

NCSU Summer Institute 2005
QTL II
Brian S. Yandell
University of Wisconsin-Madison

- Bayesian interval mapping with prior info
- model selection for multiple QTL
- data examples in detail
- multiple phenotypes & microarrays

contact information & resources

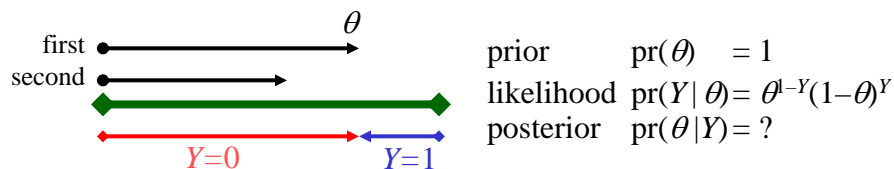
- email: byandell@wisc.edu
- web: www.stat.wisc.edu/~yandell/statgen
 - QTL & microarray resources
 - references, software, people
- thanks:
 - students: Jaya Satagopan, Pat Gaffney, Fei Zou, Amy Jin
 - faculty/staff: Alan Attie, Michael Newton, Nengjun Yi, Gary Churchill, Hong Lan, Christina Kendziorski, Tom Osborn, Jason Fine

Bayesian Interval Mapping

1. what is Bayes? Bayes theorem? 2-6
2. Bayesian QTL mapping 7-17
3. Markov chain sampling 18-25
4. sampling across architectures 26-32

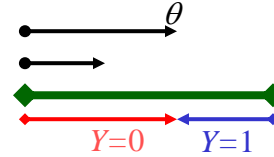
1. who was Bayes? what is Bayes theorem?

- 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?)
- basic idea (from Bayes' original example)
 - two billiard balls tossed at random (uniform) on table
 - where is first ball if the second is to its left (right)?



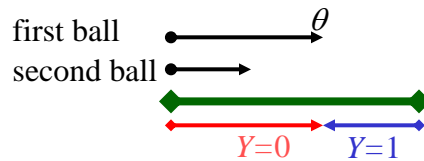
what is Bayes theorem?

- where is first ball if the second is to its **left** (right)?
- prior: probability of parameter before observing data
 - $\text{pr}(\theta) = \text{pr}(\text{parameter})$
 - equal chance of being anywhere on the table
- posterior: probability of parameter after observing data
 - $\text{pr}(\theta | Y) = \text{pr}(\text{parameter} | \text{data})$
 - more likely to left if first ball is toward the right end of table
- likelihood: probability of data given parameters
 - $\text{pr}(Y | \theta) = \text{pr}(\text{data} | \text{parameter})$
 - basis for classical statistical inference
- Bayes theorem
 - posterior = likelihood * prior / pr(data)
 - normalizing constant $\text{pr}(Y)$ often drops out of calculation

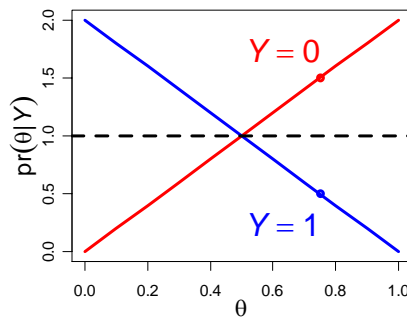


$$\text{pr}(\theta | Y) = \frac{\text{pr}(\theta, Y)}{\text{pr}(Y)} = \frac{\text{pr}(Y | \theta) \times \text{pr}(\theta)}{\text{pr}(Y)}$$

where is the second ball given the first?

