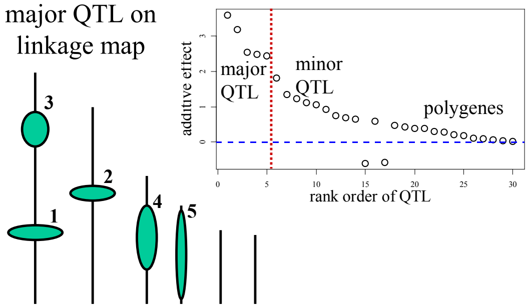
Bayesian Model Selection for Quantitative Trait Loci using Markov chain Monte Carlo in Experimental Crosses

Brian S. Yandell
University of Wisconsin-Madison
www.stat.wisc.edu/~yandell/statgen
with Chunfang “Amy” Jin, UW-Madison,
Patrick J. Gaffney, Lubrizol,
and Jaya M. Satagopan, Sloan-Kettering

Jackson Laboratory, September 2002

Pareto diagram of QTL effects

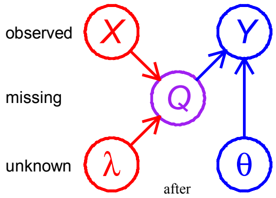


how many (detectable) QTL?

- build m = number of QTL detected into model
  - directly allow uncertainty in genetic architecture
  - model selection over number of QTL, architecture
  - use Bayes factors and model averaging
    - to identify “better” models
- many, many QTL may affect most any trait
  - how many QTL are detectable with these data?
    - limits to useful detection (Bernardo 2000)
    - depends on sample size, heritability, environmental variation
  - consider probability that a QTL is in the model
    - avoid sharp in/out dichotomy
    - major QTL usually selected, minor QTL sampled infrequently

interval mapping basics

- observed measurements
  - Y = phenotypic trait
  - X = markers & linkage map
    - i = individual index $1,\ldots,n$
- 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
  - m = number of QTL
- $\text{pr}(Q|X,\lambda,m)$ recombination model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for Q given X
- $\text{pr}(Y|Q,\theta,m)$ phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters θ (could be non-parametric)

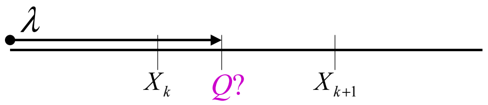


after
Sen Churchill (2001)

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

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

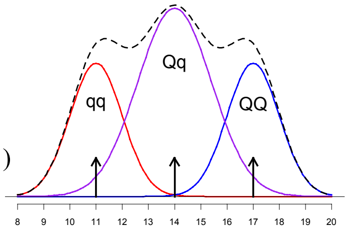


idealized phenotype model

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

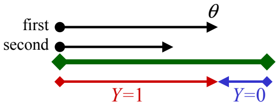
$$Y = G_Q + E$$

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



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*
  - 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 left (right)?



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

what is Bayes theorem?

- before and after observing data
  - prior: $\mathrm{pr}(\theta)$ = pr(parameters)
  - posterior: $\mathrm{pr}(\theta|Y)$ = 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

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

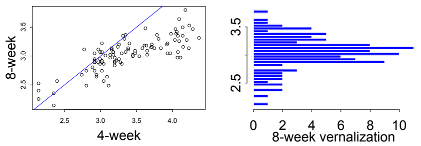
Bayesian interval mapping

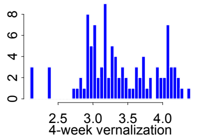
- likelihood is mixture over genotypes Q

  $L(\lambda|Y) = \mathrm{product}_i\ [\mathrm{sum}_Q\ \mathrm{pr}(Q|X_i,\lambda)\ \mathrm{pr}(Y_i|Q,\theta)]$
- Bayesian posterior includes Q as missing data
  - sample unknown data instead of averaging
    - sample unknown genotypes Q
    - prior on unknown loci λ and effects θ of interest

