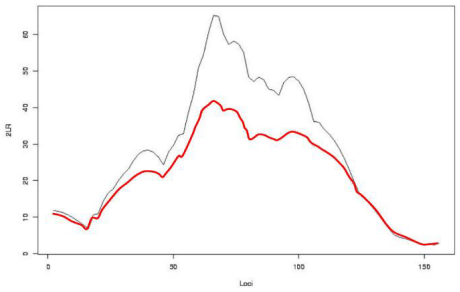
  $\text{sum}_i [I(Y_i \le y) - \text{sum}_Q \text{pr}(Q|X,\lambda)F_Q(y)]^2$
- looks complicated, but simple to implement

Non-parametric CDF Properties

- readily extended to censored data
  - time to flowering for non-vernalized plants
- nice large sample properties
  - estimates of $F(y) = \{F_Q(y)\}$ jointly normal
  - point-wise, experiment-wise confidence bands
- more robust to heavy tails and outliers
- can use to assess parametric assumptions

What QTL influence flowering time? no vernalization: censored survival

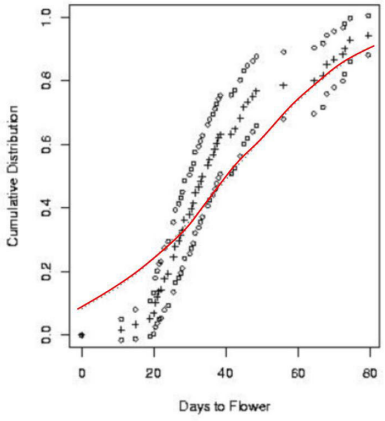
- *Brassica napus*
  - Major female
    - needs vernalization
  - Stellar male
    - insensitive
  - 99 double haploids
- Y = log(days to flower)
  - over 50% Major at QTL never flowered
  - log not fully effective



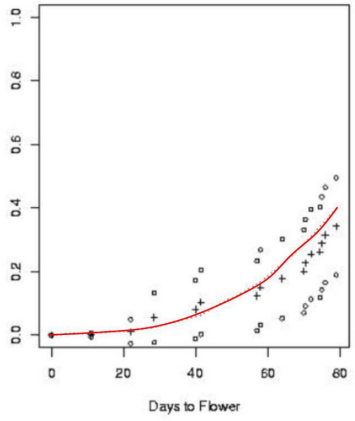
grey = normal, red = non-parametric

What shape is flowering distribution?

B. napus Stellar



B. napus Major



line = normal, + = non-parametric, o = confidence interval

Part III: Bayesian Idea

- Thomas Bayes and the original idea
- Bayes theorem
- How do frequentist & Bayesian approaches differ?
- Choice of Bayesian priors
  - normal data phenotype model
  - empirical Bayes
- Bayesian interval mapping basics

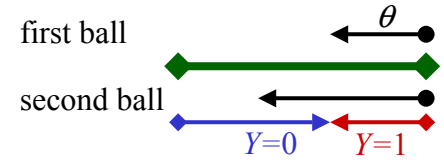
Bayesian QTL Model Selection

- 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
    - epistasis
- resampling from data
  - permutation tests
  - bootstrap, jackknife
  - MCMC
    - special Markov chain
    - Monte Carlo sampling
- show MCMC ideas
  - Gibbs sampler
  - Metropolis-Hastings
  - reversible jump MCMC

Who was Bayes?

- Reverend Thomas Bayes (1702-1761)
  - part-time mathematician
  - buried in Bunhill Cemetary, Moongate, London
  - famous paper in 1763 *Phil Trans Roy Soc London*
    - Barnard (1958 *Biometrika*), Press (1989) *Bayesian Statistics*
    - Stigler (1986) *History of Statistics*
    - Carlin Louis (1996); Gelman et al. (1995) books
  - Was Bayes the first with this idea? (Laplace)
- billiard balls on rectangular table
  - two balls tossed at random (uniform) on table
  - where is first ball if the second is to its right (left)?

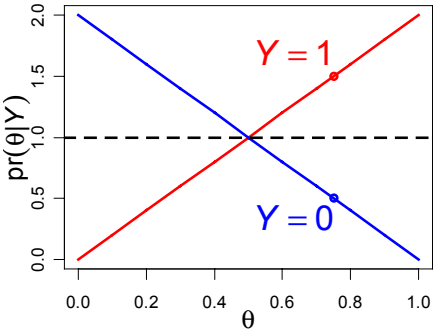
Where is the first ball?



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

likelihood $\quad \text{pr}(Y|\theta) = \theta^{Y}(1-\theta)^{1-Y}$

posterior $\quad \text{pr}(\theta|Y) = ?$



$$\text{pr}(\theta|Y) = \frac{\text{pr}(Y|\theta)\text{pr}(\theta)}{\text{pr}(Y)}$$

$$\text{pr}(Y) = \int_0^1 \theta^{Y}(1-\theta)^{1-Y}\,d\theta = \frac{1}{2}$$

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

(now throw second ball *n* times)

What is Bayes Theorem?

- before and after observing data
  - prior: $\text{pr}(\theta)$ = pr(parameters)
  - posterior: $\text{pr}(\theta|Y)$ = pr(parameters|data)
- posterior = likelihood * prior / constant
  - usual likelihood of parameters given data
  - normalizing constant pr(*Y*) depends only on data
    - constant often drops out of calculation

$$\text{pr}(\theta|Y) = \frac{\text{pr}(\theta, Y)}{\text{pr}(Y)} = \frac{\text{pr}(Y|\theta)\times\text{pr}(\theta)}{\text{pr}(Y)}$$

What is Probability?

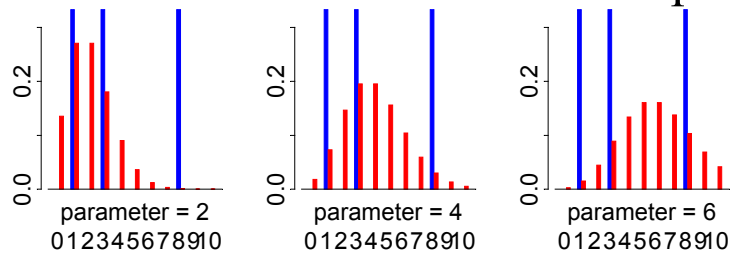
Frequentist analysis

- chance over many trials
  - long run average
    - estimates
    - confidence intervals
  - long term frequency
    - hypothesis tests
    - *p*-values
- Type I error rate
  - reject null when true
  - chance of extreme result

Bayesian analysis

- uncertainty of true value
- prior
  - uncertainty before data
  - incorporate prior knowledge/experience
- posterior
  - uncertainty after analyzing current data
  - balance prior and data

Likelihood and Posterior Example



$$\text{data} : Y = 1,3,8$$

$$\text{parameter} : \theta = ?$$

$$\text{pr}(Y = y \mid \theta) = \frac{\theta^y e^{-\theta}}{y!}$$

(M. Newton, pers. comm.)

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

Choice of Bayesian priors

- elicited priors
  - higher weight for more probable parameter values
    - based on prior empirical knowledge
  - use previous study to inform current study
    - weather prediction, previous QTL studies on related organisms
- conjugate priors
  - convenient mathematical form
  - essential before computers, helpful now to simply computation
  - large variances on priors reduces their influence on posterior
- non-informative priors
  - may have "no" information on unknown parameters
  - prior with all parameter values equally likely
    - may not sum to 1 (improper), which can complicate use
- always check sensitivity of posterior to choice of prior

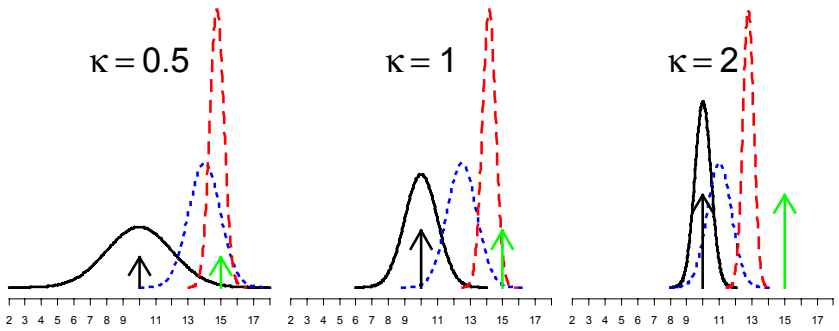
Bayes for normal data

$Y = G + E$	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	$\mathrm{pr}(G \mid \sigma^2, \mu, \kappa) = \mathrm{N}(G \mid \mu, \sigma^2/\kappa)$
posterior	$\mathrm{N}(G \mid \mu + B_1(Y - \mu), B_1\sigma^2)$
$Y_i = G + E_i$	posterior for sample of n individuals
	shrinkage weights B_n go to 1

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

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

effect of prior variance on posterior



normal prior, posterior for $n = 1$, posterior for $n = 5$, true mean

posterior by QT genetic value

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

environ $E \sim N(0,\sigma^2)$, σ^2 known

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

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

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

posterior:

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

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

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

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

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

estimates $\quad \hat{\mu} = \overline{Y}_\bullet, s^2 = \text{sum}_i (Y_i - \overline{Y}_\bullet)^2 / n$

$\kappa \le 1$ or $\kappa = \sigma^2 / s^2$

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

What if variance σ^2 is unknown?

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

joint effects posterior details

$Y_i = G(Q_i) + E_i$

genetic Q_i = qq,Qq,QQ

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

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

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

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

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

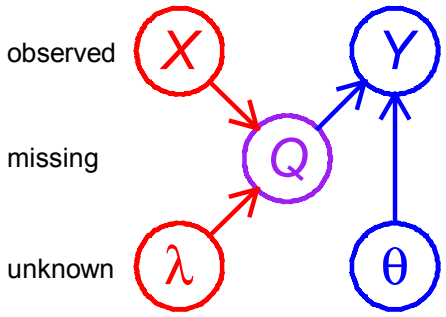
posterior: $\text{pr}(G_Q \mid Y, Q, \sigma^2, \mu, \kappa) = N\left(G_Q \middle| \bar{Y}_\cdot + B_Q(\bar{Y}_Q - \bar{Y}_\cdot), B_Q \frac{\sigma^2}{n_Q}\right)$

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

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

Bayesian Idea for QTLs

- key idea
  - sample missing genotypes Q
    - using recombination model
  - phenotype model given Q
    - see previous slides
- methods and philosophy
  - EM & MCMC
  - Frequentists & Bayesians



- review interval maps & profile LODs
- case study: simulated single QTL

QTL Full Posterior

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

pr(loci, genos, effects | trait, map)

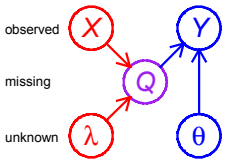
$$\text{pr}(\lambda, Q, \theta \mid Y, X) = \frac{\text{pr}(\theta)\text{pr}(\lambda)\text{pr}(Q \mid X, \lambda)\text{pr}(Y \mid Q, \theta)}{\text{constant}}$$

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

$$\text{pr}(Q \mid X, \lambda)\text{pr}(Y \mid Q, \theta) = \text{product}_i \left[\text{pr}(Q \mid X_i, \lambda)\text{pr}(Y_i \mid Q, \theta)\right]$$

marginal posteriors

- joint posterior
  - pr($\lambda, Q, \theta \mid Y, X$) = pr($\theta$)pr($\lambda$)pr($Q \mid X, \lambda$)pr($Y \mid Q, \theta$) /constant
- genetic effects
  - pr($\theta \mid Y, X$) = sum$_Q$ pr($\theta \mid Y, Q$) pr($Q \mid Y, X$)
- QTL locus
  - pr($\lambda \mid Y, X$) = sum$_Q$ pr($\lambda \mid X, Q$) pr($Q \mid Y, X$)
- QTL genotypes more complicated
  - pr($Q \mid Y, X$) = sum$_{\lambda, \theta}$ pr($Q \mid Y, X, \lambda, \theta$) pr($\lambda, \theta \mid Y, X$)
  - impossible to separate λ and θ in sum



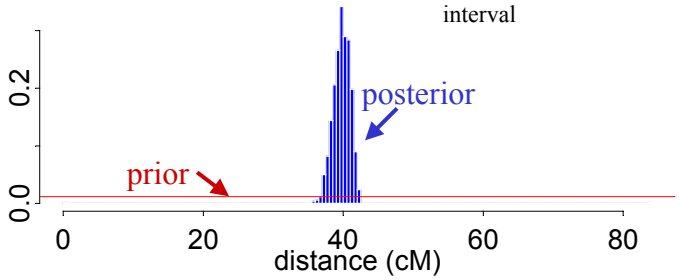
prior & posterior for genotypes Q

- prior is recombination model
  $\text{pr}(Q|X_i,\lambda)$
- can explicitly decompose by individual i
  - binomial (or trinomial) probability
- posterior for genotype depends on
  - effects via trait model
  - locus via recombination model
- posterior agrees exactly with interval mapping
  - used in EM: estimation step
  - but need to know locus λ and effects θ

$$P_{Qi} = \text{pr}(Q \mid Y_i, X_i, \lambda, \theta) = \frac{\text{pr}(Y_i \mid Q, \theta)\text{pr}(Q \mid X_i, \lambda)}{\text{sum}_Q \left[\text{pr}(Y_i \mid Q, \theta)\text{pr}(Q \mid X_i, \lambda)\right]}$$

prior & posterior for QT 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



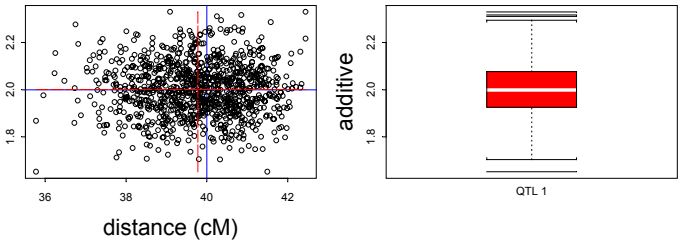
QTL Bayesian Inference

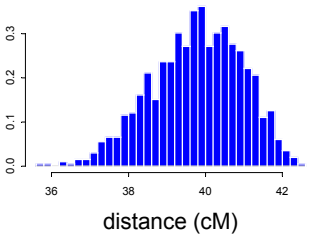
- study posterior distribution of locus & effects
  - sample joint distribution
    - locus, effects & genotypes
  - study marginal distribution of
    - locus
    - effects
      - overall mean, genotype difference, variance
    - locus & effects together
- estimates & confidence regions
  - histograms,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

