Model Selection for Quantitative Trait Loci in Experimental Crosses

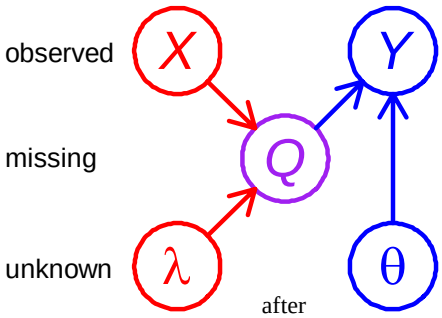
Brian S. Yandell
University of Wisconsin-Madison
www.stat.wisc.edu/~yandell/statgen
with Chunfang "Amy" Jin, UW-Madison,
Patrick J. Gaffney, Lubrizol,
and Jaya M. Satagopan, Sloan-Kettering
Jackson Laboratory, October 2003

Outline

- Non-normal phenotypes
  - Limitations of normal assumption
  - Quick fixes: transformations
  - Fancy approaches: semi-parametric "families"
  - Bottom line: normal OK for location, not effects
- Bayesian interval mapping
  - Multiple QTL model selection
  - Graphical diagnostic tools

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
  - m = number of QTL
- pr($Q|X,\lambda,m$) recombination model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for Q given X
- pr($Y|Q,\theta,m$) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters θ (could be non-parametric)

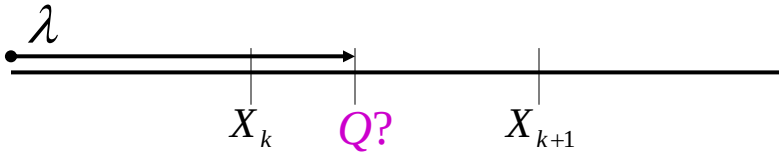


after
Sen Churchill (2001)

recombination model 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

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



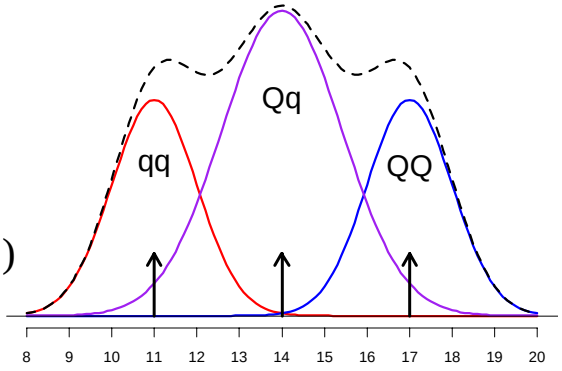
idealized phenotype model

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

$$Y = G_Q + E$$

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

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



limitations of normal assumption

- measurements not normal
  - categorical traits: counts (*e.g.* number of tumors)
    - use methods specific for counts
    - binomial, Poisson, negative binomial
  - traits measured over time and/or space
    - survival time (*e.g.* days to flowering)
    - developmental process; signal transduction between cells
    - TP Speed (pers. comm.); Ma, Casella, Wu (2002)
- false positives due to miss-specified model
  - how to check model assumptions?
- want more robust estimates of effects
  - parametric: only center (mean), spread (SD)
  - shape of distribution may be important

quick fixes: 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

QTL for binomial data

- approximate methods: marker regression
  - Zeng (1993,1994); Visscher et al. (1996); McIntyre et al. (2001)
- interval mapping, CIM
  - Xu Atchley (1996); Yi Xu (2000)
  - $Y \sim$ binomial(1,π), π depends on genotype Q
  - pr($Y|Q$) = $(\pi_Q)^Y (1 - \pi_Q)^{(1-Y)}$
  - substitute this phenotype model in EM iteration
- or just map it as another marker!
  - but may have complex

threshold or latent variable idea

- "real", unobserved phenotype *Z* is continuous
- observed phenotype *Y* is ordinal value
  - no/yes; poor/fair/good/excellent
  - pr($Y = j$) = pr($\tau_{j-1} < Z \leq \tau_j$)
  - pr($Y \leq j$) = pr($Z \leq \tau_j$)
- use logistic regression idea (Hackett Weller 1995)
  - substitute new phenotype model in to EM algorithm
  - or use Bayesian posterior approach
  - extended to multiple QTL (papers in press)

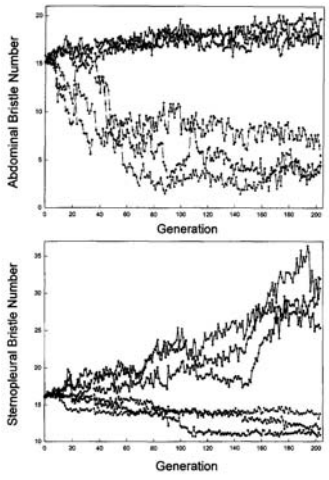
$$\text{pr}(Y \leq j \mid Q) = \text{pr}(Z \leq \tau_j \mid Q) = [1 + \exp(\mu + G_Q - \tau_j)]^{-1}$$

quantitative & qualitative traits

- Broman (2003): spike in phenotype
  - large fraction of phenotype has one value
  - map binary trait (is/is not that value)
  - map continuous trait given not that value
- multiple traits
  - Williams et al. (1999)
    - multiple binary & normal traits
    - variance component analysis
  - Corander Sillanpaa (2002)
    - multiple discrete & continuous traits
    - latent (unobserved) variables

other parametric approaches

- Poisson counts
  - Mackay Fry (1996)
    - trait = bristle number
  - Shepel et al (1998)
    - trait = tumor count
- negative binomial
  - Lan *et al.* (2001)
    - number of tumors
- exponential
  - Jansen (1992)



Mackay Fry (1996 *Genetics*)

marker density & sample size: 2 QTL

modest sample size
dense vs. sparse markers

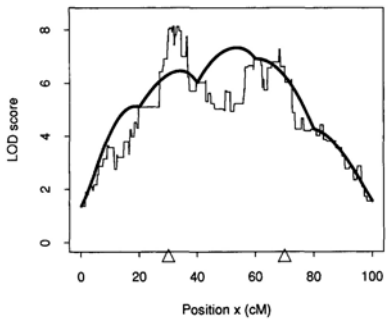


FIGURE 1.—The two-QTL true model with a QTL at 30 cM and a second QTL of somewhat smaller effect at 70 cM (true locations indicated by △). A normal single-QTL model is assumed and the LOD score for 100 simulated individuals is given for dense markers (thin curve) and markers at 20-cM intervals (bold curve).

Wright Kong (1997 *Genetics*)

large sample size
dense vs. sparse markers

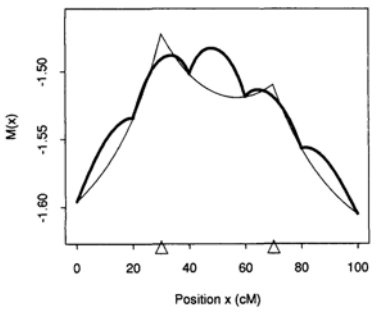


FIGURE 4.—$M(x)$ for a normal single-QTL assumed model under a two-QTL true model when both of the genes lie on the chromosome under study. This scenario was originally depicted in Figure 1. With dense markers (thin curve), $M(x)$ peaks at exactly 30 cM, the location of the QTL of stronger effect. With nondense markers at 20-cM intervals, $M(x)$ peaks at 47 cM in an incorrect interval (bold curve). Note the similarity in shape between the LODs in Figure 1 and the limiting forms depicted here.

robust locus estimate for non-normal phenotype

large sample size &
dense marker map:
no need for normality

but what happens for
modest sample sizes?