  $\mathrm{pr}(\lambda,Q,\theta|Y,X) = [\mathrm{product}_i\ \mathrm{pr}(Q_i|X_i,\lambda)\ \mathrm{pr}(Y_i|Q_i,\theta)]\ \mathrm{pr}(\lambda,\theta|X)$
  - marginal summaries provide key information
    - loci: $\mathrm{pr}(\lambda|Y,X) = \mathrm{sum}_{Q,\theta}\ \mathrm{pr}(\lambda,Q,\theta|Y,X)$
    - effects: $\mathrm{pr}(\theta|Y,X) = \mathrm{sum}_{Q,\lambda}\ \mathrm{pr}(\lambda,Q,\theta|Y,X)$

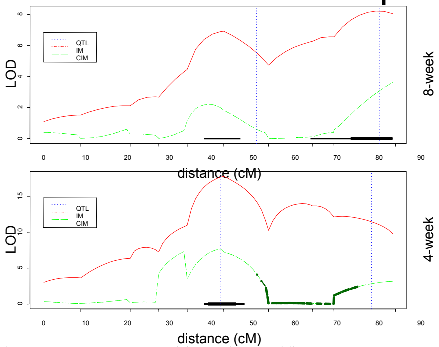
Brassica 4- & 8-week Data



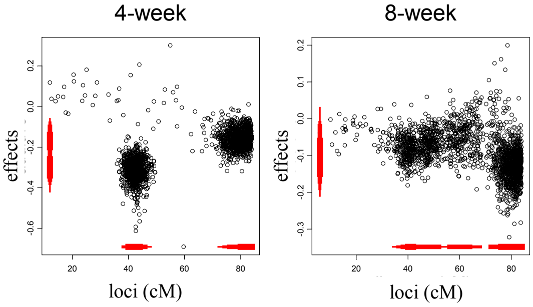


- •4-week & 8-week vernalization
  - –log(days to flower)
- •genetic cross of
  - –Stellar (annual canola)
  - –Major (biennial rapeseed)
- •105 double haploid (DH) lines
  - –homozygous at every locus
- •10 markers on chromosome N2

Brassica Data LOD Maps



Bayesian samples for *Brassica*



multiple QTL phenotype model

- phenotype influenced by genotype & environment

  pr($Y|Q,\theta$) ~ $N(G_Q, \sigma^2)$, or $Y = G_Q$ + environment
- partition mean into separate QTL effects

  G_Q = mean + main effects + epistatic interactions

  $G_Q = \mu + \theta_{1Q} + \ldots + \theta_{mQ} + \theta_{12Q} + \ldots$
- priors on mean and effects

$G_Q \sim N(\mu_0, \kappa\sigma^2)$	model independent genotypic value
$\mu \sim N(\mu_0, \kappa_0\sigma^2)$	grand mean
$\theta_{jQ} \sim N(0, \kappa_1\sigma^2/m)$	effects down-weighted by m
$\theta_{j2Q} \sim N(0, \kappa_2\sigma^2/m_2)$	interactions down-weighted by m_2

- determine hyper-parameters via Empirical Bayes

$$\mu_0 \approx \overline{Y}, \kappa - \kappa_0 \approx \frac{h^2}{1-h^2} = \frac{\sigma_G^2}{\sigma^2}, \kappa = \kappa_0 + \kappa_1 + \kappa_2$$

phenotype posterior mean

- phenotype influenced by genotype & environment

  pr($Y|Q,\theta$) ~ $N(G_Q, \sigma^2)$, or $Y = G_Q$ + environment
- relation of posterior mean to LS estimate