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)

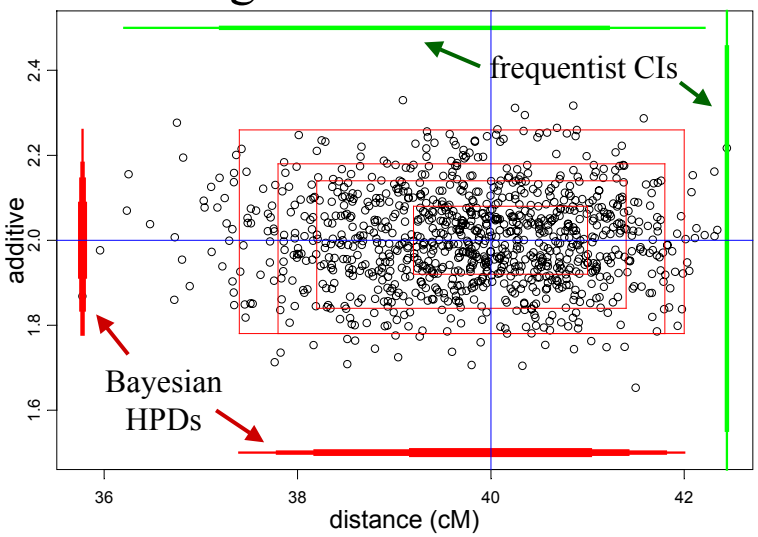
Posterior for locus & effect





summaries of MC samples
joint scatter plot
(lines for true, estimates)
marginal histogram
marginal box-plot

HPD region for locus & effect



Part IV: Monte Carlo Sampling from the Bayesian posterior

- How can we study Bayesian posteriors?
  - frequentist & Bayesian methods
  - EM & MCMC duality
- What is a Markov chain?
- How does Markov chain Monte Carlo work?
  - use Markov chain to sample from posterior
  - Gibbs sampler for effects θ, genotypes Q
  - Metropolis-Hastings for loci λ

How can we study Bayesian posteriors?

- exact methods if possible
  - manipulate math formula
  - can be difficult or impossible to analyze
- approximate methods
  - importance sampling
  - numerical integration
  - Monte Carlo & other
- Monte Carlo methods
  - easy to implement
  - independent samples
  - just draw large sample
- MCMC methods
  - handle hard problems
  - art to efficient use
  - iterative: Markov chain
  - correlated samples

How to study the QTL likelihood?

- frequentist approach
  - maximize (*IM)
    - find the peak
    - avoid local maxima
  - profile LOD across locus
    - maximize for effects
- approximate methods
  - EM (Lander Botstein 1989)
  - MCMC (Guo Thompson 1994)
- Bayesian approach
  - sample from posterior
  - examine whole posterior
    - summarize later
    - joint locus & effects
- approximate methods
  - MCMC (Satagopan et al. 1996)
  - imputation (Sen Churchill 2001)

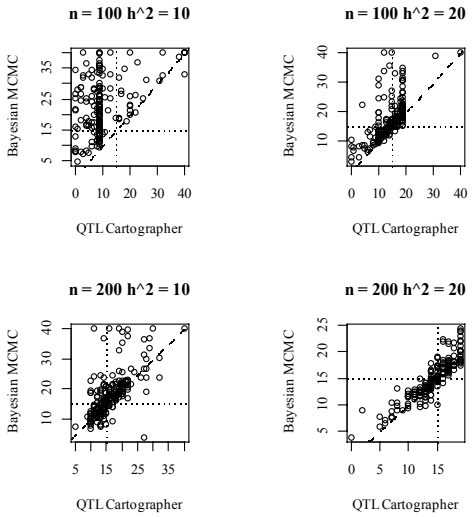
EM-MCMC duality

- EM approaches can be redone with MCMC
  - EM estimates & maximizes
  - MCMC draws random samples
    - simulated annealing: gradually cool towards peak
  - 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! (math, computing tricks)
  - very handy for hard problems in genetics

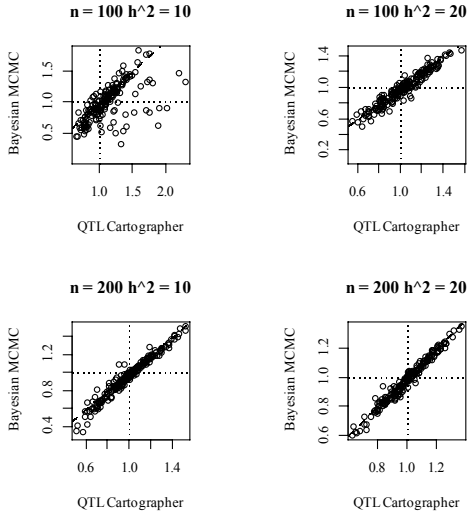
Simulation Study

- 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 heritability h^2

200 Simulations: Locus



200 Simulations: Effect



Why not Ordinary Monte Carlo?

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

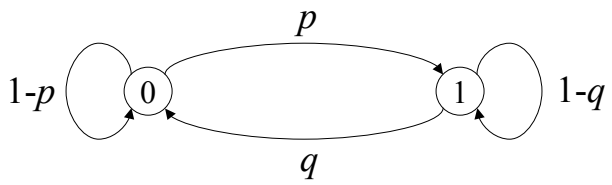
$$\text{pr}(\theta|Y,Q) = \text{pr}(G_Q \mid Y,Q,\sigma^2)\ \text{pr}(\sigma^2 \mid Y,Q)$$

- 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
  - $\text{pr}(\theta|Y,X) = \text{sum}_Q\ \text{pr}(\theta|Y,Q)\ \text{pr}(Q|Y,X)$

What is a Markov chain?

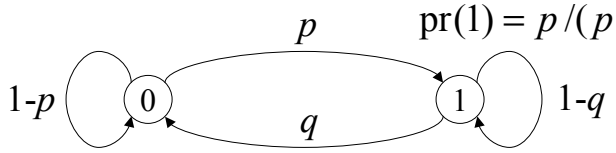
- future given present is independent of past
- update chain based on current value
  - can make chain arbitrarily complicated
  - chain converges to stable pattern $\pi()$ we wish to study

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



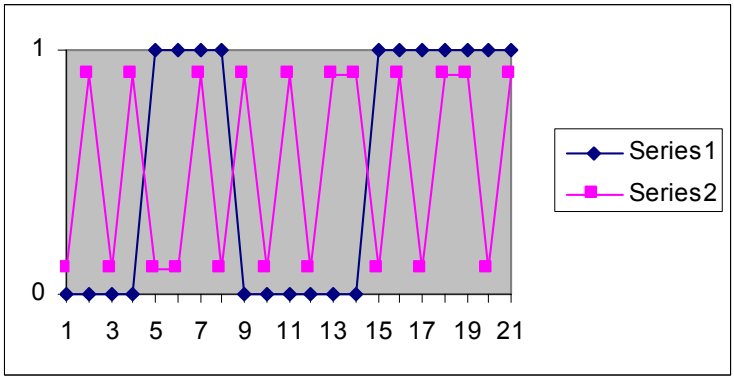
Markov chain idea

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



Mitch's

other pubs



Markov chain Monte Carlo

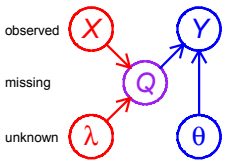
- can study arbitrarily complex models
  - need only specify how parameters affect each other
  - can reduce to specifying full conditionals
- construct Markov chain with "right" model
  - joint posterior of unknowns as limiting "stable" distribution
  - update unknowns given data and all other unknowns
    - 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

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

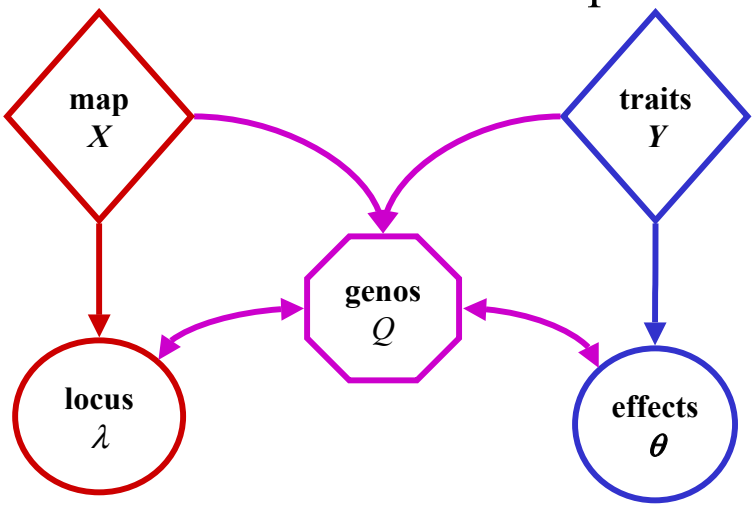
$$(\lambda, Q, \theta) \sim \text{pr}(\lambda, Q, \theta \mid Y, X)$$
$$(\lambda, Q, \theta)_1 \rightarrow (\lambda, Q, \theta)_2 \rightarrow \cdots \rightarrow (\lambda, Q, \theta)_N$$

sample from full conditionals



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

MCMC full conditional updates



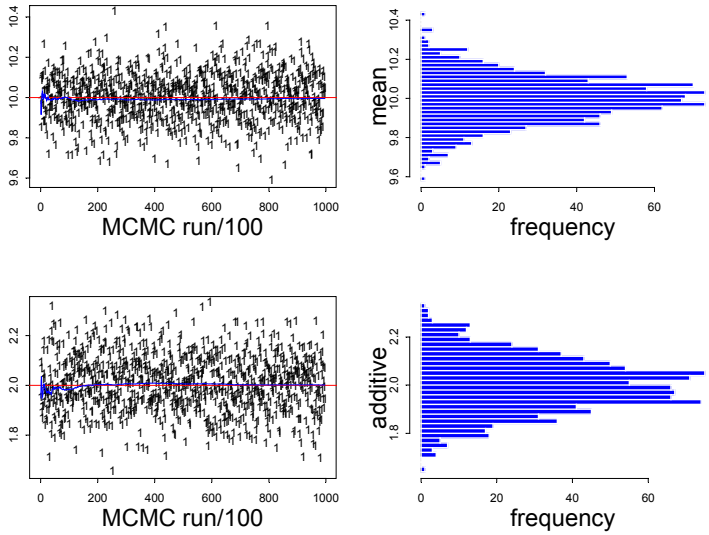
Gibbs Sampler: effects & genotypes

- for given locus λ, can sample effects θ and genotypes Q
  - effects parameter vector $\theta = (G,\sigma^2)$ with $G=(G_{qq},G_{Qq},G_{QQ})$
  - 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

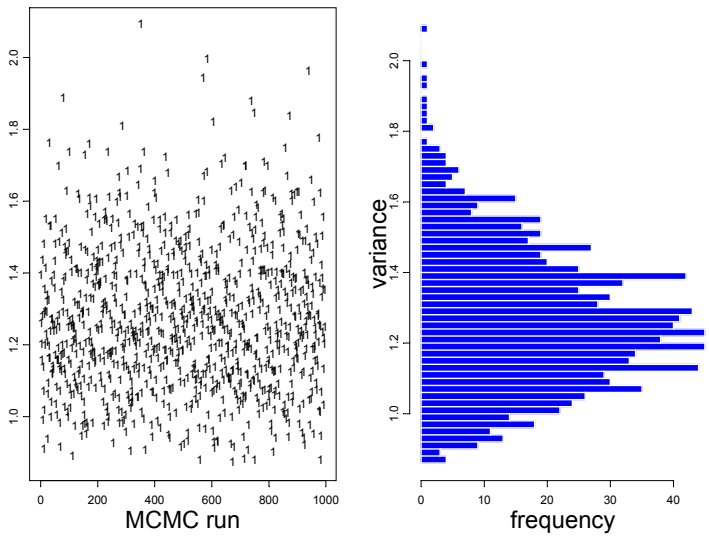
phenotype model: alternate form

- genetic value $G(Q)=G_Q$ 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
  - $G_{qq} = \mu - a$, $G_{Qq} = \mu$, $G_{QQ} = \mu + a$
- 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