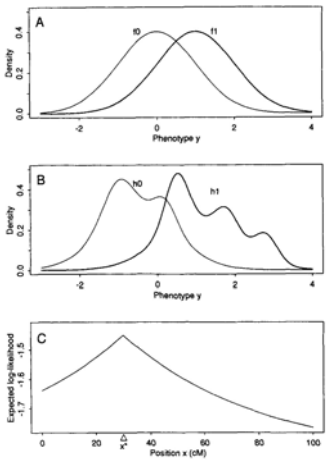


FIGURE 2.—Misspecification of the phenotype model. (A) The assumed distributions f_0 and f_1. (B) The true distributions h_0, h_1. (C) The expected log-likelihood across the chromosome when the markers are dense. Despite the misspecification, the function is maximized at exactly the true location $x^* = 30$ cM (indicated by △).

Wright Kong (1997 *Genetics*)

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 pr($Y|X,\lambda,\theta$) = sum$_Q$ pr($Q|X,\lambda$) pr($Y|Q,\theta$)
- what shape is your phenotype model?
  - parametric pr($Y|Q,\theta$) = $f(Y \mid \mu, G_Q, \sigma^2)$
  - semi-parametric pr($Y|Q,\theta$) = $f(Y)$ exp($Y\beta_Q$)
  - non-parametric pr($Y|Q,\theta$) = $F_Q(Y)$

semi-parametric QTL

- phenotype model $\text{pr}(Y|Q,\theta) = f(Y)\exp(Y\beta_Q)$
  - $f(Y)$ is some unknown distribution shape (density)
  - $\exp(Y\beta_Q)$ `tilts' f based on genotype Q and phenotype Y
- test for QTL at locus λ
  - $\beta = (\beta_{qq}, \beta_{Qq}, \beta_{QQ})$ unknown effect parameters
  - no QTL: $\beta_Q = 0$ for all Q, or $\text{pr}(Y|Q,\theta) = f(Y)$
- includes many standard phenotype models *without* having to choose among them

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

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

  exponential, binomial, …, but not negative binomial

semi-parametric empirical likelihood

- likelihood: basis for scanning the genome

  product over $i = 1,\ldots,n$ individuals

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

  $= \text{product}_i \text{ sum}_Q \text{ pr}(Q|X_i,\lambda) \text{ pr}(Y_i|Q,\theta)$
- empirical likelihood (Owen 1988)

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

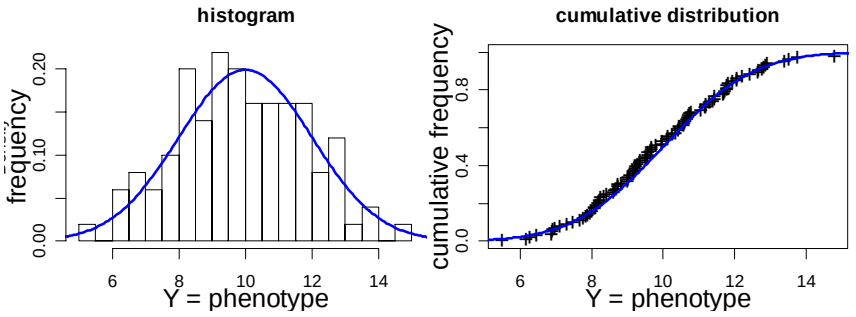
  $= \text{product}_i f(Y_i) w_i$

  with weights $w_i = \text{sum}_Q \text{ pr}(Q|X_i,\lambda) \exp(Y_i\beta_Q)$
  - relies only on flanking markers X_i
  - 4 possible values for BC, 9 for F2, etc.
- profile of likelihood: $L(\lambda|Y,X) = \max_\theta L(\theta,\lambda|Y,X)$ (rescaled LOD score)

semi-parametric formal tests

- clever tricks
  - partial and conditional empirical LOD
    - Zou, Fine, Yandell (2002 *Biometrika*); Zou Fine (2003 *Bioka*)
    - Jin, Fine, Yandell (in prep)
  - generalized estimating equations
    - Lange, Whittaker (2001 *Genetics*)
- tests similar to normal LOD
  - single locus test: approximate χ^2 with 1 d.f.
  - genome-wide scan: same critical values
  - permutation test: possible with some extra work

histograms and CDFs

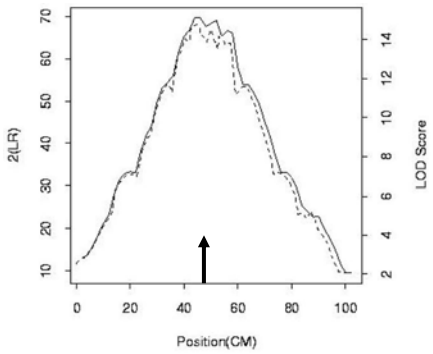


histograms capture shape
but are not very accurate

CDFs are more accurate
but not always intuitive

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

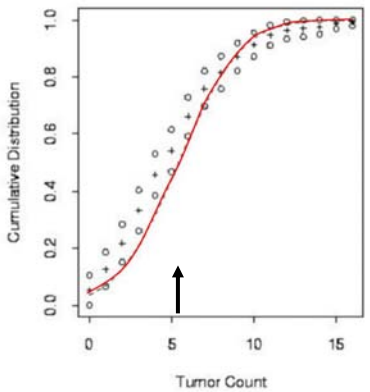
- rat backcross
  - two inbred strains
    - Wistar-Furth susceptible
    - Wistar-Kyoto resistant
  - backcross to WF
  - 383 females
  - chromosome 5, 58 markers
- search for resistance genes
- Y = # mammary carcinomas
- where is the QTL?