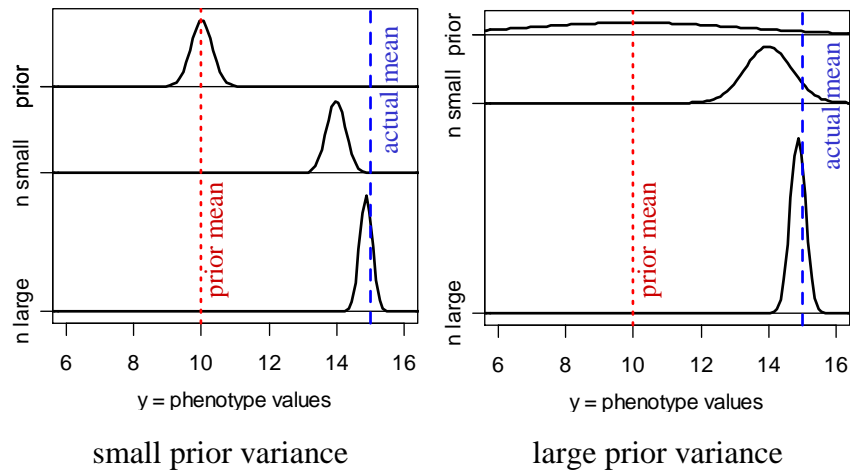


- prior $\text{pr}(\theta) = 1$
- likelihood $\text{pr}(Y | \theta) = \theta^l (1 - \theta)^Y$
- posterior $\text{pr}(\theta | Y) = ?$



- prior : $\text{pr}(\theta) = 1$
- likelihood : $\text{pr}(Y | \theta) = \begin{cases} \theta & \text{if } Y = 0 \\ 1 - \theta & \text{if } Y = 1 \end{cases}$
- marginal : $\text{pr}(Y) = \frac{1}{2}$
- posterior : $\text{pr}(\theta | Y) = \frac{\text{pr}(Y | \theta) \text{pr}(\theta)}{\text{pr}(Y)}$
 $= \begin{cases} 2\theta & \text{if } Y = 0 \\ 2(1 - \theta) & \text{if } Y = 1 \end{cases}$

Bayes posterior for normal data



Bayes

NCSU QTL II: Yandell © 2005

5

Bayes posterior for normal data

model	$Y_i = \mu + E_i$
environment	$E \sim N(0, \sigma^2), \sigma^2 \text{ known}$
likelihood	$Y \sim N(\mu, \sigma^2)$
prior	$\mu \sim N(\mu_0, \kappa\sigma^2), \kappa \text{ known}$
posterior:	mean tends to sample mean
single individual	$\mu \sim N(\mu_0 + B_1(Y_1 - \mu_0), B_1\sigma^2)$
sample of n individuals	$\mu \sim N(B_n \bar{Y}_\bullet + (1 - B_n)\mu_0, B_n\sigma^2 / n)$
	with $\bar{Y}_\bullet = \sum_{i=1, \dots, n} Y_i / n$
fudge factor (shrinks to 1)	$B_n = \frac{\kappa n}{\kappa n + 1} \rightarrow 1$