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

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

full conditional for locus

- cannot easily sample from locus full conditional

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

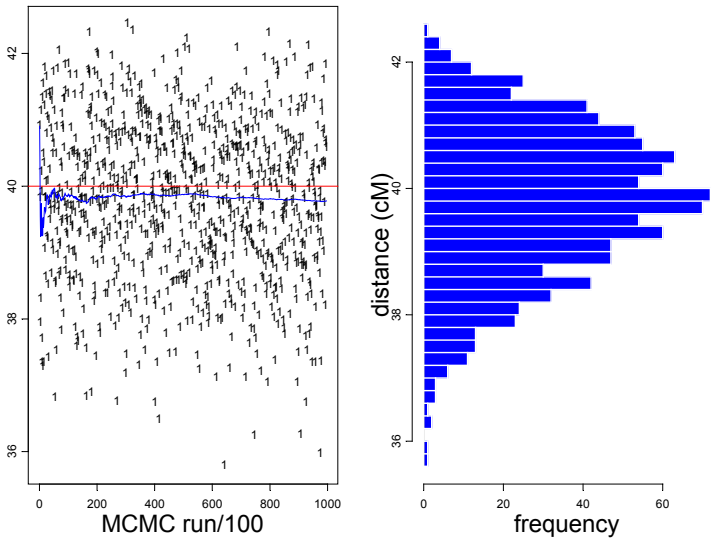
- cannot explicitly determine full conditional
  - difficult to normalize
  - need to average over all possible genotypes over entire map
- Gibbs sampler will not work
  - but can use method based on ratios of probabilities…

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



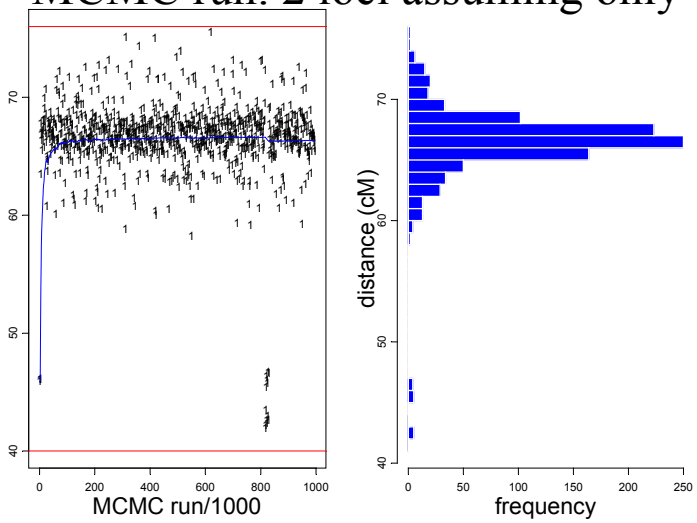
Care & Use of MCMC

- 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"

Part V: Multiple QTL

- multiple QTL phenotype model
- issues for 2 QTL
- MCMC sampling from the posterior
- Simulated data for 0,1,2 QTL
- *Brassica* data on days to flowering

MCMC run: 2 loci assuming only 1



Multiple QTL model

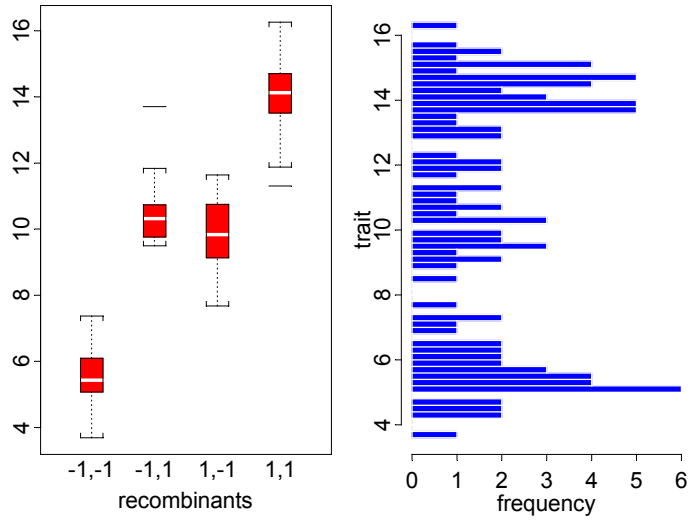
- trait = mean + add1 + add2 + error
- trait = genetic effect + error
- pr(trait | genos, effects)

$$Y_i = \mu + a_1 Q_{1i} + a_2 Q_{2i} + e_i$$

$$Y_i = \mu + \sum_{r=1}^{m} a_r Q_{ri} + e_i$$

$$\mu \sim \mathrm{N}(\overline{Y}_\cdot, s^2), a_r \sim \mathrm{N}(0, 2\beta s^2 / m), \beta = ?$$

Simulated Data with 2 QTL

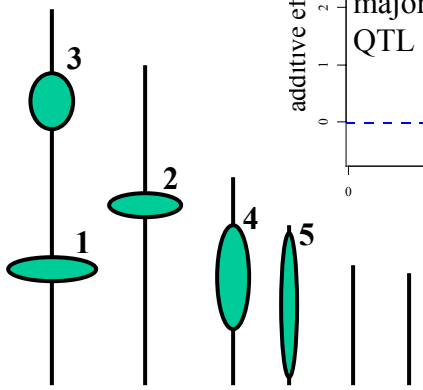


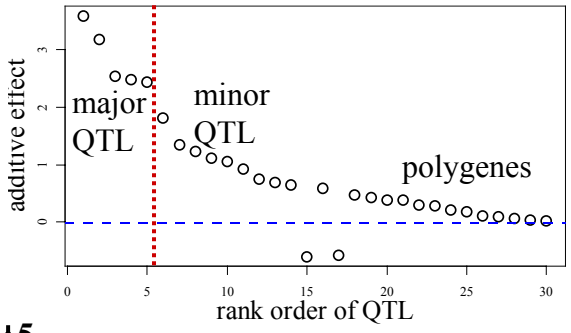
Issues for Multiple QTL

- how many QTL influence a trait?
  - 1, several (oligogenic) or many (polygenic)?
  - or do *most* genes influence most complex traits?
  - how many are supported by the data?
    - effects can be localized
- 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

Pareto diagram of QTL effects

major QTL on linkage map

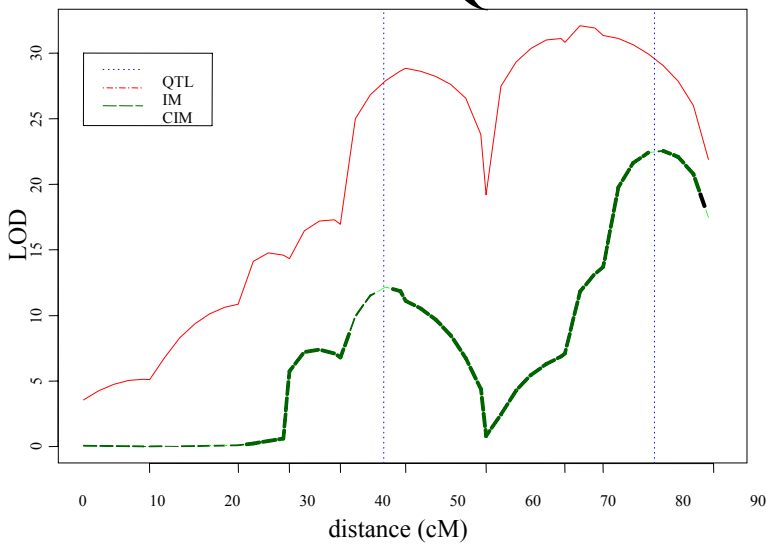




interval mapping approach

- 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)
  - use CIM as basic model
  - forward selection/backward elimination

LOD for 2 QTL



Bayesian approach

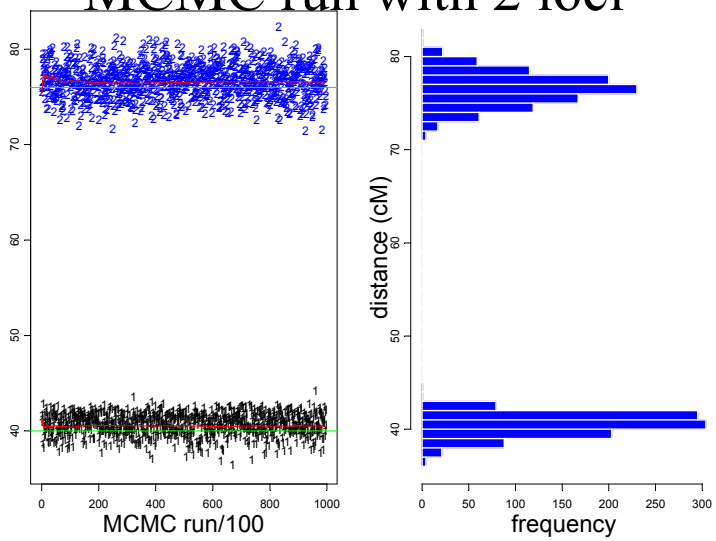
- simultaneous search for multiple QTL
- use Bayesian paradigm
  - easy to consider joint distributions
  - easy to modify later for other types of data
    - counts, proportions, etc.
  - employ MCMC to estimate posterior dist
- study estimates of loci and effects
- model selection and assessment
 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

MCMC for multiple QTLs

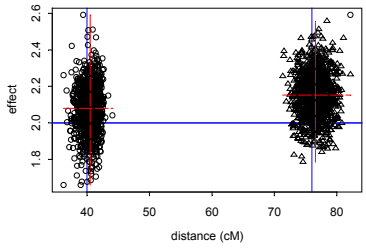
- posterior now has *m* loci rather than 1 locus
  - just change interpretation of unknowns
  - add extra subscript to keep track of loci
- construct Markov chain around posterior
  - "easy" extension of 1 QTL approach
  - now randomly pick which loci to update
- update all terms for each locus at one time?
  - open questions of efficient mixing

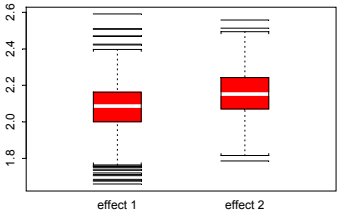
$$(\lambda, Q, \theta) \sim \text{pr}(\lambda, Q, \theta \mid Y, X)$$
$$(\lambda, Q, \theta)_1 \rightarrow (\lambda, Q, \theta)_2 \rightarrow \cdots \rightarrow (\lambda, Q, \theta)_N$$

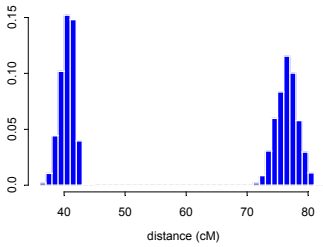
MCMC run with 2 loci



effects for 2 simulated QTL





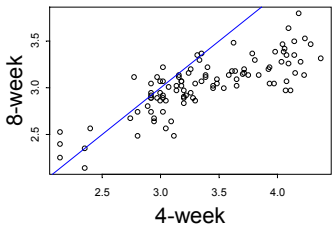


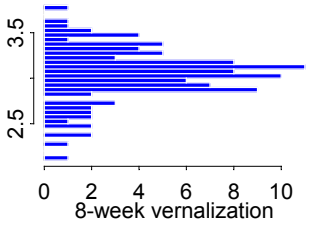
summaries of MC samples
joint scatter plots by loci
(lines for true, estimates)
marginal histograms
marginal box-plots