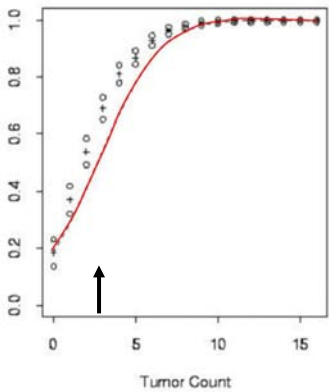
dash = normal
solid = semi-parametric

what shape histograms by genotype?
Zou et al. *Biometrika* (2002)

WF/WF



WKy/WF



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

non-parametric interval mapping

- phenotype model pr($Y|Q,\theta$) = $F_Q(Y)$
  - $\theta = F = (F_{qq},F_{Qq},F_{QQ})$ arbitrary distribution functions
- interval mapping Wilcoxon rank-sum test
  - replaced Y by rank(Y)
  - Kruglyak Lander (1995); Poole Drinkwater (1996); Broman (2003)
- estimator of horizontal shift in distribution
  - Hodges-Lehmann estimator: pr($Y|Q,\theta$) = $F_Q(Y) = F(Y+Q\beta)$
  - Zou, Yandell Fine (2003 *Genetics*, in press)
- non-parametric cumulative distribution
  - Fine, Zou, Yandell (2001 ms)
- stochastic ordering
  - Hoff et al. (2002)

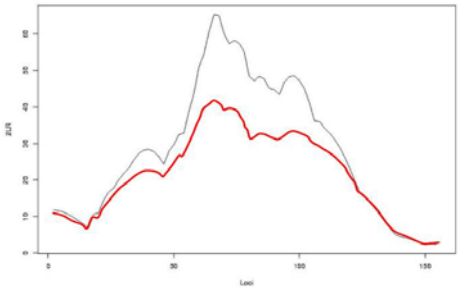
non-parametric QTL CDFs

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

  pr($Y \le y\ |X,\lambda$) = sum$_Q$ 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.
- readily extended to censored data
  - time to flowering for non-vernalized plants

what QTL influence flowering time? no vernalization: censored survival

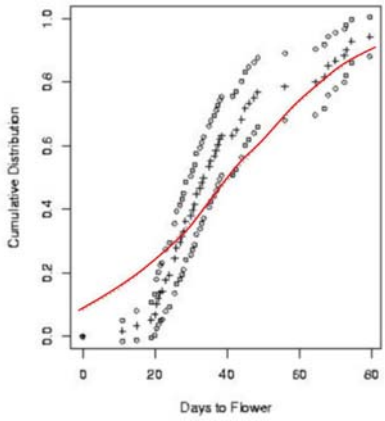
- *Brassica napus*
  - Major female
    - needs vernalization
  - Stellar male
    - insensitive
  - 99 double haploids
- *Y* = log(days to flower)
  - over 50% Major at QTL never flowered
  - log not fully effective



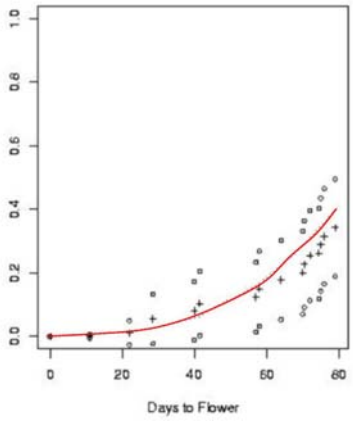
grey = normal, red = non-parametric

what shape is flowering distribution?

B. napus Stellar



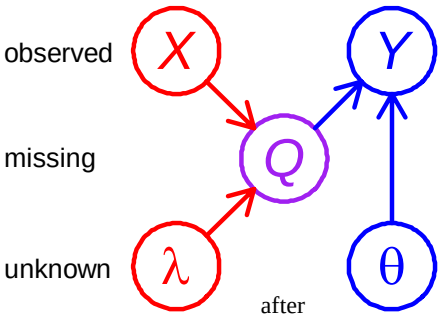
B. napus Major



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

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
  - m = number of QTL
- pr($Q|X,\lambda,m$) recombination model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for Q given X
- pr($Y|Q,\theta,m$) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters θ (could be non-parametric)



after
Sen Churchill (2001)

Bayesian interval mapping

- likelihood mixes over genotypes Q

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

  – maximize likelihood to estimate loci & effects
- Bayesian posterior samples Q as missing data

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

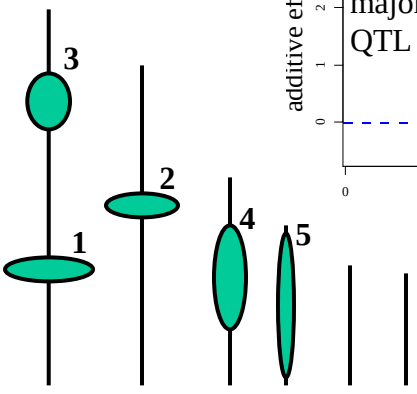
  – marginal summaries to estimate loci & effects
    - loci: $\text{pr}(\lambda|Y,X) = \text{sum}_{Q,\theta}\ \text{pr}(\lambda,Q,\theta|Y,X)$
    - effects: $\text{pr}(\theta|Y,X) = \text{sum}_{Q,\lambda}\ \text{pr}(\lambda,Q,\theta|Y,X)$
  – sample unknowns from posterior
    - prior beliefs built into $\text{pr}(\lambda,\theta)$

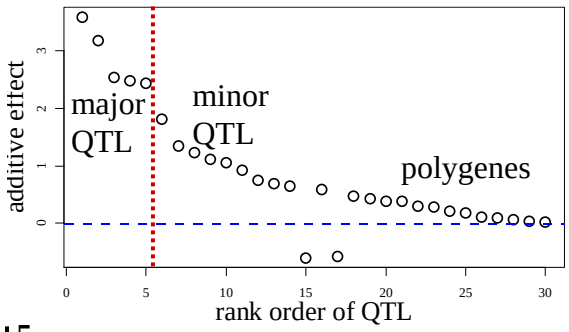
why worry about multiple QTL?

- 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
- build *M* = model = genetic architecture into model
  - *M* = {loci 1,2,…,*m*, plus interactions 12,13,…}
  - directly allow uncertainty in genetic architecture
  - model selection over number of QTL, genetic architecture
  - use Bayes factors and model averaging
    - to identify "better" models

Pareto diagram of QTL effects

major QTL on linkage map





multiple QTL phenotype model

- phenotype affected by genotype & environment

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

  $Y = G_Q$ + environment
- partition genotypic mean into QTL effects

  $G_Q = \mu + \beta_1(Q) + \ldots + \beta_m(Q) + \beta_{12}(Q) + \ldots$

  G_Q = mean + main effects + epistatic interactions
- general form of QTL effects for model M

  $G_Q = \mu + \text{sum}_{j \text{ in } M}\, \beta_j(Q)$

  $|M|$ = number of terms in model $M < 2^m$

prior for phenotype model

- want prior for $E(Y)$ that does not depend on model M
  - watch out for bias toward larger models
  - typically no prior information on genetic architecture
- priors on mean and effects for model M

$G_Q \sim N\left(\bar{Y}, (\kappa + h^2)s^2\right)$ model-independent expectation

$\mu \sim N(\bar{Y}, \kappa s^2)$ model-independent grand mean

$\beta_j(Q) \sim N\left(\bar{Y}, \frac{h^2 s^2}{|M|}\right)$ effects down-weighted by size of M

posterior mean & LS estimate

- posterior depends
  - prior specification
  - classical (least squares) estimates