$$G_Q \mid Y, m \sim N(\mu_0 + B_Q(\hat{G}_Q - \mu_0), B_Q C_Q \sigma^2)$$

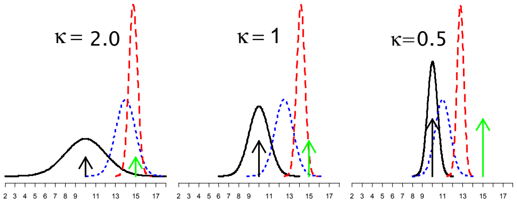
$$\approx N(\hat{G}_Q, C_Q\sigma^2)$$

LS estimate $\hat{G}_Q = \hat{\mu} + \sum_i \sum_j \hat{\theta}_{ijQ} = \sum_i w_{iQ} Y_i$

variance $V(\hat{G}_Q) = \sum_i w_{iQ}^2 \sigma^2 = C_Q \sigma^2$

shrinkage $B_Q = \kappa / (\kappa + C_Q) \to 1$

effect of prior variance on posterior



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

prior & posterior for genotypes *Q*

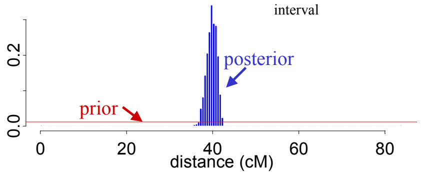
- prior is recombination model

  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

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

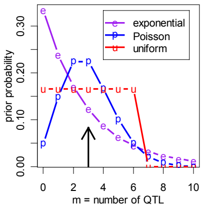
prior & posterior for QT locus

- 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



prior & posterior on number of QTL

- what prior on number of QTL?
  - uniform over some range
  - Poisson with prior mean
  - geometric with prior mean
- prior influences posterior
  - good: reflects prior belief
    - push data in discovery process
  - bad: skeptic revolts!
    - “answer” depends on “guess”

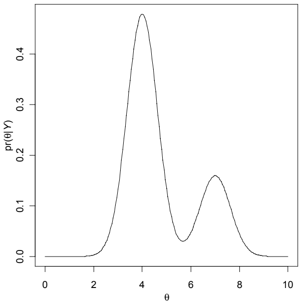


Markov chain Monte Carlo idea

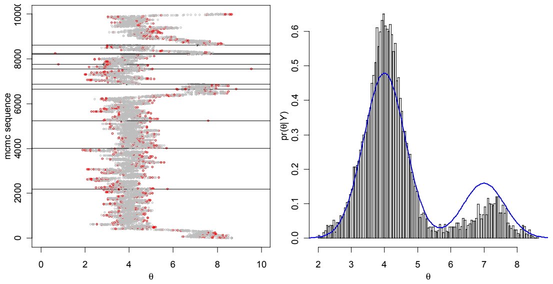
have posterior pr($\theta | Y$)
want to draw samples

propose $\theta \sim$ pr($\theta | Y$)
(ideal: Gibbs sample)

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



MCMC realization



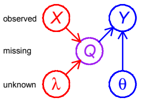
added twist: occasionally propose from whole domain

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 *m*-QTL model 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

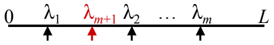
$$(\lambda, Q, \theta, m) \sim \text{pr}(\lambda, Q, \theta, m \mid Y, X)$$
$$(\lambda, Q, \theta, m)_1 \to (\lambda, Q, \theta, m)_2 \to \cdots \to (\lambda, Q, \theta, m)_N$$

sample from full conditionals for model with *m* QTL



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

reversible jump MCMC



action steps: draw one of three choices

- update *m*-QTL model with probability 1-*b*(*m*+1)-*d*(*m*)
  - update current model using full conditionals
  - sample *m* QTL loci, effects, and genotypes
- add a locus with probability *b*(*m*+1)
  - propose a new locus along genome
  - innovate new genotypes at locus and phenotype effect
  - decide whether to accept the "birth" of new locus
- drop a locus with probability *d*(*m*)
  - propose dropping one of existing loci
  - decide whether to accept the "death" of locus