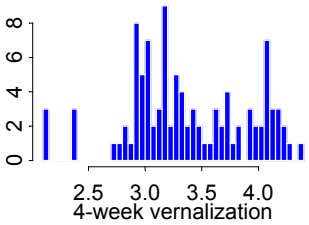
Brassica napus data

- 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

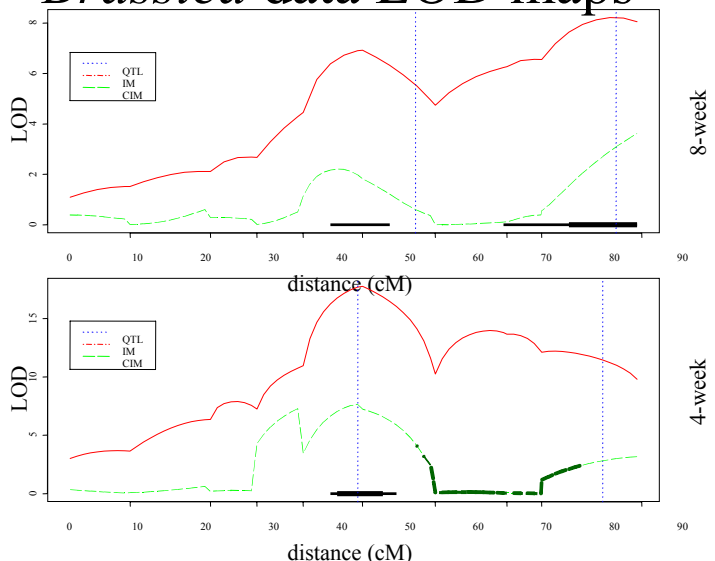




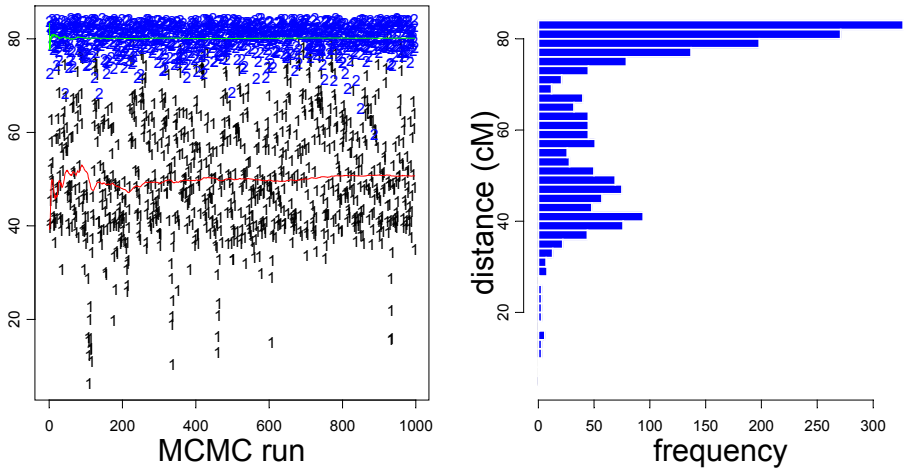


summaries of raw data
joint scatter plots
(identity line)
separate histograms

Brassica data LOD maps



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

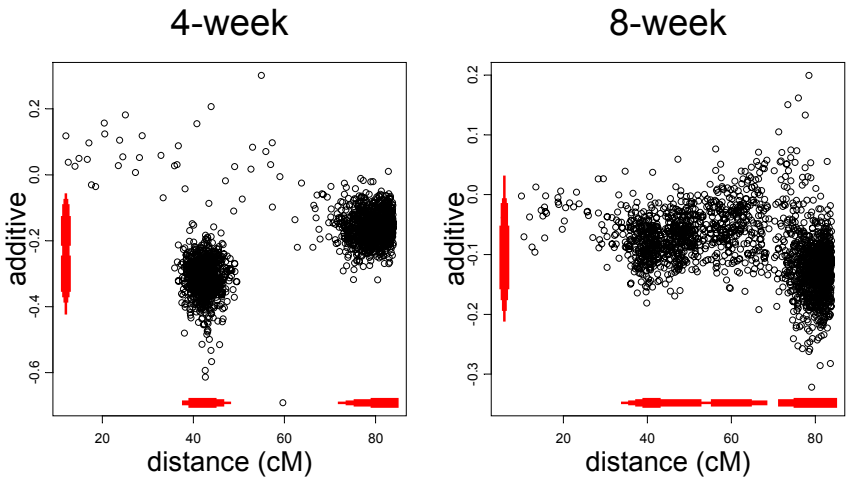
cM	add
40	.30
80	.16

8-week vernalization

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

cM	add
40	.06
80	.13

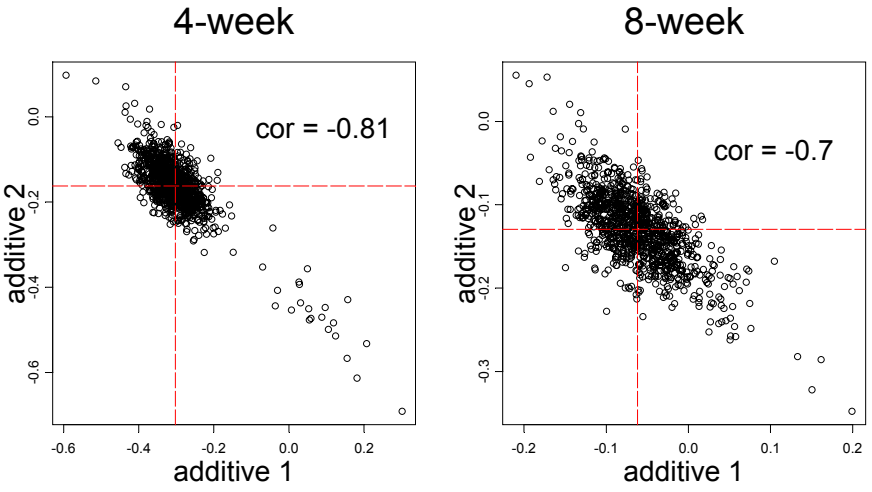
Brassica credible regions



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

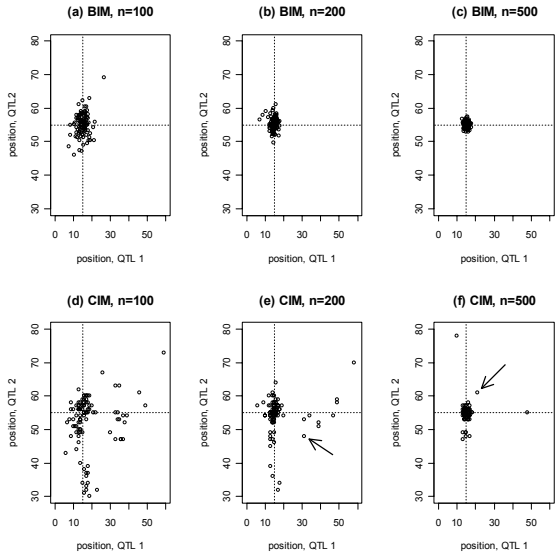
Brassica effect correlations



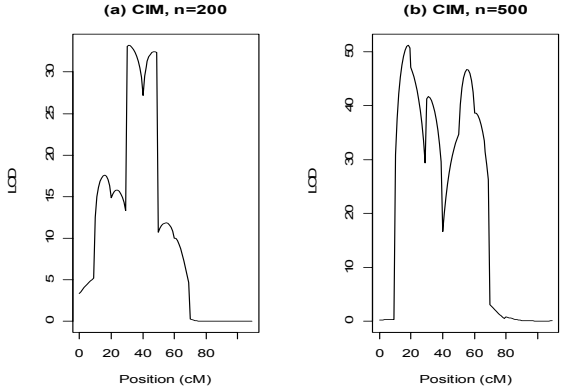
simulation: Bayesian vs composite IM

- 11 markers at 10cM spacing
- 2 QTL at 15cM, 55cM
  - effect size 1, variance 1, heritability 74%
- sample sizes n = 100, 200, 500
  - 100 independent trials
- comparison of methods
  - Bayesian interval mapping, 400,000 scans
  - composite interval mapping (QTL Cart)

Bayesian vs. composite IM: 2 QTL



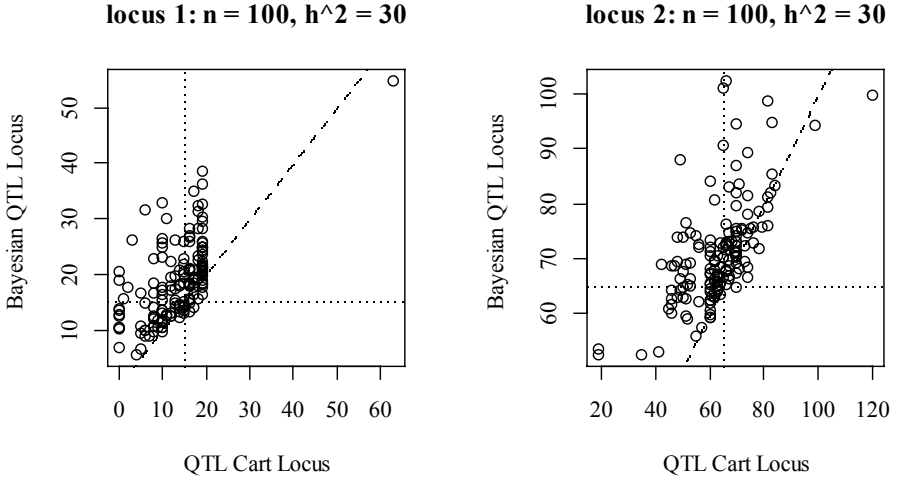
ghost QTL detected with CIM two selected examples



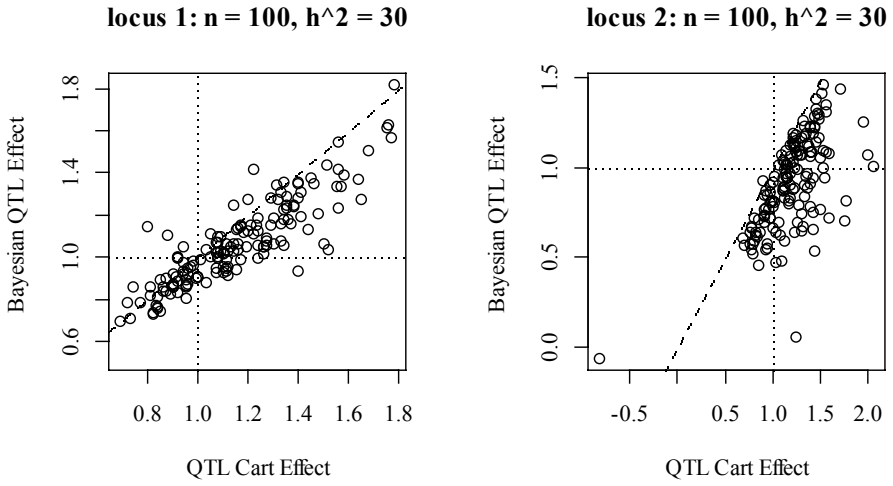
another 2-QTL simulation

- CIM vs. Bayesian QTL estimates
  - locus: 15, 65cM
  - effect: 1, 1
- sample sizes and heritabilities
  - $n = 100$, $h^2 = 30$
  - $n = 200$, $h^2 = 25, 30, 40$
  - 100 independent trials
- examine loci and effects

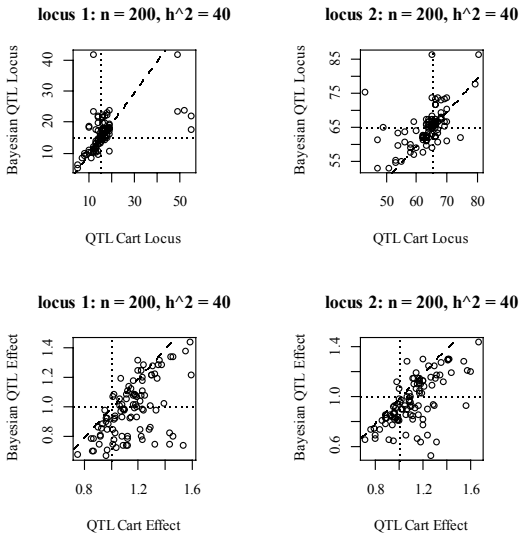
2 QTL: loci estimates



2 QTL: effect estimates



2 QTL: loci and effects



Part VI: How many QTLs?

- How many QTLs?
  - number of QTL is uncertain
  - estimate the number *m*
- What is the genetic architecture?
  - model architecture is uncertain
    - number, gene action, interaction among QTL
  - estimate the model *M*
- How does reversible jump MCMC work?
  - basic idea of Green(1995)
  - model selection in regression

QTL Full Posterior

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

pr(loci, genos, effects, model | trait, map)

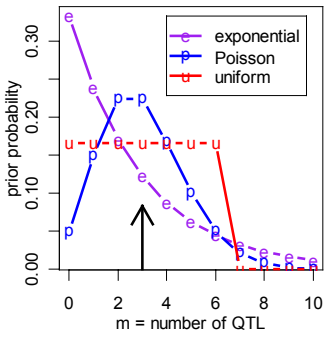
$$\text{pr}(\lambda, Q, \theta, M \mid Y, X) = \frac{\text{pr}(\theta \mid M)\text{pr}(\lambda \mid M)\text{pr}(Q \mid X, \lambda, M)\text{pr}(Y \mid Q, \theta, M)\text{pr}(M)}{\text{constant}}$$

$$\text{constant} = \text{pr}(Y) = \text{sum}_M \text{pr}(Y \mid M)$$

$$\text{pr}(Q \mid X, \lambda, M)\text{pr}(Y \mid Q, \theta, M) = \text{product}_i \left[\text{pr}(Q \mid X_i, \lambda, M)\text{pr}(Y_i \mid Q, \theta, M)\right]$$

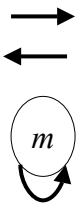
jumping the number of QTL

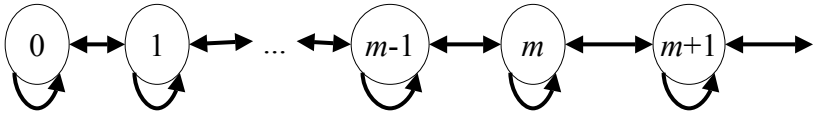
- model changes with number of QTL
  - analogous to stepwise regression if Q known
  - use reversible jump MCMC to change number
    - book keeping to compare models
    - change of variables between models
- what prior on number of QTL?
  - uniform over some range
  - Poisson with prior mean
  - exponential with prior mean