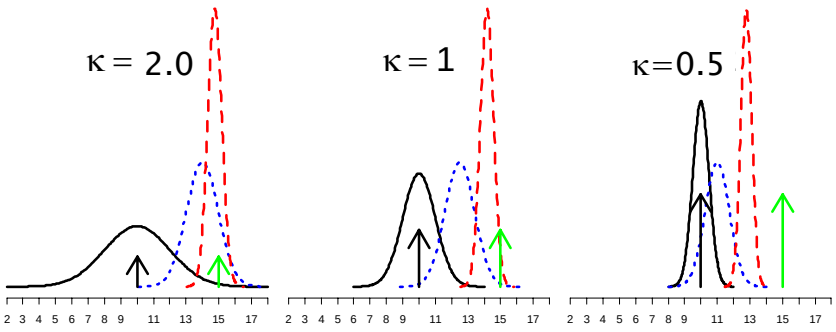
$$G_Q \mid Y, m \sim N(\bar{Y} + B_Q(\hat{G}_Q - \bar{Y}), B_Q C_Q \sigma^2)$$
$$\approx N(\hat{G}_Q, C_Q \sigma^2)$$

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

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

shrinkage $B_Q = \left(1 + \frac{C_Q^{\ 2}}{\kappa + h^2}\right)^{-1} \to 1$

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)

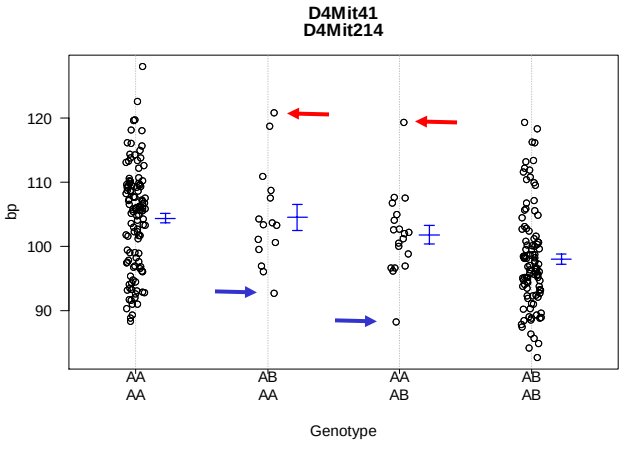
prior & posterior for genotypes *Q*

- prior is recombination model

  $\text{pr}(Q|X_i,\lambda)$
- 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

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

how does phenotype *Y* affect *Q*?



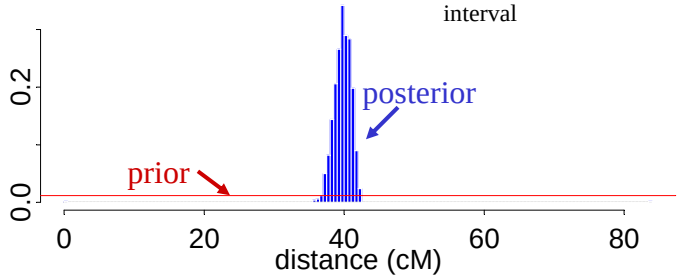
what are probabilities for genotype *Q* between markers?

recombinants AA:AB

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

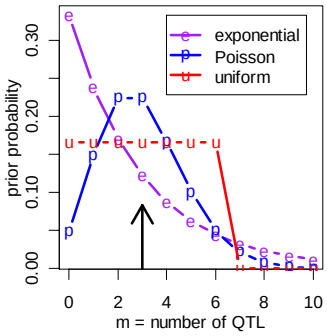
prior & posterior for QT locus

- 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”



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)

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

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

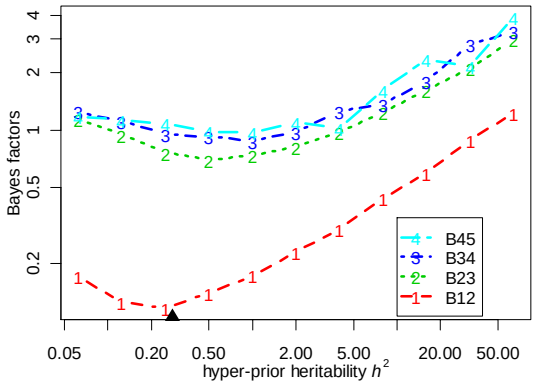
QTL Bayes factors & RJ-MCMC

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

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

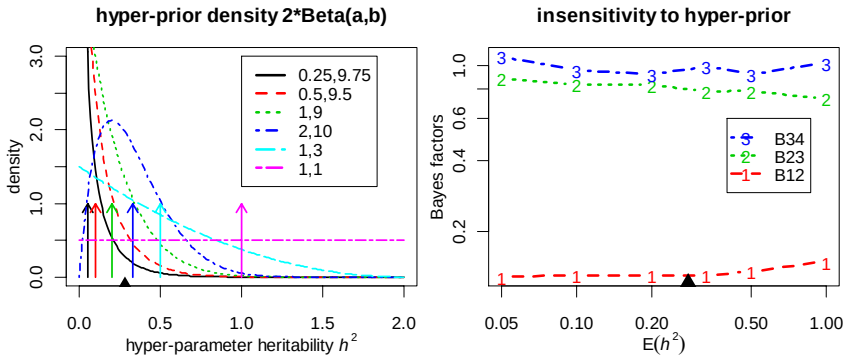
- *BF* insensitive to shape of prior
  - geometric, Poisson, uniform
  - precision improves when prior mimics posterior
- *BF* sensitivite to prior variance on effects θ
  - prior variance should reflect data variability
  - apparently resolved by using hyper-priors
    - automatic algorithm; no need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

BF sensitivity to fixed prior for effects



$$\beta_j(Q) \sim \mathrm{N}\left(0, \frac{h^2 s^2}{|M|}\right), h^2 \text{ fixed}$$

BF insensitivity to random effects prior



$$\beta_j(Q) \sim \mathrm{N}\left(0, \frac{h^2 s^2}{|M|}\right), \frac{h^2}{2} \sim \mathrm{Beta}(a,b)$$

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 locus λ given Q,X (using Metropolis-Hastings step)
  - update genotypes Q given λ,θ,Y,X (using Gibbs sampler)
  - update effects θ given Q,Y (using Gibbs sampler)

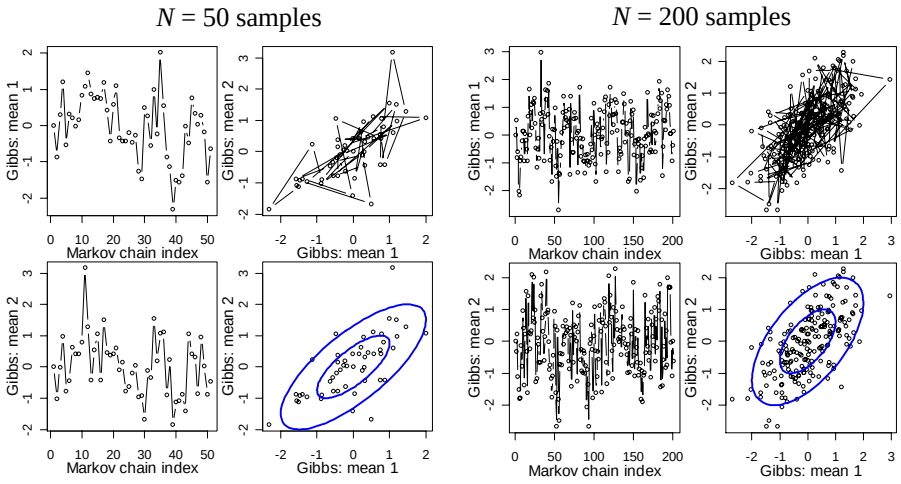
$$(\lambda,Q,\theta,m) \sim \text{pr}(\lambda,Q,\theta,m \mid Y,X)$$
$$(\lambda,Q,\theta,m)_1 \rightarrow (\lambda,Q,\theta,m)_2 \rightarrow \cdots \rightarrow (\lambda,Q,\theta,m)_N$$