sampling the number of QTL

- use reversible jump MCMC to change *m*
  - bookkeeping helps in comparing models
  - adjust to change of variables between models
  - Green (1995); Richardson Green (1997)
  - other approaches out there these days…
- think model selection in multiple regression
  - but regressors (QT genotypes) are unknown
  - linked loci = collinear regressors = correlated effects
  - consider additive effects with coding Q_{ij}= -1,0,1

$$\theta_{ijQ} = \alpha_j (Q_{ij} - \bar{Q}_j)$$

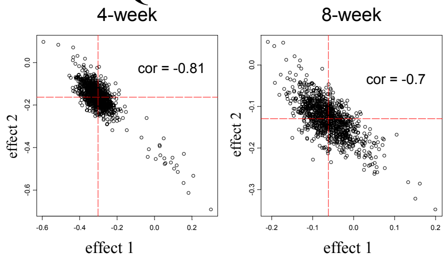
Model Selection in Regression

- consider known genotypes (Q)
  - models with 1 or 2 QTL at known loci
- jump between 1-QTL and 2-QTL models
  - adjust posteriors when model changes
  - due to collinearity of QTL genotypes

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

$$m = 2 : Y_i = \mu + \alpha_1(Q_{i1} - \overline{Q}_1) + \alpha_2(Q_{i1} - \overline{Q}_1) + e_i$$

collinear QTL = correlated effects

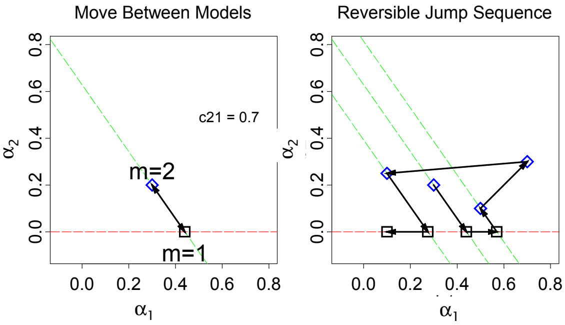


•linked QTL: collinear genotypes & correlated effect estimates

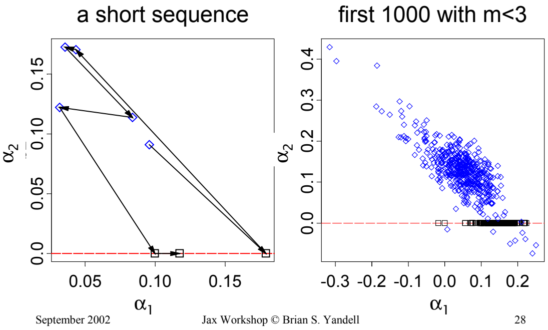
–sum of linked effects usually well determined

•which QTL to go after in breeding, genome walking?

Geometry of Reversible Jump



QT additive Reversible Jump



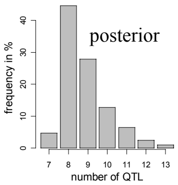
a complicated simulation

•simulated F2 intercross, 8 QTL

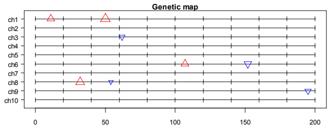
- (Stephens, Fisch 1998)
- n=200, heritability = 50%
- detected 3 QTL

•increase to detect all 8

- n=500, heritability to 97%



QTL	chr	loci	effect
1	1	11	–3
2	1	50	–5
3	3	62	+2
4	6	107	–3
5	6	152	+3
6	8	32	–4
7	8	54	+1
8	9	195	+2



loci pattern across genome

- notice which chromosomes have persistent loci
- best pattern found 42% of the time

Chromosome

m	**1**	2	**3**	4	5	**6**	7	**8**	**9**	10	**Count of 8000**
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	2	0	2	1	0	218
9	2	0	1	0	0	*3*	0	2	1	0	218
9	2	0	1	0	0	2	0	2	*2*	0	198