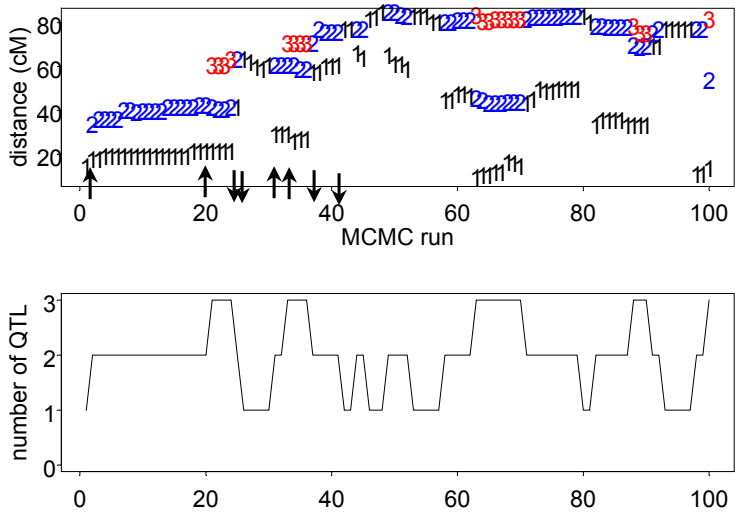
Markov chain for number m

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

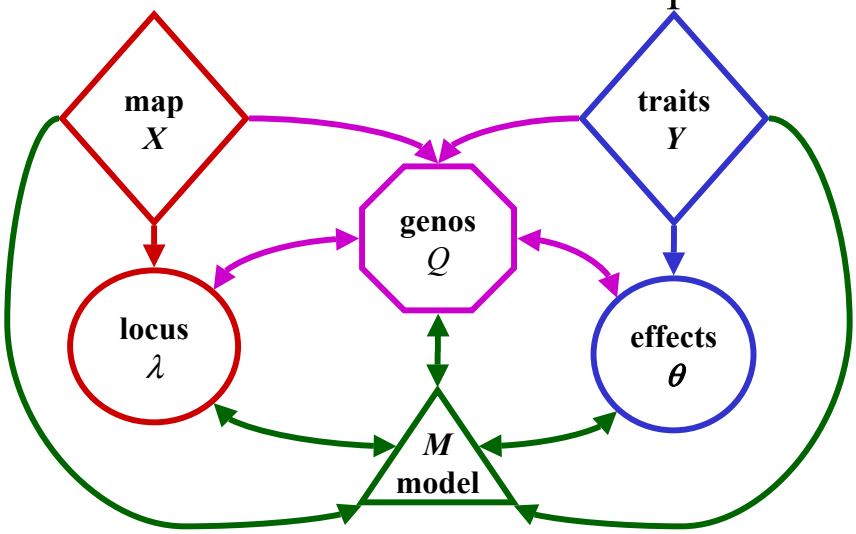




jumping QTL number and loci



RJ-MCMC full conditional updates

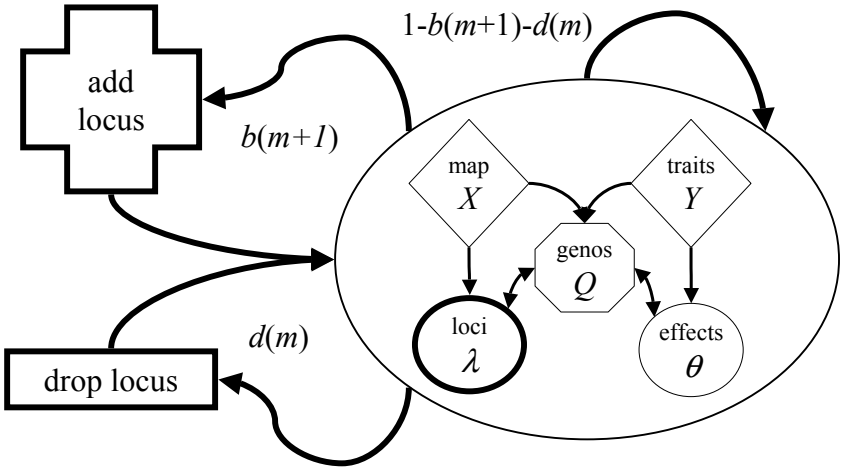


reversible jump choices

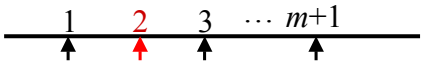
action step: draw one of three choices
(m = number of QTL in model)

- update step with probability 1-$b(m+1)$-$d(m)$
  - 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

RJ-MCMC updates



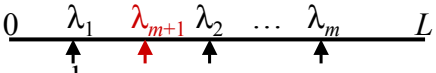
propose to drop a locus



- choose an existing locus
  - 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

$$q_d(r;m+1)=\frac{1}{m+1}$$

propose to add a locus



- propose a new locus
  - uniform chance over genome
  - actually need to be more careful (R van de Ven, pers. comm.)
    - choose interval between loci already in model (include 0,L)
      - probability proportional to interval length $(\lambda_2-\lambda_1)/L$
    - uniform chance within this interval $1/(\lambda_2-\lambda_1)$
  - 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 ...

$$q_b(\lambda)=1/L$$

acceptance of reversible jump

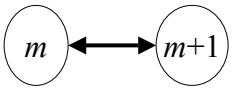
- accept birth of new locus with probability
$\min(1, A)$
- accept death of old locus with probability
$\min(1, 1/A)$

$$A = \frac{\text{pr}(\theta_{m+1}, m+1 \mid Y, X)}{\text{pr}(\theta_m, m \mid Y, X)} \times \frac{d(m+1)}{b(m)} \frac{q_b(\lambda_{m+1})}{q_d(r; m+1)} \frac{1}{J}$$

$$\theta_m = (Q, \theta, \lambda, m)$$

acceptance of reversible jump

- move probabilities



$$\frac{d(m+1)}{b(m)}$$

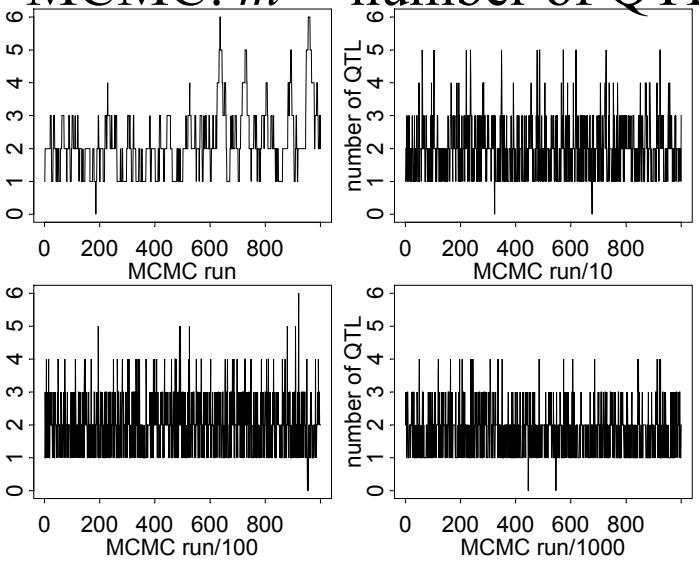
- birth & death proposals

$$\frac{q_b(\lambda_{m+1})}{q_d(r; m+1)}$$

- Jacobian between models
  - –fudge factor
  - –see stepwise regression example

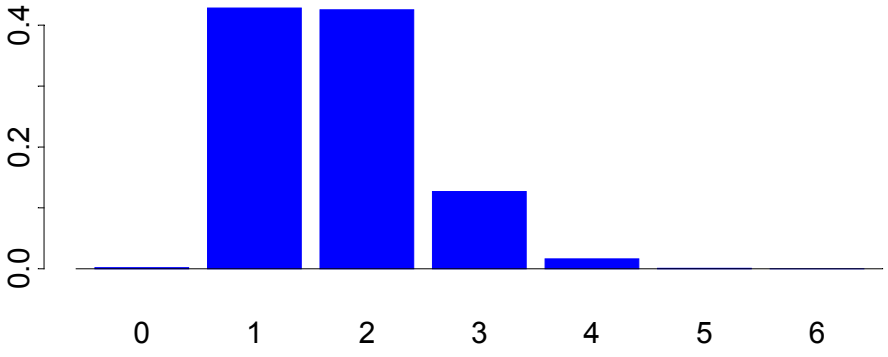
$$J = \frac{\sigma}{s_{r|others}\sqrt{n}}$$

RJ-MCMC: m = number of QTL



posterior for m for 8-week data

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



Part 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

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

$$m = 1 : Y_i = \mu + a(Q_{i1} - \overline{Q}_1) + e_i$$

$$m = 2 : Y_i = \mu + a_1(Q_{i1} - \overline{Q}_1) + a_2(Q_{i2} - \overline{Q}_2) + e_i$$

slope estimate for 1 regressor

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

$$\hat{a} = \frac{r_{1y}\, s_y}{s_1}, \quad r_{1y} = \frac{\sum_{i=1}^{n}(Q_{i1} - \overline{Q}_1)(Y_i - \overline{Y})/n}{s_1\, s_y}$$

$$s_1^2 = \sum_{i=1}^{n}(Q_{i1} - \overline{Q}_1)^2/n, \; s_y^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2/n$$

2 correlated regressors

slopes adjusted for other regressors

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

$$\hat{a}_2 = \frac{(r_{2y} - r_{12}r_{1y})s_y}{s_2}, s^2_{2\cdot 1} = \frac{\sum_{i=1}^{n}(Q_{i2} - \overline{Q}_2 - c_{21}(Q_{i1} - \overline{Q}_1))^2}{n}$$

Gibbs Sampler for Model 1

- mean $$\mu \sim \phi\left(\eta + B_n(\overline{Y}_\cdot - \eta), B_n\frac{\sigma^2}{n}\right), B_n = \frac{n}{n+\kappa}$$
- slope $$a \sim \phi\left(B_n\frac{\sum_{i=1}^{n}(Q_{i1} - \overline{Q}_1)(Y_i - \overline{Y}_\cdot)}{ns_1^2}, B_n\frac{\sigma^2}{ns_1^2}\right)$$
- variance $$\sigma^2 \sim \text{inv-}\chi^2\left(\nu + n, \frac{\nu\tau^2 + \sum_{i=1}^{n}\left(Y_i - \overline{Y}_\cdot - a(Q_{i1} - \overline{Q}_1)\right)^2}{\nu + n}\right)$$

Gibbs Sampler for Model 2

- mean $\mu \sim \phi\left(\eta + B_n(\overline{Y}_\cdot - \eta), B_n \frac{\sigma^2}{n}\right)$
- slopes $a_2 \sim \phi\left(B_n \frac{\sum_{i=1}^{n}(Q_{i2} - \overline{Q}_2)(Y_i - \overline{Y}_\cdot - a_1(Q_{i1} - \overline{Q}_1))}{ns_{2\cdot 1}^2}, B_n \frac{\sigma^2}{ns_{2\cdot 1}^2}\right)$
- variance $\sigma^2 \sim \text{inv-}\chi^2\left(v + n, \frac{v\tau^2 + \sum_{i=1}^{n}\left(Y_i - \overline{Y}_\cdot - \sum_{k=1}^{2} a_k(Q_{ik} - \overline{Q}_k)\right)^2}{v+n}\right)$

updates from 2->1

- drop 2nd regressor
- adjust other regressor

$$a \rightarrow a_1 + a_2 c_{21}$$
$$a_2 \rightarrow 0$$

updates from 1->2

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

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

$$a_2 \to \hat{a}_2 + z \times J, \quad \hat{a}_2 = \frac{\sum_{i=1}^{n} (Q_{i2} - \overline{Q}_2)\left(Y_i - \hat{\mu} - \hat{a}_1 (Q_{i1} - \overline{Q}_1)\right)}{n s_{2\cdot1}^2}$$

$$a_1 \to a - a_2 c_{21} = a - z \times c_{21} J - \hat{a}_2 c_{21}$$

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
$1 \to 2$	$(\mu, a, \sigma^2; z)$	$z \sim \phi(0,1)$	$\begin{Bmatrix} a_2 \to \hat{a}_2 + z \times J \\ a_1 \to a - a_2 c_{21} \end{Bmatrix}$
$2 \to 1$	$(\mu, a_1, a_2, \sigma^2)$		$\begin{Bmatrix} a \to a_1 + a_2 c_{21} \\ z \to 0 \end{Bmatrix}$

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)

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{bmatrix} 1 & -c_{21}J & -c_{21} \\ 0 & J & 1 \end{bmatrix} \times \begin{pmatrix} a \\ z \\ \hat{a}_2 \end{pmatrix} = g(a;z \mid Y, Q_1, Q_2)$$

$$\int \pi(a_1, a_2 \mid Y, Q_1, Q_2) da_1 da_2 = \int \pi(a;z \mid Y, Q_1, Q_2) J da dz$$

Jacobian & the calculus

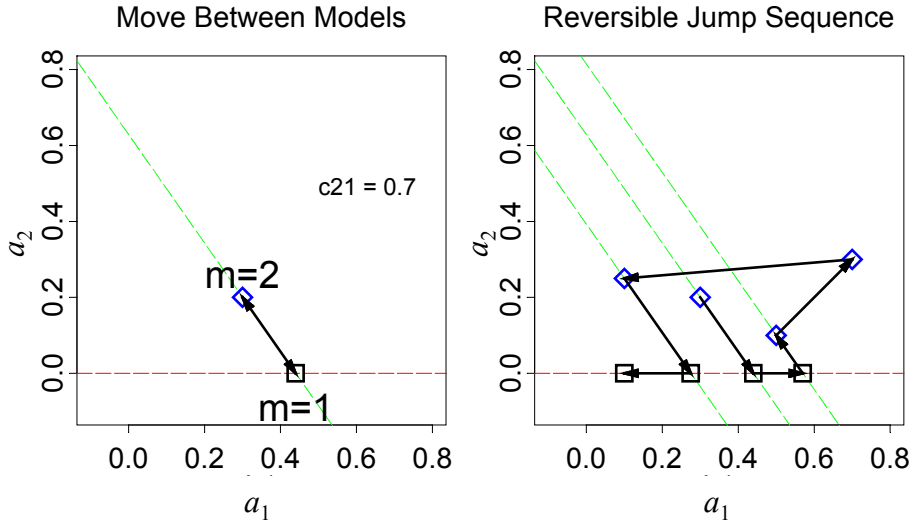
- Jacobian sorts out change of variables
  - careful: easy to mess up here!