Gibbs sampler idea

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

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}\Big|\,\mu,\rho \sim N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$
$$\theta_1 \mid \theta_2,\mu,\rho \sim N\left(\mu_1+\rho(\theta_2-\mu_2),1-\rho^2\right)$$
$$\theta_2 \mid \theta_1,\mu,\rho \sim N\left(\mu_2+\rho(\theta_1-\mu_1),1-\rho^2\right)$$

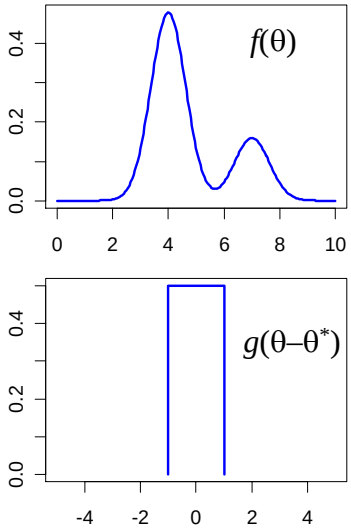
Gibbs sampler samples: $\rho = 0.6$



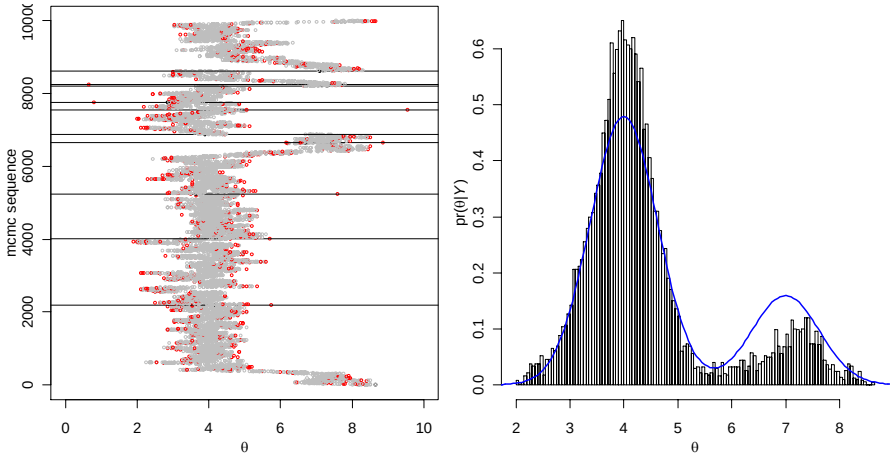
Metropolis-Hastings idea

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

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

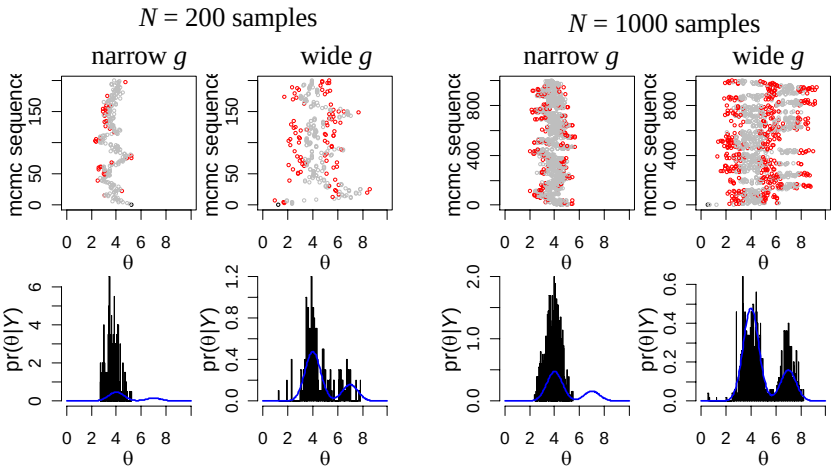


MCMC realization

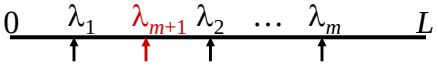


added twist: occasionally propose from whole domain

Metropolis-Hastings samples



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

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

Bayesian software for QTLs

- R/bim (Satagopan Yandell 1996; Gaffney 2001)
  - www.stat.wisc.edu/~yandell/qtl/software/Bmapqtl
  - www.r-project.org contributed package
  - version available within WinQTLCart (statgen.ncsu.edu/qtlcart)
- Bayesian IM with epistasis (Nengjun Yi, U AB)
  - separate C++ software (papers with Xu)
  - plans in progress to incorporate into R/bim
- R/qtl (Broman et al. 2003)
  - biosun01.biostat.jhsph.edu/~kbroman/software
  - www.r-project.org contributed package
- Pseudomarker (Sen Churchill 2002)
  - www.jax.org/staff/churchill/labsite/software
- Bayesian QTL / Multimapper
  - Sillanpää Arjas (1998)
  - www.rni.helsinki.fi/~mjs
- Stephens & Fisch (email)

Bmapqtl: our RJ-MCMC software

- www.stat.wisc.edu/~yandell/qtl/software/Bmapqtl
  - R contributed library (www.r-project.org)
    - library(bim) is cross-compatible with library(qtl)
  - module within WinQTLCart format
- Bayes factor & reversible jump MCMC
  - 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
  - upgrade (H Wu, PJ Gaffney, CF Jin, BS Yandell 2003)

shape phenotype in BC study indexed by PC1

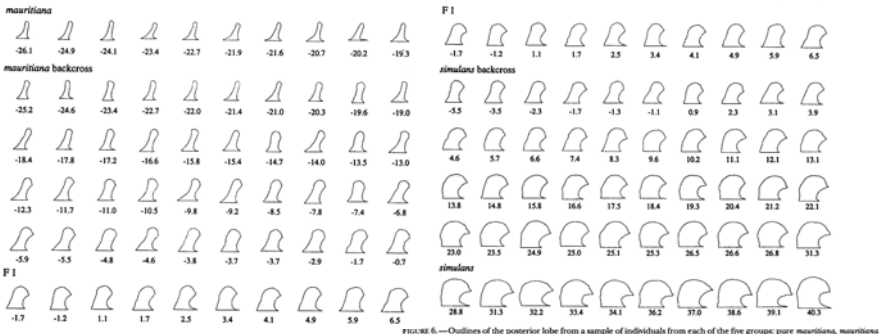


FIGURE 6.—Outlines of the posterior lobe from a sample of individuals from each of the five groups: pure *mauritiana*, *mauritiana* backcross, F_1, *simulans* backcross, and pure *simulans*. Within each group, the outlines are presented in order of their PC1 score (sampled at even intervals from the range of variation). The number below each specimen is its PC1 score. The outlines are drawn to scale with the origin at the centroid of each outline and with all baselines parallel.