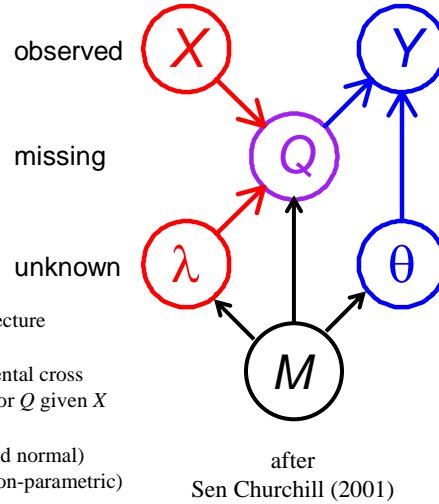
Bayes

NCSU QTL II: Yandell © 2005

6

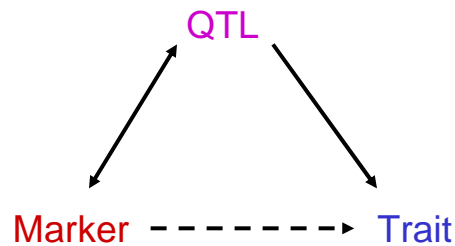
2. Bayesian QTL mapping

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



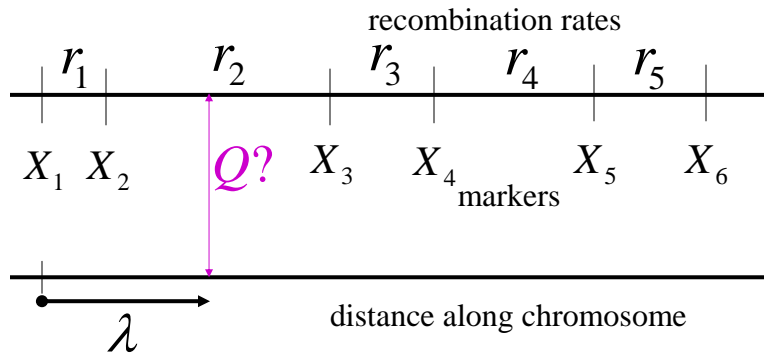
QTL Mapping (Gary Churchill)

Key Idea: Crossing two inbred lines creates linkage disequilibrium which in turn creates associations and linked segregating QTL



pr(Q/X, λ) recombination model

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



pr(Y|Q, θ) phenotype model

- pr(trait | genotype, effects)
- trait = genetic + environment
- assume genetic uncorrelated with environment

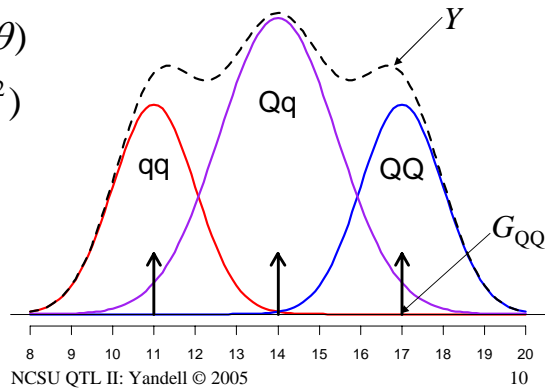
$$\text{pr}(Y \mid Q = q, \theta)$$

$$Y \sim N(G_q, \sigma^2)$$

$$Y = G_q + e$$

$$\text{var}(G_q) = \sigma_G^2$$

$$\text{var}(e) = \sigma^2$$



Bayesian priors for QTL

- missing genotypes Q
 - $\text{pr}(Q | X, \lambda)$
 - recombination model is formally a prior
- effects $\theta = (G, \sigma^2)$
 - $\text{pr}(\theta) = \text{pr}(G_q | \sigma^2) \text{pr}(\sigma^2)$
 - use conjugate priors for normal phenotype
 - $\text{pr}(G_q | \sigma^2) = \text{normal}$
 - $\text{pr}(\sigma^2) = \text{inverse chi-square}$
- each locus λ may be uniform over genome
 - $\text{pr}(\lambda | X) = 1 / \text{length of genome}$
- combined prior
 - $\text{pr}(Q, \theta, \lambda | X) = \text{pr}(Q | X, \lambda) \text{pr}(\theta) \text{pr}(\lambda | X)$

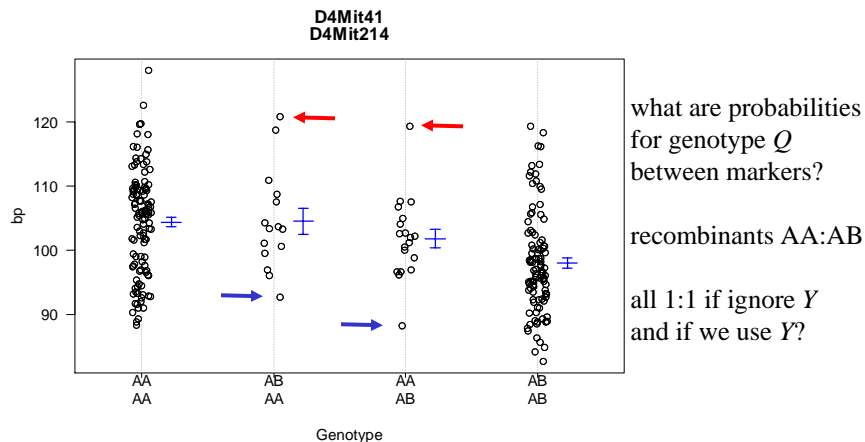
Bayesian model posterior

- augment data (Y, X) with unknowns Q
- study unknowns (θ, λ, Q) given data (Y, X)
 - properties of posterior $\text{pr}(\theta, \lambda, Q | Y, X)$
- sample from posterior in some clever way
 - multiple imputation or MCMC

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

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

how does phenotype Y improve posterior for genotype Q ?