$$g(a;z) = (a_1, a_2), \quad \frac{\partial g(a;z)}{\partial a \partial z} = \begin{bmatrix} 1 & -c_{21}J \\ 0 & J \end{bmatrix}$$

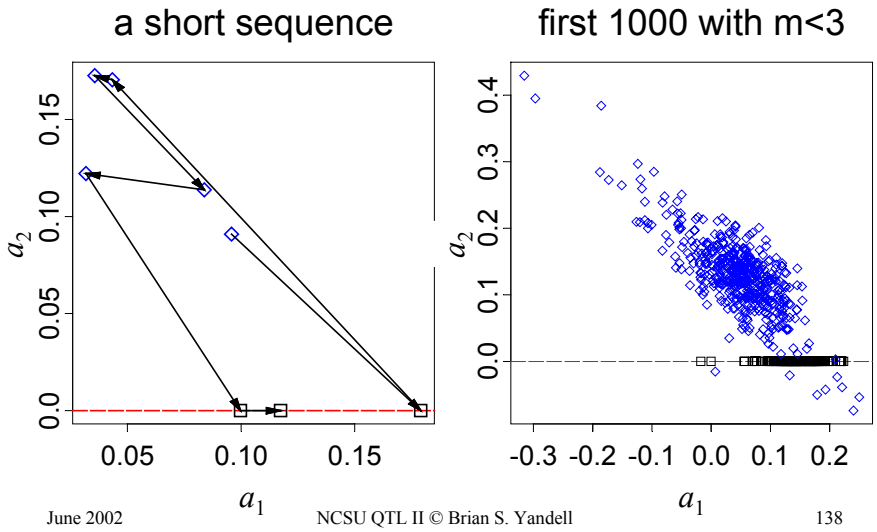
$$\left| \det\left(\begin{bmatrix} 1 & -c_{21}J \\ 0 & J \end{bmatrix} \right) \right| = \left| 1 \times J - 0 \times (-c_{21}J) \right| = J$$

$$da_1 da_2 = \left| \det\left(\frac{\partial g(\mu, a, \sigma^2; z)}{\partial a \partial z} \right) \right| da_1 da_2 = J da dz$$

geometry of reversible jump



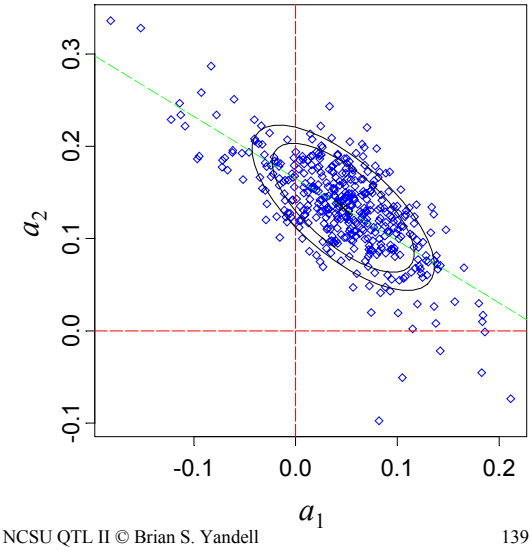
QT additive reversible jump



credible set for additive

90% & 95% sets based on normal

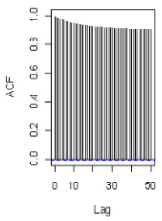
regression line corresponds to slope of updates



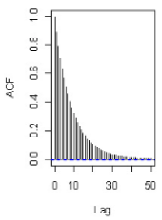
multivariate updating of effects

- more computations when $m > 2$
- avoid matrix inverse
  - Cholesky decomposition of matrix
- simultaneous updates
  - effects at all loci
- accept new locus based on
  - sampled new genos at locus
  - sampled new effects at all loci
- also long-range positions updates

before



after



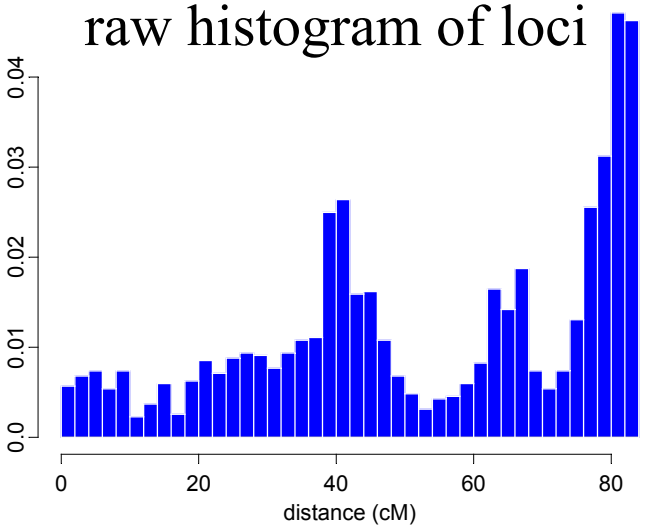
Part VII: Model Assessment

- posterior histograms
  - number of QTL
    - effect of prior on posterior
  - genetic architecture across genome
  - model averaging: loci and effects
- Bayes factors
  - What are Bayes factors?
  - Bayes factors from RJ-MCMC chain
  - components of Bayes factors

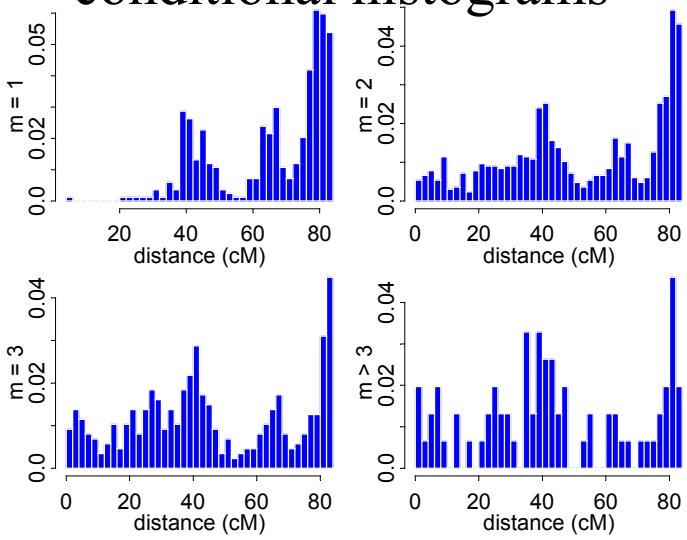
model selection and assessment

- how many QTL are supported by data?
  - parsimony: balance model fit to model complexity
- posterior as tool for model insight
  - number and chromosome distribution of QTL
  - loci and effects across genome
  - heritability, variance and LOD by number of QTL
- Bayes factors
  - advantages
    - related to likelihood ratio test
    - modest sensitivity to prior
    - intuitive interpretation
  - disadvantages/criticisms
    - do not work for improper priors
    - theoretical problems (Gelfand Day 1994; Draper 1995)

8-week vernalization raw histogram of loci



conditional histograms



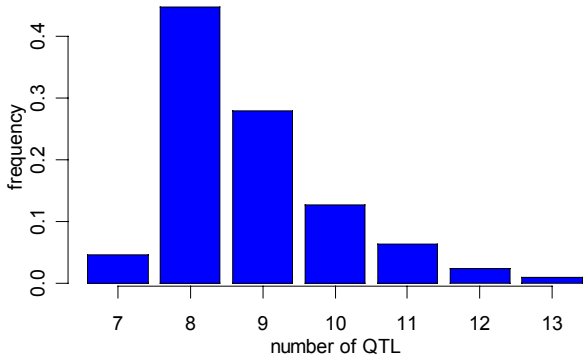
8 QTL simulation (Stevens Fisch 1998)

- $n = 200$, $h^2 = .5$
  - SF detected 3 QTL
- Bayesian IM

n	h^2	detect
200	.5	2
200	.8	4
500	.9	7
500	.97	8

QTL No. j	Location, λ_j			Additive effect α_j	Dominance Effect δ_j
	Chrom. λ_j^c	Marker λ_j^m	Position (cM)		
1	1	1	11	-3	0
2	1	3	10	-5	0
3	3	4	2	2	0
4	6	6	7	-3	0
5	6	8	12	3	0
6	8	2	12	-4	0
7	8	3	14	1	0
8	9	10	15	2	0

posterior number of QTL

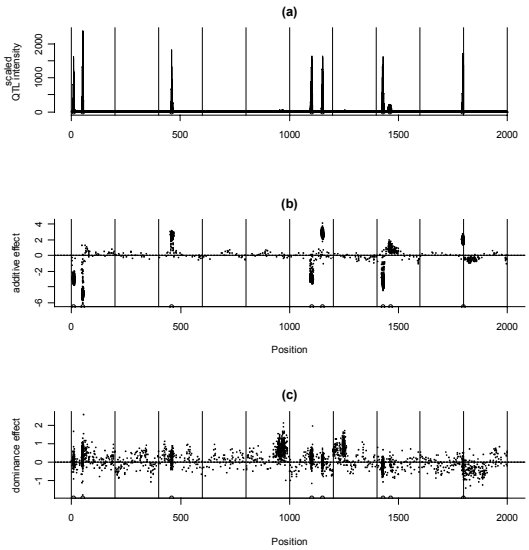


geometric prior with mean 0.5
seems to have no influence on posterior here

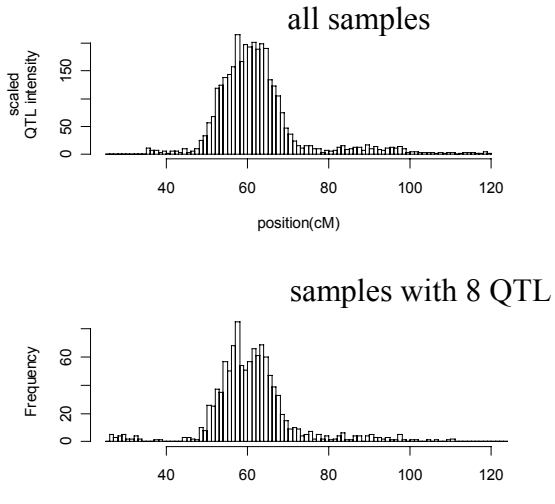
posterior genetic architecture

	Chromosome count vector										
m	1	2	3	4	5	6	7	8	9	10	Count
8	2	0	1	0	0	2	0	2	1	0	3371
9	*3*	0	1	0	0	2	0	2	1	0	751
7	2	0	1	0	0	2	0	*1*	1	0	377
9	2	0	1	0	0	*3*	0	2	1	0	218
9	2	0	1	0	0	2	0	2	*2*	0	198

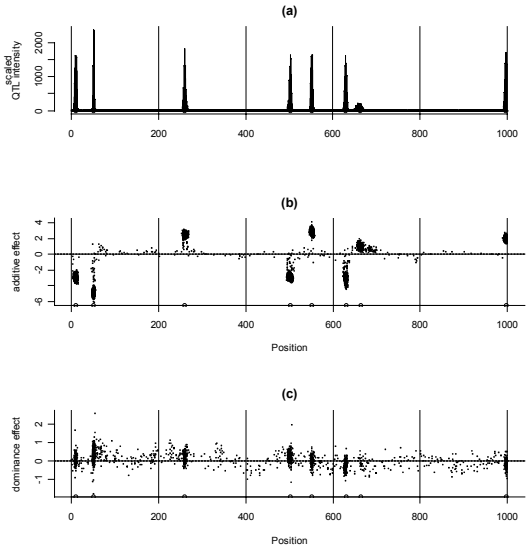
model averaging for 8 QTL



model averaging: focus on chr. 8



focus on key chromosomes



B. napus 8-week vernalization whole genome study

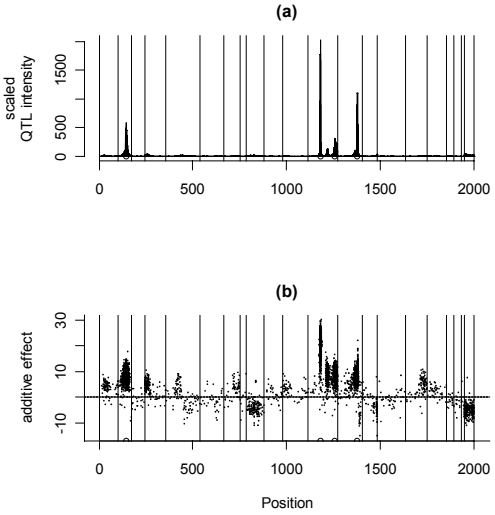
chrom	position	LOD	effect
n2	66.4	25.87	21.3
n3	106.8	13.33	12.95
n10	43.3	13.14	12.77
n2	154.0	10.69	11.3
n13	126.7	32.4	-5.78

Table 8.5: Result of CIM analysis for *B. napus dataset.*

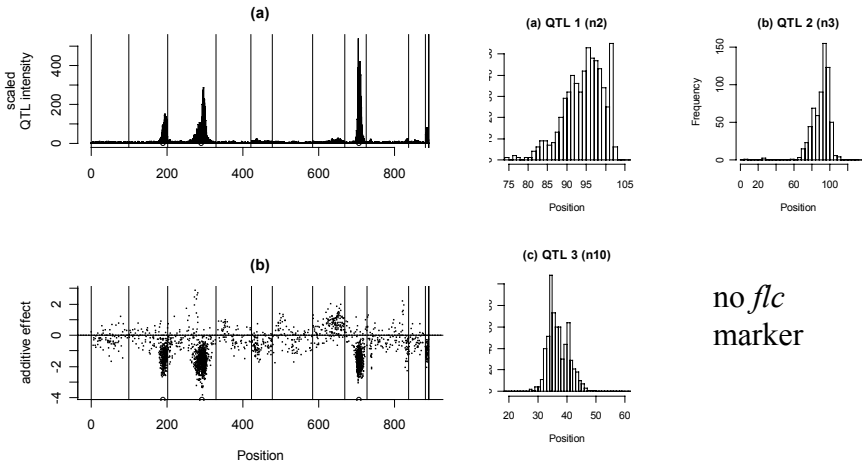
chrom	position	effect
n10	45.0	9.24
n2	66.9	22.4
n2	142.6	9.01
n3	103.4	8.36

Table 8.6: Estimates of QTL location and effect using BIM.