Liu et al. (1996) *Genetics*

shape phenotype via PC

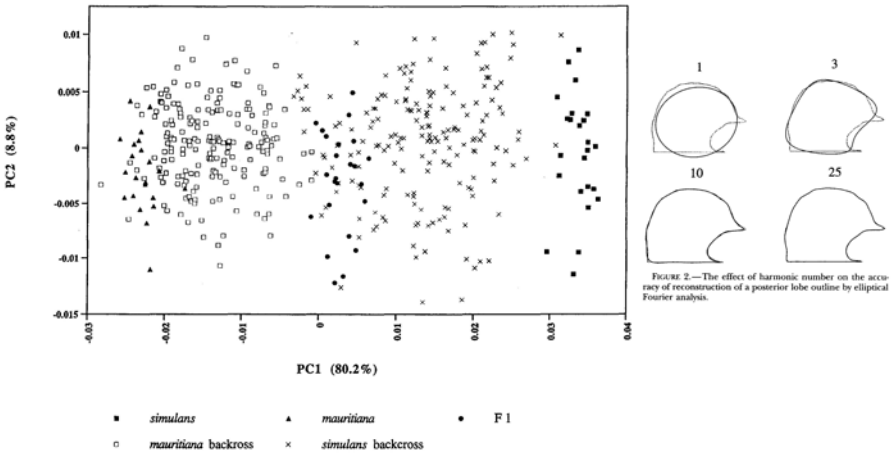


FIGURE 5.—A plot of the first two principal components of the Fourier coefficients from posterior lobe outlines. Many individuals from each of five genotypic classes are represented. Each point represents an average of scores from the left and right sides of an individual (with a few exceptions for which the score is from one side only). The percentage of variation in the Fourier coefficients accounted for by each principal component is given in parentheses.

FIGURE 2.—The effect of harmonic number on the accuracy of reconstruction of a posterior lobe outline by elliptical Fourier analysis.

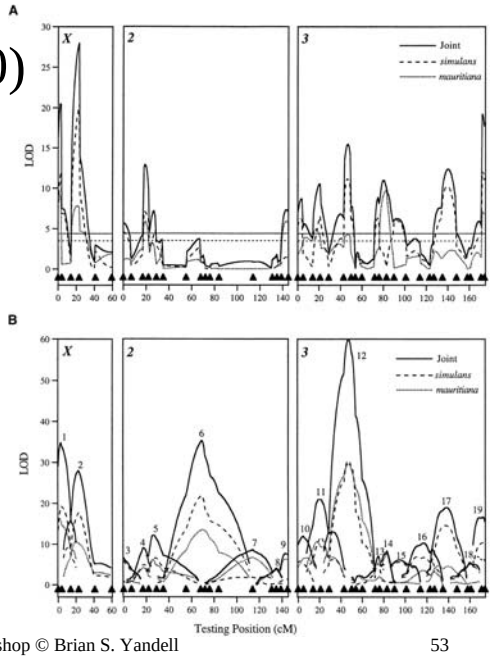
Liu et al. (1996) *Genetics*

Zeng et al. (2000) CIM vs. MIM

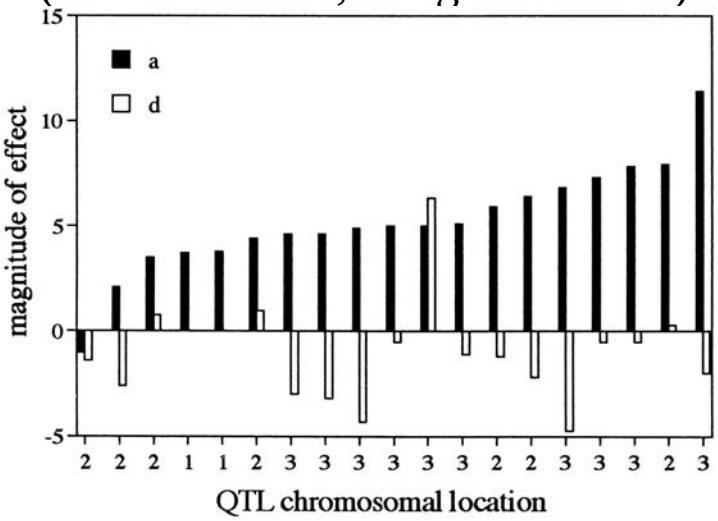
composite interval mapping
(Liu et al. 1996)
narrow peaks
miss some QTL

multiple interval mapping
(Zeng et al. 2000)
triangular peaks

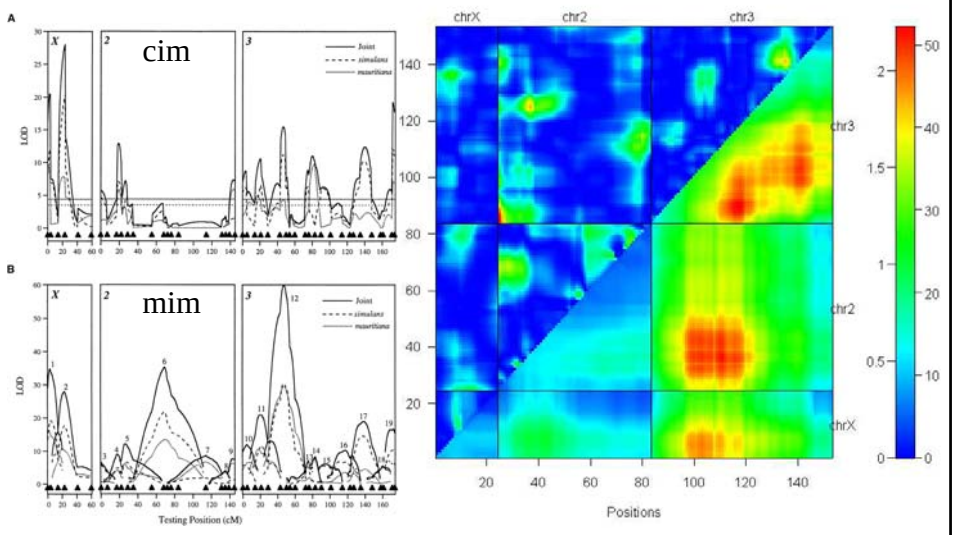
both conditional 1-D scans
fixing all other "QTL"



MIM effects for gonad shape
(Liu et al. 1996; Zeng et al. 2000)