Bayes factors to assess models

- Bayes factor: which model best supports the data?
  - ratio of posterior odds to prior odds
  - ratio of model likelihoods
- equivalent to *LR* statistic when
  - comparing two nested models
  - simple hypotheses (e.g. 1 vs 2 QTL)
- Bayes Information Criteria (BIC)
  - Schwartz introduced for model selection in general settings
  - penalty to balance model size (p = number of parameters)

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

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

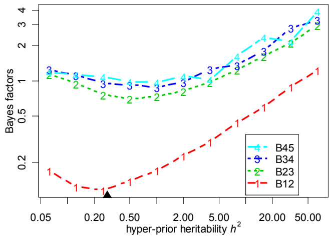
QTL Bayes factors & RJ-MCMC

- easy to compute Bayes factors from samples
  - posterior pr(m|Y,X) is marginal histogram
  - posterior affected by prior pr(m)

$$BF_{m,m+1} = \frac{\mathrm{pr}(m \mid Y, X) / \mathrm{pr}(m)}{\mathrm{pr}(m+1 \mid Y, X) / \mathrm{pr}(m+1)}$$

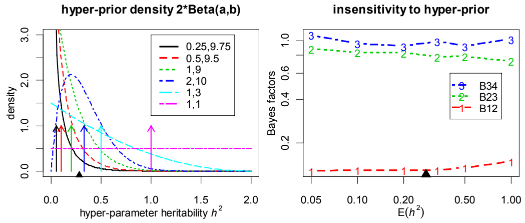
- *BF* insensitive to shape of prior
  - geometric, Poisson, uniform
  - precision improves when prior mimics posterior
- *BF* sensitivity to prior variance on effects θ
  - prior variance should reflect data variability
  - resolved by using hyper-priors
    - automatic algorithm; no need for tuning by user

BF sensitivity to fixed prior for effects



$$\theta_{jQ} \sim \mathrm{N}\left(0, \kappa_1\sigma^2 / m\right), \kappa_1\sigma^2 = h^2\sigma^2_{\text{total}}, h^2 \text{ fixed}$$

BF insensitivity to random effects prior



$$\theta_{jQ} \sim \mathrm{N}\left(0, \kappa_1\sigma^2 / m\right),\ \kappa_1\sigma^2 = h^2\sigma^2_{\text{total}},\ \frac{h^2}{2} \sim \mathrm{Beta}(a,b)$$

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
    - library(bim) is cross-compatible with library(qtl)
- 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

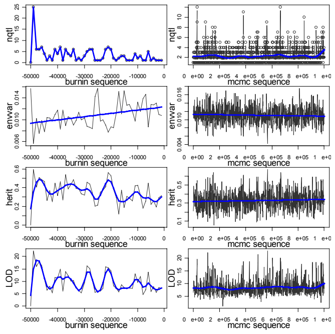
B. napus 8-week vernalization whole genome study

- 108 plants from double haploid
  - similar genetics to backcross: follow 1 gamete
  - parents are Major (biennial) and Stellar (annual)
- 300 markers across genome
  - 19 chromosomes
  - average 6cM between markers
    - median 3.8cM, max 34cM
  - 83% markers genotyped
- phenotype is days to flowering
  - after 8 weeks of vernalization (cooling)
  - Stellar parent requires vernalization to flower

Markov chain Monte Carlo sequence

burnin (sets up chain)
mcmc sequence

number of QTL
environmental variance
h^2 = heritability
(genetic/total variance)
LOD = likelihood