8-week vern: model averaging

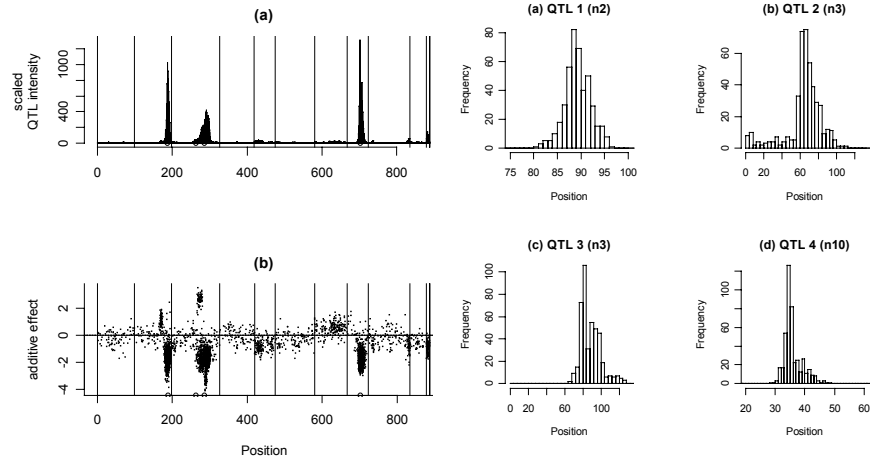


B. rapa: effect of *flc* marker
Kole et al. (1997)



no *flc* marker

B. rapa with added *flc* marker



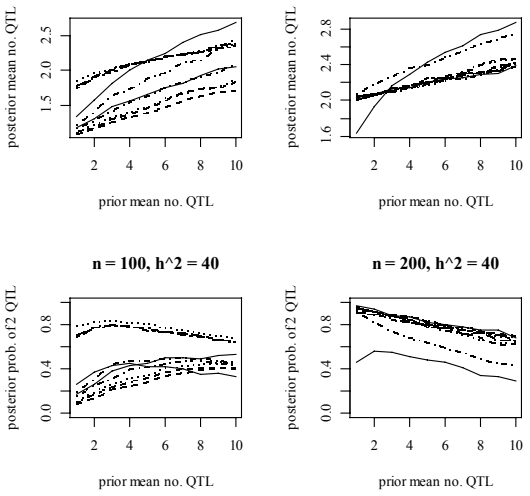
How sensitive is posterior to choice of prior?

- 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

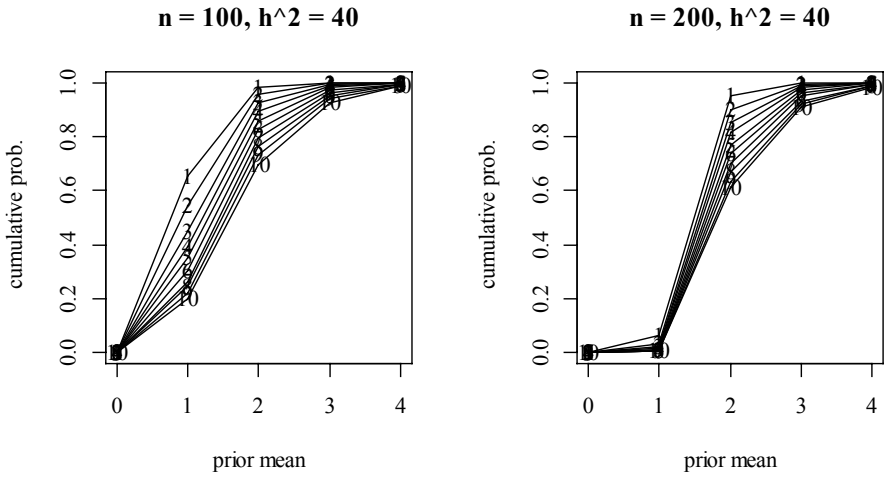
simulation study: prior

- 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

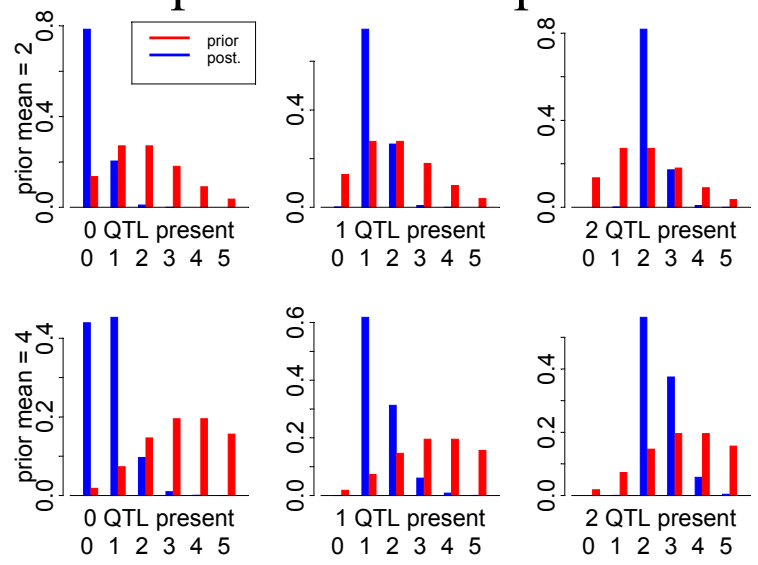
posterior *m* depends on prior



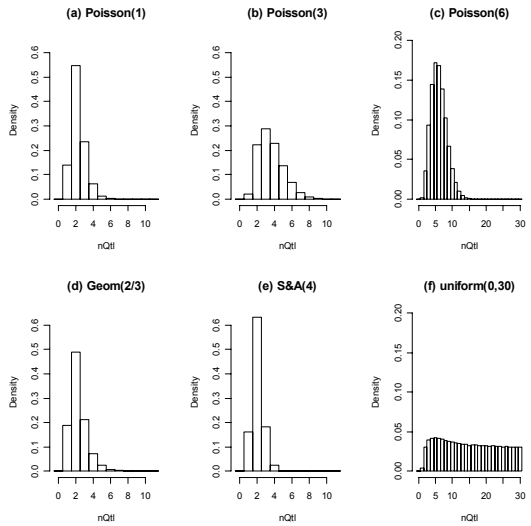
cumulative posterior as prior mean changes



effect of prior mean on posterior *m*



prior effect on posterior



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

$$B_{12} = \frac{\text{pr}(\text{model}_1 \mid Y)/\text{pr}(\text{model}_2 \mid Y)}{\text{pr}(\text{model}_1)/\text{pr}(\text{model}_2)} = \frac{\text{pr}(Y \mid \text{model}_1)}{\text{pr}(Y \mid \text{model}_2)}$$

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

computing marginal means

$$\text{pr}(Y \mid \text{model}_k) = \int \text{pr}(Y \mid \theta_k, \text{model}_k)\text{pr}(\theta_k \mid \text{model}_k)d\theta_k$$

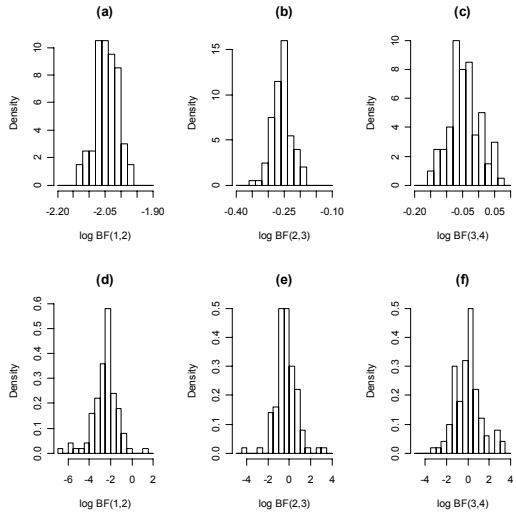
- very difficult based on separate model runs
  - run MCMC for model k
  - average pr(Y|θ_k) across model parameters θ_k
    - arithmetic mean
      - can be inefficient if prior differs from posterior
    - weighted harmonic mean
      - more efficient but less stable
    - stabilized harmonic mean (SHM)
      - average over "nuisance parameters" (*e.g.* variance)
      - more work, but estimate is more stable (Satagopan et al. 2000)
- easy when model itself is a parameter
  - RJ-MCMC: marginal summaries of number of QTL
  - posterior/prior provides Bayes factor yardstick

estimating Bayes factors

combined RJ-MCMC
posterior freq/prior

8-wk vern data
100 MCMC repeats
(note different scales)

separate MCMCs
stabilized
harmonic mean



Bayes factors & likelihood ratio

- 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*

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

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
  - geometric marginal means for pr(*Y*|model)
  - log posterior odds (LPD; see Sen Churchill 2000)

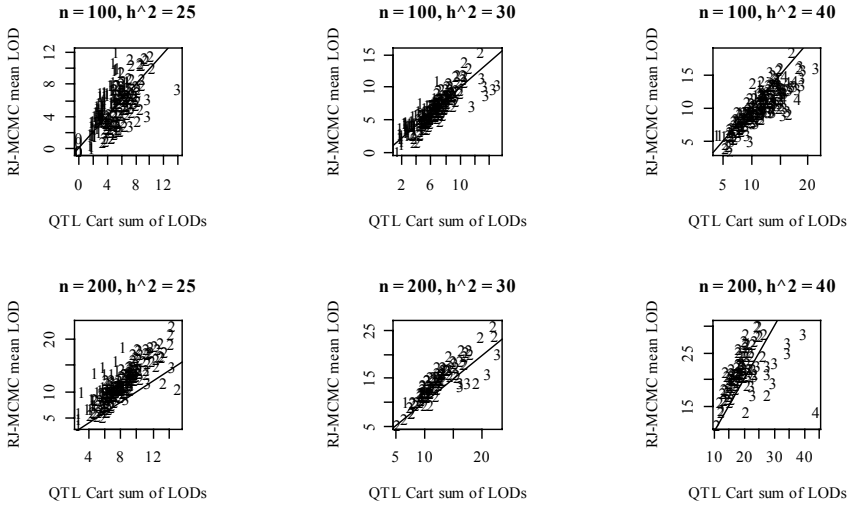
$$LOD(\lambda) = (\log_{10} e)\sum_{i=1}^{n} \ln\left(\frac{\sum_{Q} \Pr(Y_i \mid Q;\hat{\theta})\Pr(Q \mid X_i,\lambda)}{\Pr(Y_i \mid \tilde{\theta})} \right)$$

$$BLOD = (\log_{10} e)\sum_{i=1}^{n} \ln\left(\frac{\Pr(Y_i \mid Q_i,\theta)}{\Pr(Y_i \mid \tilde{\theta})} \right)$$

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



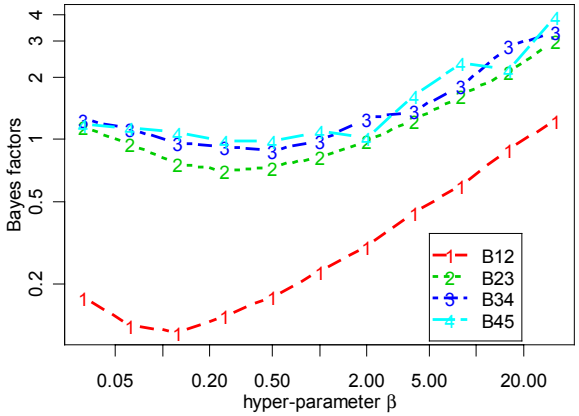
prior sensitivity of Bayes factor

- adjust effects prior as *m* grows
  - otherwise prior dominates for large models
- BF sensitive to fixed prior for effects
  - use hyperprior to soften effect

$$Y_i = \mu + \sum_{r=1}^{m} a_r Q_{ri} + e_i$$

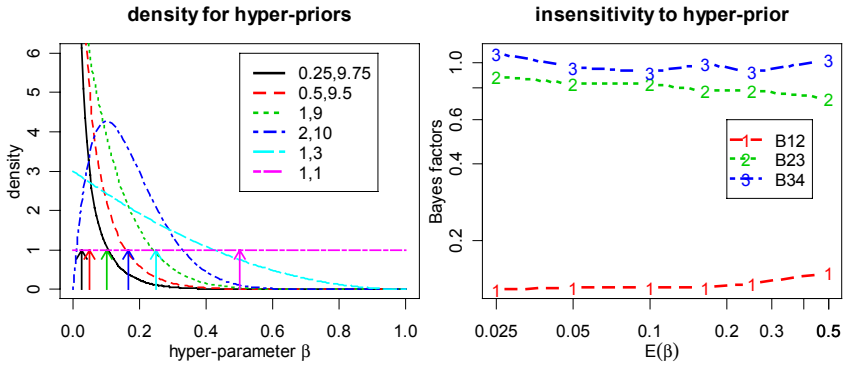
$$a_r \sim \mathrm{N}(0, 2\beta s^2 / m), \beta \sim \mathrm{Beta}(?, ?)$$