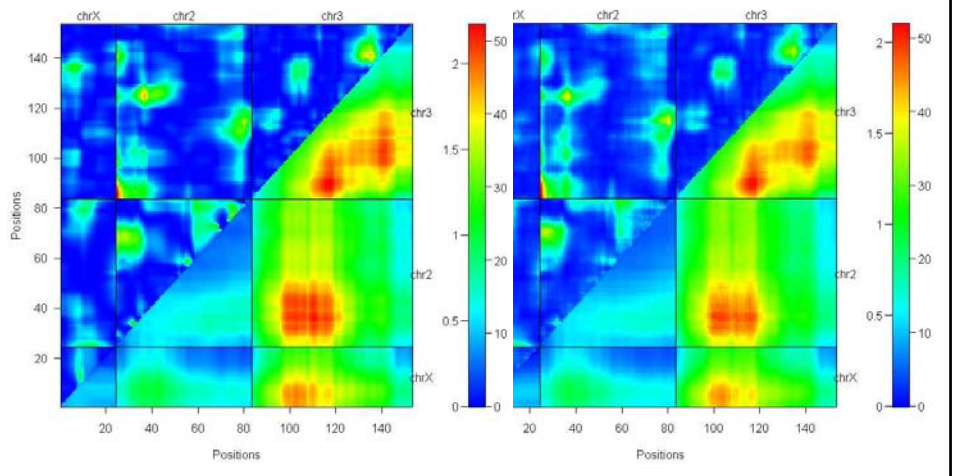
CIM, MIM and IM pairscan



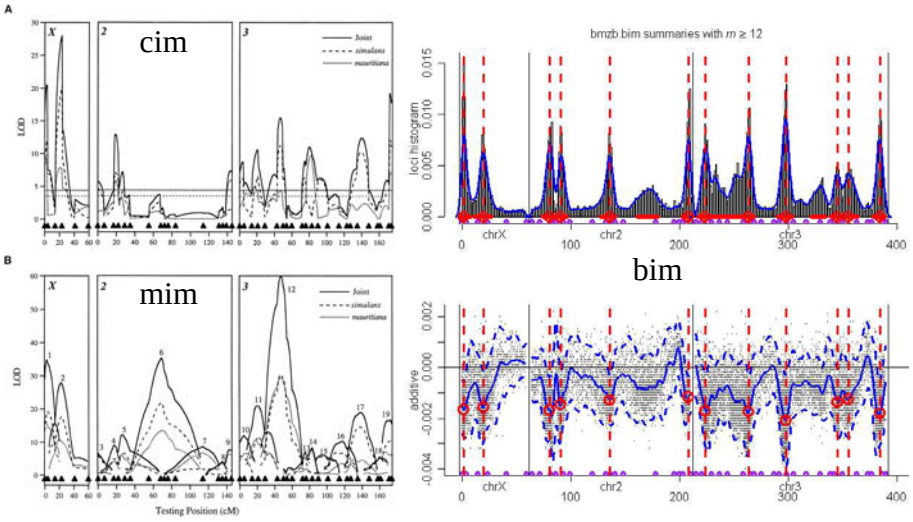
2 QTL + epistasis: IM versus multiple imputation

IM pairscan

multiple imputation



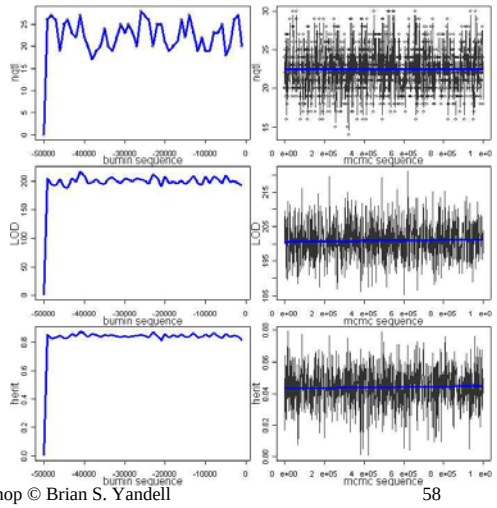
multiple QTL: CIM, MIM and BIM



MCMC diagnostics for *Dm* shape

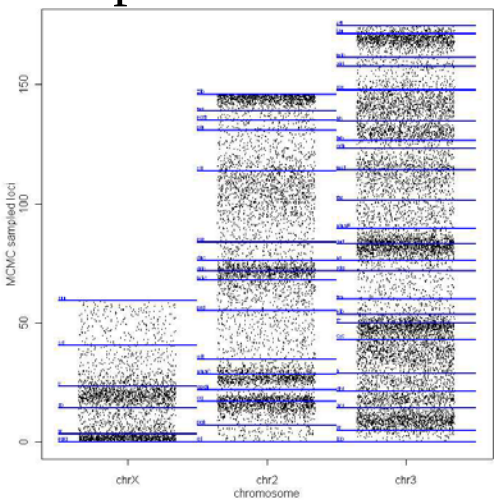
- *m* ~ Poisson(15) prior on number of QTL
- Bayesian LOD (log posterior density)
- Heritability

- 5% burnin
- 1,000,000 samples
  - every 1000th recorded
- note stable mean



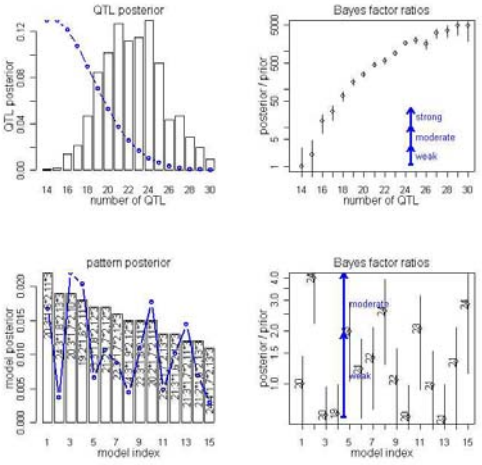
MCMC sampled loci

- markers as blue lines
  - horizontal jittering
- note denser regions
  - 10-11 broad regions
- jointly sampling
  - 15-30 QTL at once



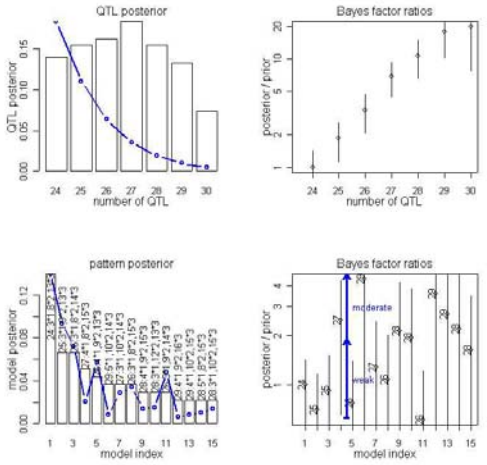
MCMC model selection

- m = number of QTL
  - prior: Poisson(15)
    - rescaled in blue
  - posterior: mean 22.4
  - Bayes factor increases
- pattern across genome
  - prior depends on m and length of chromosomes
  - posterior mode: m=20
  - Bayes factor favors
    - m = 24
    - 3*1, 8*2, 13*3