Bayes

NCSU QTL II: Yandell © 2005

13

posterior on QTL genotypes

- full conditional for Q depends data for individual i
 - proportional to prior $\text{pr}(Q | X_i, \lambda)$
 - weight toward Q that agrees with flanking markers
 - proportional to likelihood $\text{pr}(Y_i | Q, \theta)$
 - weight toward $Q=q$ so that group mean $G_q \approx Y_i$
- phenotype and prior recombination may conflict
 - posterior recombination balances these two weights

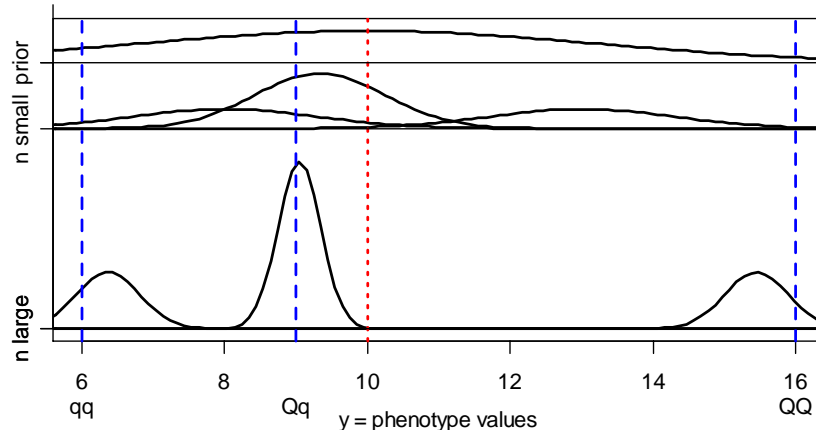
$$\text{pr}(Q | Y_i, X_i, \theta, \lambda) = \frac{\text{pr}(Q | X_i, \lambda) \text{pr}(Y_i | Q, \theta)}{\text{pr}(Y_i | X_i, \theta, \lambda)}$$

Bayes

NCSU QTL II: Yandell © 2005

14

posterior genotypic means G_q



posterior genotypic means G_q

posterior centered on sample genotypic mean
but shrunken slightly toward overall mean

prior: $G_q \sim N(\bar{Y}_\bullet, \kappa\sigma^2)$

posterior: $G_q \sim N(B_q \bar{Y}_q + (1 - B_q) \bar{Y}_\bullet, B_q \sigma^2 / n_q)$

$$n_q = \text{count}\{Q_i = q\}, \bar{Y}_q = \frac{\sum_{\{Q_i=q\}} Y_i}{n_q}$$

fudge factor: $B_q = \frac{\kappa n_q}{\kappa n_q + 1} \rightarrow 1$

What if variance σ^2 is unknown?

- sample variance is proportional to chi-square
 - $ns^2/\sigma^2 \sim \chi^2(n)$
 - likelihood of sample variance s^2 given n, σ^2
- conjugate prior is inverse chi-square
 - $v\tau^2/\sigma^2 \sim \chi^2(v)$
 - prior of population variance σ^2 given v, τ^2
- posterior is weighted average of likelihood and prior
 - $(v\tau^2 + ns^2)/\sigma^2 \sim \chi^2(v+n)$
 - posterior of population variance σ^2 given n, s^2, v, τ^2
- empirical choice of hyper-parameters
 - $\tau^2 = s^2/3, v=6$
 - $E(\sigma^2/v, \tau^2) = s^2/2, \text{Var}(\sigma^2/v, \tau^2) = s^4/4$