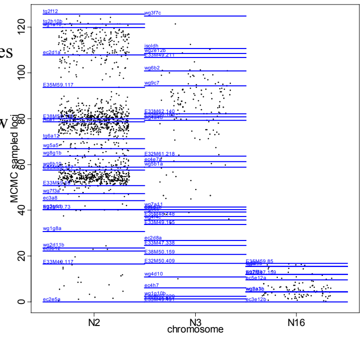


MCMC sampled loci

subset of chromosomes
N2, N3, N16

points jittered for view
blue lines at markers

note concentration
on chromosome N2

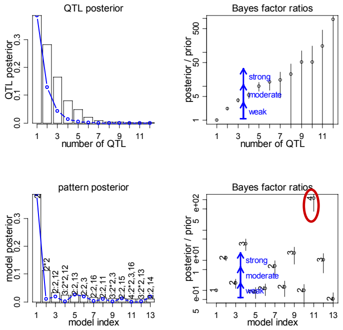


Bayesian model assessment

row 1: # QTL
row 2: pattern

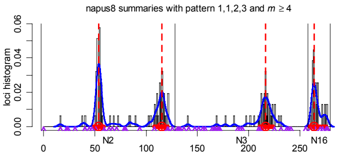
col 1: posterior
col 2: Bayes factor
note error bars on bf

evidence suggests
4-5 QTL
N2(2-3),N3,N16

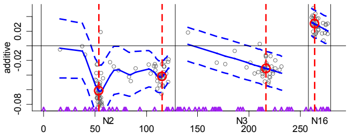


Bayesian estimates of loci & effects

histogram of loci
blue line is density
red lines at estimates



estimate additive effects
(red circles)
grey points sampled
from posterior
blue line is cubic spline
dashed line for 2 SD

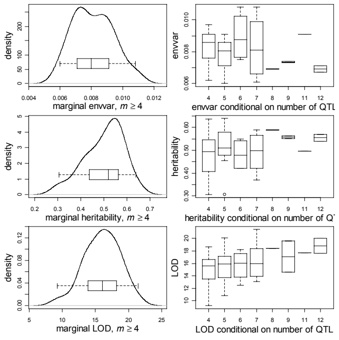


Bayesian model diagnostics

pattern: N2(2),N3,N16
col 1: density
col 2: boxplots by m

environmental variance
$\sigma^2 = .008$, $\sigma = .09$
heritability
$h^2 = 52\%$
LOD = 16
(highly significant)

but note change with m



some QTL references

- Bernardo R (2000) What if we knew all the genes for a quantitative trait in hybrid crops? *Crop Sci.* (submitted).
- Gaffney PJ (2001) An efficient reversible jump Markov chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses. *PhD thesis, UW-Madison Statistics*.
- Heath S (1997) Markov chain Monte Carlo segregation and linkage analysis for oligenic models, *Am J Hum Genet* 61: 748-760.
- Satagopan JM, Yandell BS, Newton MA, Osborn TC (1996) A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* 144: 805-816.
- Satagopan JM, Yandell BS (1996) Estimating the number of quantitative trait loci via Bayesian model determination. *Proc JSM Biometrics Section*.

more QTL references

- Sillanpaa MJ, Arjas E (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data., *Genetics* 148: 1373-1388.
- Stephens DA, Fisch RD (1998) Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics* 54: 1334-1347.
- Uimari P and Hoeschele I (1997) Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms, *Genetics* 146: 735-743.
- Zou F, Fine JP, Yandell BS (2001) On empirical likelihood for a semiparametric mixture model. *Biometrika 00*: 000-000.
- Zou F, Yandell BS, Fine JP (2001) Threshold and power calculations for QTL analysis of combined lines. *Genetics 00*: 000-000.

reversible jump MCMC references

- Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82: 711-732.
- Kuo L, Mallick B (1998) Variable selection for regression models. *Sankhya, Series B, Indian J Statistics 60*: 65-81
- Mallick BK (1998) Bayesian curve estimation by polynomial of random order. *J Statistical Planning and Inference 70*: 91-109
- Richardson S, Green PJ (1997) On Bayesian analysis of mixture with an unknown of components. *J Royal Statist Soc B* 59: 731-792.

many thanks

Michael Newton	Tom Osborn
Daniel Sorensen	David Butruille
Daniel Gianola	Marcio Ferrera
Yang Song	Josh Udahl
Fei Zou	Pablo Quijada
Liang Li	Alan Attie
Hong Lan	Jonathan Stoehr

USDA Hatch Grants