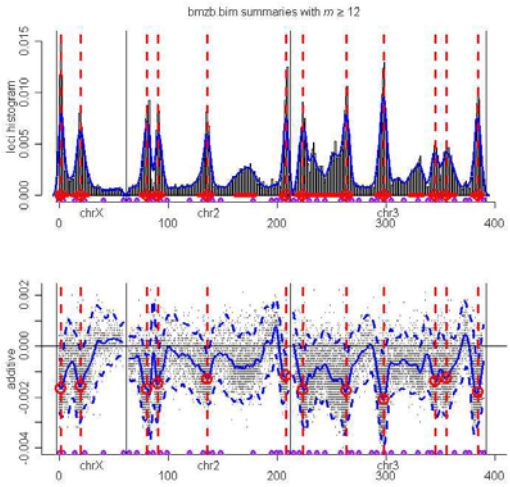
MCMC model selection restricted to "better models"

- models with minimum
  - $m \geq 24$
  - pattern ≥ 3*1, 8*2, 13*3
- note uncertainty in BF
  - estimate ± 2 SE
- mode is chosen pattern
  - ~14% of samples
- BF similar to more complicated patterns
  - parsimony: simpler model
  - 2SE intervals overlap



MCMC loci and effects

- model averaging
  - over all models
  - 1000 samples
- histogram of loci
  - marginal posteriors
  - superimposed on genome
  - 12 peaks identified
- scatterplot: loci & effects
  - smoothed mean ± 2 SE



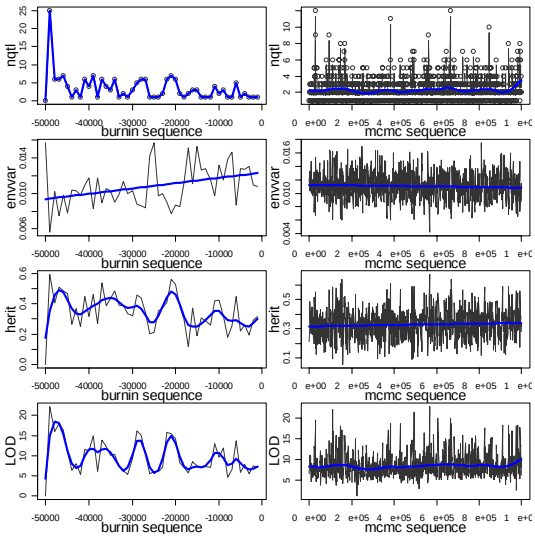
B. napus 8-week vernalization whole genome study

- 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
- available in R/bim package

Markov chain Monte Carlo sequence

burnin (sets up chain)
mcmc sequence

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

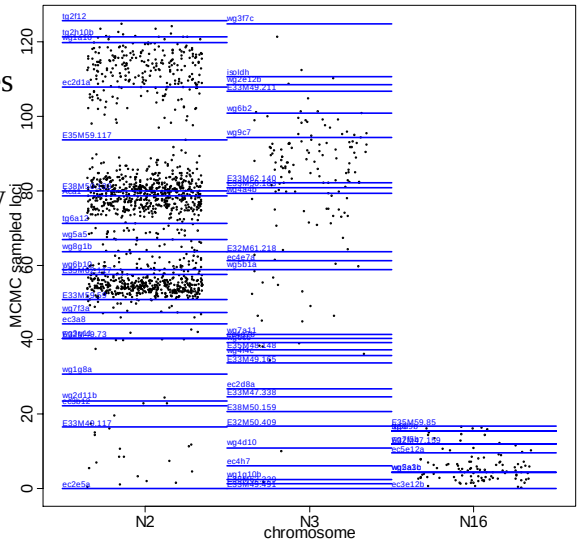


MCMC sampled loci

subset of chromosomes
N2, N3, N16

points jittered for view
blue lines at markers

note concentration
on chromosome N2

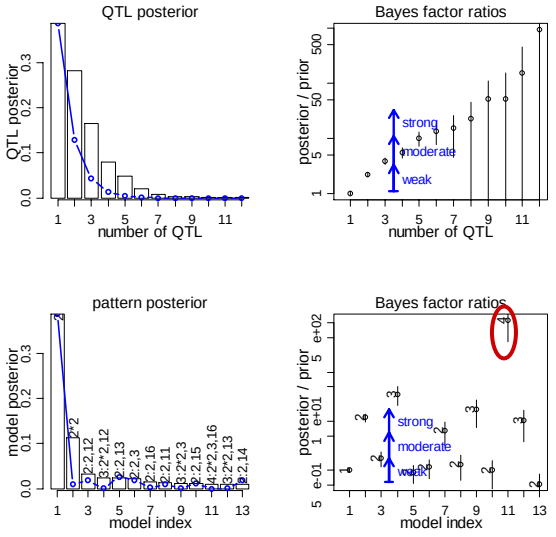


Bayesian model assessment

row 1: # QTL
row 2: pattern

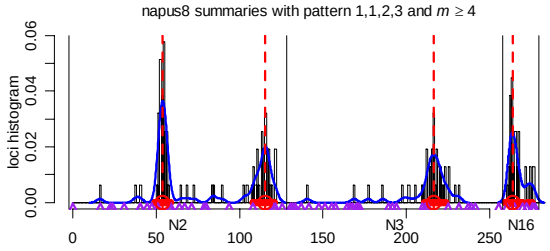
col 1: posterior
col 2: Bayes factor
note error bars on bf

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

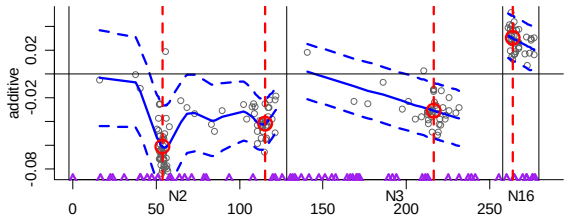


Bayesian estimates of loci & effects

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

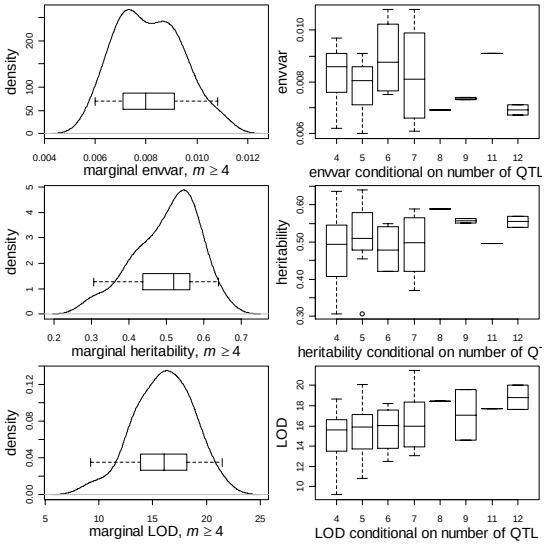


Bayesian model diagnostics

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*



many thanks

Michael Newton
Daniel Sorensen
Daniel Gianola
Yang Song
Fei Zou
Liang Li
Hong Lan
Hao Wu

Tom Osborn
David Butruille
Marcio Ferrera
Josh Udahl
Pablo Quijada
Alan Attie
Jonathan Stoehr
Gary Churchill

USDA Hatch, NIH/NIDDK, Jackson Labs