3. Markov chain sampling of architectures

- 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
- hard to sample (Q, θ, λ, M) from joint posterior
 - update (Q, θ, λ) from full conditionals for model M
 - update genetic model M

$$(Q, \theta, \lambda, M) \sim \text{pr}(Q, \theta, \lambda, M | Y, X)$$

$$(Q, \theta, \lambda, M)_1 \rightarrow (Q, \theta, \lambda, M)_2 \rightarrow \dots \rightarrow (Q, \theta, \lambda, M)_N$$

Gibbs sampler idea

- toy problem
 - want to study two correlated effects
 - could sample directly from their bivariate distribution
- instead use Gibbs sampler:
 - sample each effect from its full conditional given the other
 - pick order of sampling at random
 - repeat many times

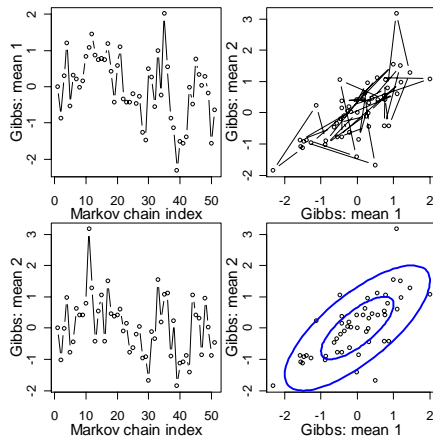
$$\begin{pmatrix} G_1 \\ G_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

$$G_1 \sim N(\mu_1 + \rho(G_2 - \mu_2), 1 - \rho^2)$$

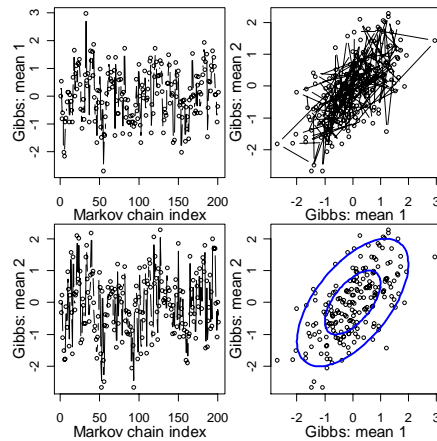
$$G_2 \sim N(\mu_2 + \rho(G_1 - \mu_1), 1 - \rho^2)$$

Gibbs sampler samples: $\rho = 0.6$

$N = 50$ samples



$N = 200$ samples



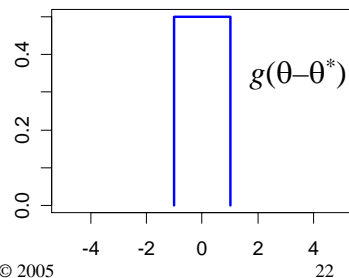
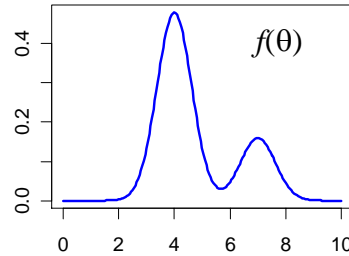
MCMC sampling of (λ, Q, θ)

- Gibbs sampler
 - genotypes Q
 - effects $\theta = (G, \sigma^2)$
 - *not* loci λ
- $$Q \sim \frac{\text{pr}(Q | Y_i, X_i, \theta, \lambda)}{\text{pr}(Y | Q)}$$
- $$\theta \sim \frac{\text{pr}(Y | Q, \theta) \text{pr}(\theta)}{\text{pr}(Y | Q)}$$
- $$\lambda \sim \frac{\text{pr}(Q | X, \lambda) \text{pr}(\lambda | X)}{\text{pr}(Q | X)}$$
- Metropolis-Hastings sampler
 - extension of Gibbs sampler
 - does not require normalization
 - $\text{pr}(Q | X) = \sum_{\lambda} \text{pr}(Q | X, \lambda) \text{pr}(\lambda)$