BF sensitivity to fixed prior for effects



Bayes factors for 8-week vernalization data

BF insensitivity to random effects prior



$$a_r \sim \mathrm{N}(0, 2\beta s^2/m), \beta \sim \mathrm{Beta}(?,?)$$

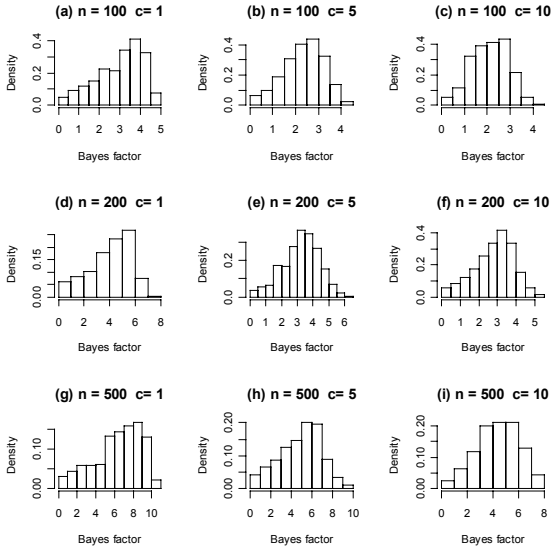
BF approximately invariant to form of prior

Prior, pr(*m*)	B_{12}	B_{23}	B_{34}	B_{45}
Geometric(2/3)	0.129	0.773	0.954	1.019
Poisson(1)	0.128	0.775	0.941	1.013
Poisson(3)	0.130	0.766	0.954	1.003
Poisson(6)	0.132	0.775	0.963	1.009
Fast-decay poisson(1)	0.128	0.764	0.941	1.022
Fast-decay Poisson(4)	0.129	0.773	9.963	1.032
Uniform	0.133	0.774	0.960	0.99

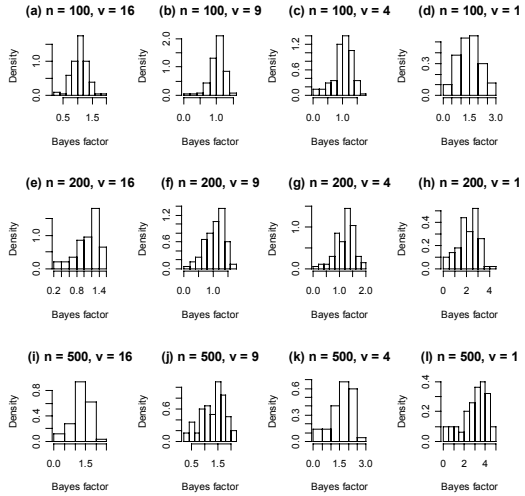
power study of Bayes factor

- large B_{01} favors model with 1 QTL
- n = 100, 200, 500
- size of genome
  - c = 1, 5, 10 number of chromosomes
- environmental variance
  - V = 1, 4, 9 with effect of size 1
- 11 markers per chromosome, 10cM apart
- 100 independent trials

B_{01} "power" vs. n & genome size c



B_{12} "size of test" vs. n & variance v



Part IX: Software & References

- MCMC software
  - general
  - specific to QTL problems
- References
  - MCMC, reversible jump
  - Bayesian approach
  - Bayesian QTL analysis

RJ-MCMC software

- General MCMC software
  - U Bristol links
    - www.stats.bris.ac.uk/MCMC/pages/links.html
  - BUGS (Bayesian inference Using Gibbs Sampling)
    - www.mrc-bsu.cam.ac.uk/bugs/
- MCMC software for QTLs
  - Bmapqtl (Satagopan Yandell 1996; Gaffney 2001)
    - www.stat.wisc.edu/~yandell/qtl/software/Bmapqtl
  - Bayesian QTL / Multimapper (Sillanpää Arjas 1998)
    - www.rni.helsinki.fi/~mjs
  - Yi, Xu (shxu@citrus.ucr.edu)
  - Stephens, Fisch (email)

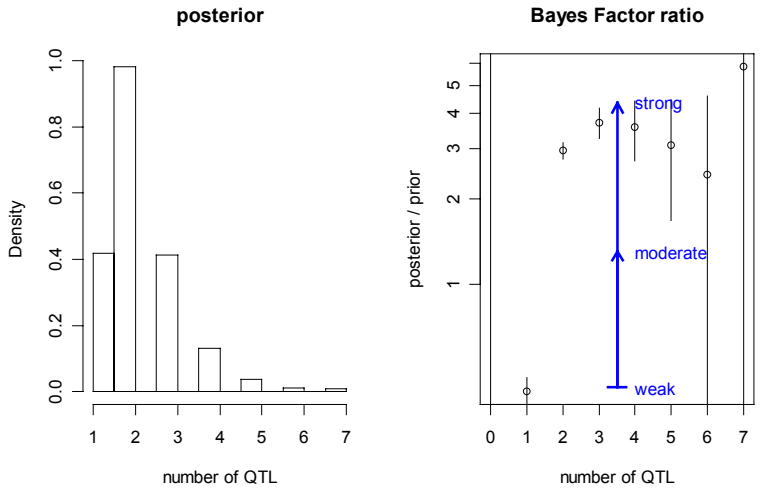
Bmapqtl: our RJ-MCMC software

- www.stat.wisc.edu/~yandell/qtl/software/Bmapqtl
  - module using QtlCart format
  - compiled in C for Windows/NT
  - extensions in progress
  - R post-processing graphics (Yandell)
- Bayes factor and reversible jump MCMC computation
- enhances MCMCQTL and revjump software
  - 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

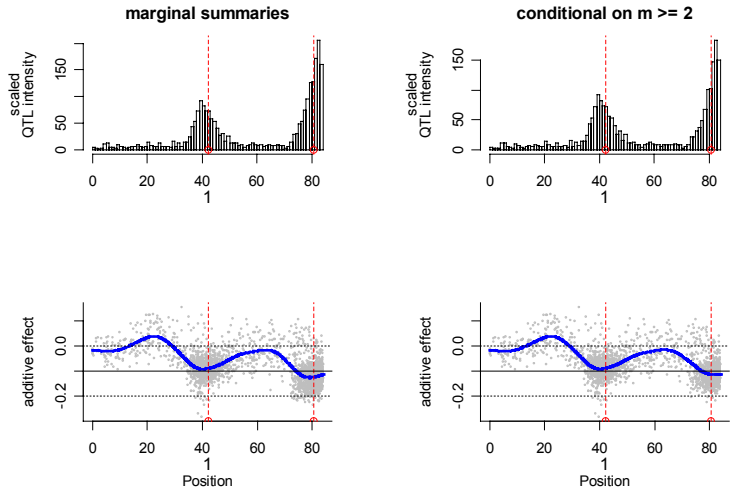
R library(Bmapqtl)

```
> library(Bmapqtl)
> Bmapqtl.plot( vern8.exp, "exp", nqtl = 2 )
posterior for number of QTL
   1    2    3    4    5    6    7
0.21 0.49 0.21 0.07 0.02 0.00 0.00
Loading required package: modreg
2 estimated loci: 42.22 80.64
marginal heritability 0.356
conditional heritability
    1     2     3     4     5     6     7
0.325 0.360 0.373 0.363 0.361 0.368 0.357
marginal LOD 9.726
conditional LOD
     1      2      3      4      5      6      7
 8.576  9.933 10.193 10.079 10.233 10.823  9.489
```

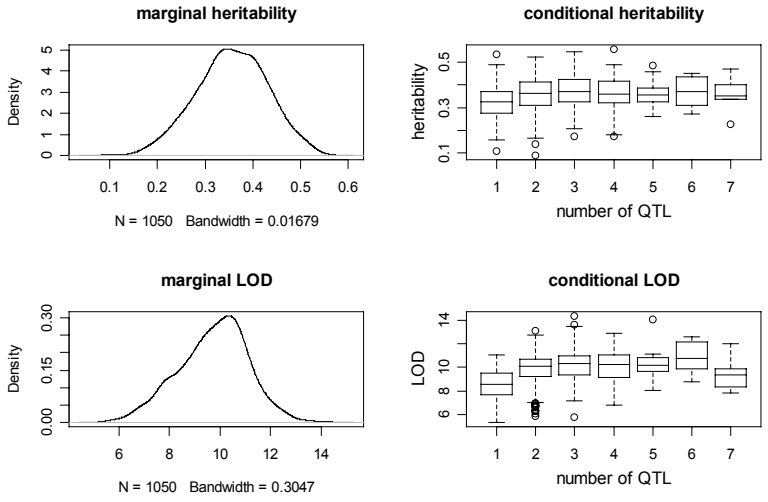
Bmapqtl.nqtl(data, prior)



Bmapqtl.loci(data, nqtl)



Bmapqtl.hist(data, histograms, mains)



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)
  - model averaging

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 Raftery (2000) Easy estimation of normalizing constants and Bayes factors from posterior simulation: Stabilizing the harmonic mean estimator. Technical Report 1028, Department of Statistics, University of Wisconsin.

reversible jump MCMC references

- PJ Green (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* 82: 711-732.
- L Kuo & B Mallick (1998) Variable selection for regression models, *Sankhya B* 60: 65-81.
- BK Mallick & AE Gelfand (1994) Generalized linear models with unknown number of components, *Biometrika* 81: 237-245.
- S Richardson & PJ Green (1997) On Bayesian analysis of mixture with an unknown of components, *J Royal Statist Soc B* 59: 731-792.

QTL reversible jump MCMC: inbred lines

- Gaffney PJ (2001) An efficient reversible jump Markov chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses. PhD dissertation, Department of Statistics, UW-Madison.
- 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 Sillanpää, 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.
- N Yi, S Xu (2002) Mapping quantitative trait loci with epistatic effects. *Genet. Res. Camb*. 00: 000-000.

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.

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.

Part X: Combining Multiple Crosses

- combining inbred lines in search of QTL
  - most IM methods limited to single cross
  - animal model largely focuses on polygenes
  - individuals no longer independent given Q
- recent work in plant sciences
  - Bernardo (1994) Wright's relationship matrix A
  - Rebai *et al.* (1994) regression method
  - Xu, Atchley (1995) IBD & A for QTLs & polygenes
  - Liu, Zeng (2000) multiple inbred lines, fixed effect IM
  - Zou, Yandell, Fine (2001 *Genetics*) power, threshold
  - Yi, Xu (2002) *Genetica*

thresholds for multiple crosses

- permutation test
  - Churchill Doerge (1994); Doerge Churchill (1996)
  - computationally intensive
  - difficult to compare different designs
- theoretical approximation
  - Lander Botstein (1989) Dupuis Siegmund (1999)
    - single cross, dense linkage map
  - Rebai *et al.* (1994, 1995) approximate extension
    - Piepho (2001) improved calculation of efficiency
  - Zou, Yandell, Fine (2001) extend original theory

extension of threshold theory

- likelihood for multiple crosses of inbred lines with m founders
  - approximately χ^2 with m degrees of freedom
  - genome-wide threshold theory
    - extends naturally based on Ornstein-Uhlenbeck
  - threshold based on dense or sparse linkage map
- some calculations based on BC1, F2, BC2
  - Liu Zeng (2000) ECM method to estimate $Y_j \sim \text{Normal}(G_{Qj}, \sigma_j^2), j = \text{cross}$

literature on outbred studies

- Interval Mapping for Outbred Populations
  - Haley, Knott & Elsen (1994) *Genetics*
  - Thomas & Cortessis (1992) *Hum. Hered.*
  - Hoeschele & vanRanden (1993ab) *Theor. Appl. Genet.* (etc.)
  - Guo & Thompson (1994) *Biometrics*
- Experimental Outbred Crosses (BC, F2, RI)
  - collapse markers from 4 to 2 alleles
- Multiple Cross Pedigrees
  - polygenic effects not modeled here
  - related individuals are correlated (via coancestry)
  - Liu & Zeng (2000) *Genetics*
  - Zou, Yandell & Fine (2001) *Genetics*

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.

many thanks

Michael Newton
Daniel Sorensen
Daniel Gianola
Jaya Satagopan
Patrick Gaffney
Fei Zou
Liang Li
Yang Song
Chunfang Jin

Tom Osborn
David Butruille
Marcio Ferrera
Josh Udahl
Pablo Quijada
Alan Attie
Jonathan Stoehr
Hong Lan

USDA Hatch Grants