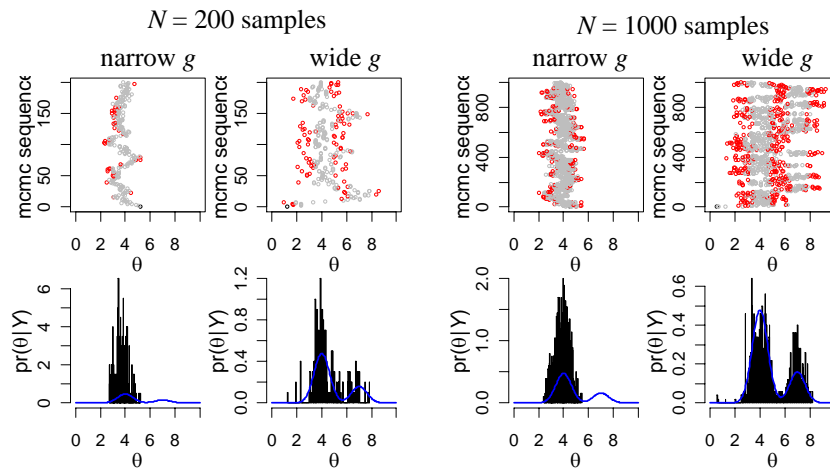
Metropolis-Hastings idea

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

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



Metropolis-Hastings samples



Bayes

NCSU QTL II: Yandell © 2005

23

full conditional for locus

- cannot easily sample from locus full conditional

$$pr(\lambda | Y, X, \theta, Q) = pr(\lambda | X, Q)$$

$$= pr(Q / X, \lambda) pr(\lambda) / \text{constant}$$
- constant is very difficult to compute explicitly
 - must average over all possible loci λ over genome
 - must do this for every possible genotype Q
- Gibbs sampler will not work in general
 - but can use method based on ratios of probabilities
 - Metropolis-Hastings is extension of Gibbs sampler

Bayes

NCSU QTL II: Yandell © 2005

24

Metropolis-Hastings Step

- pick new locus based upon current locus
 - propose new locus from some distribution $g(\cdot)$
 - pick value near current one? (usually)
 - pick uniformly across genome? (sometimes)
 - accept new locus with probability A
 - otherwise stick with current value

$$A(\lambda_{old}, \lambda_{new}) = \min\left(1, \frac{\text{pr}(\lambda_{new})\text{pr}(Q | X, \lambda_{new})g(\lambda_{new}, \lambda_{old})}{\text{pr}(\lambda_{old})\text{pr}(Q | X, \lambda_{old})g(\lambda_{old}, \lambda_{new})}\right)$$

4. sampling across architectures

- search across genetic architectures M of various sizes
 - allow change in $m =$ number of QTL
 - allow change in types of epistatic interactions
- methods for search
 - reversible jump MCMC
 - Gibbs sampler with loci indicators
- complexity of epistasis
 - Fisher-Cockerham effects model
 - general multi-QTL interaction & limits of inference


search across genetic architectures

- think model selection in multiple regression
 - but regressors (QTL genotypes) are unknown
- linked loci are collinear regressors
 - correlated effects
- consider only main genetic effects here
 - epistasis just gets more complicated

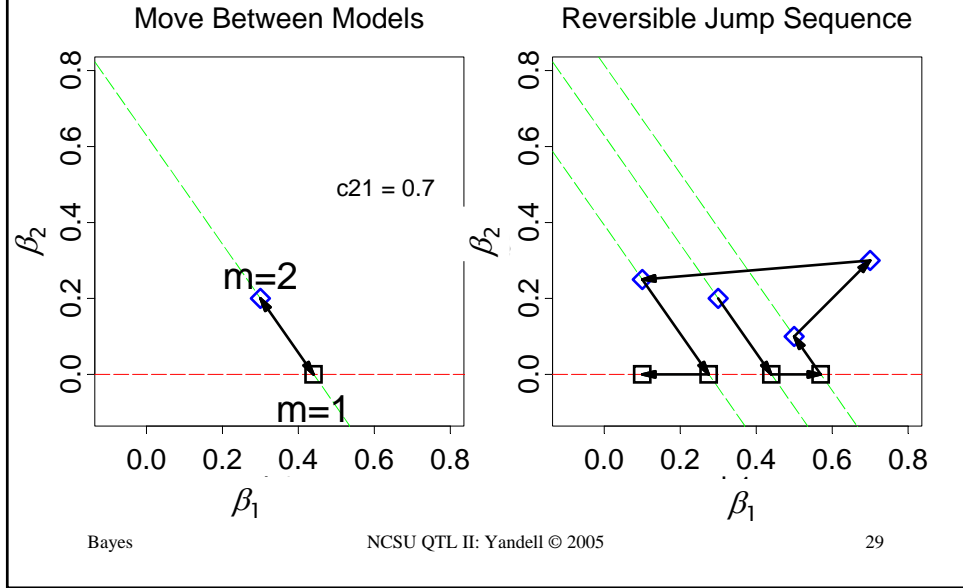
$$G_q = \mu + \beta_{q1} \text{ or } G_q = \mu + \beta_{q1} + \beta_{q2}$$

model selection in regression

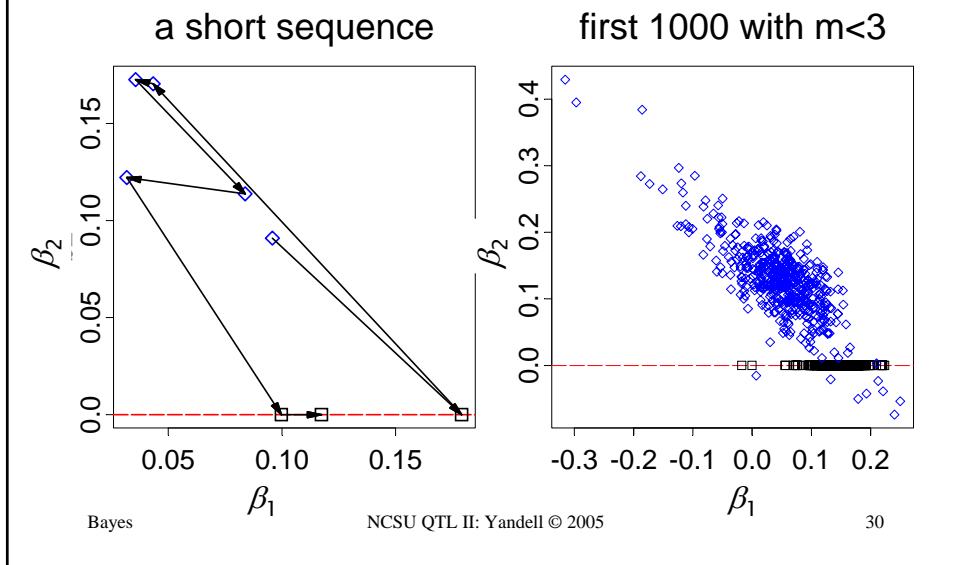
- consider known genotypes Q at 2 known loci λ
 - models with 1 or 2 QTL
- jump between 1-QTL and 2-QTL models
- adjust parameters when model changes
 - β_{q1} estimate changes between models 1 and 2
 - due to collinearity of QTL genotypes


$$m = 1 : G_q = \mu + \beta_{q1}$$
$$m = 2 : G_q = \mu + \beta_{q1} + \beta_{q2}$$

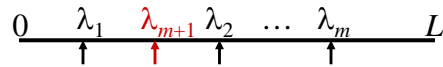
geometry of reversible jump



geometry allowing Q and λ to change



reversible jump MCMC idea



- Metropolis-Hastings updates: 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 and innovate new genotypes & genotypic 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
- Satagopan Yandell (1996, 1998); Sillanpaa Arjas (1998); Stevens Fisch (1998)
 - these build on RJ-MCMC idea of Green (1995); Richardson Green (1997)

Gibbs sampler with loci indicators

- consider only QTL at pseudomarkers
 - every 1-2 cM
 - modest approximation with little bias
- use loci indicators in each pseudomarker
 - $\delta = 1$ if QTL present
 - $\delta = 0$ if no QTL present
- Gibbs sampler on loci indicators δ
 - relatively easy to incorporate epistasis
 - Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005 *Genetics*)
 - (see earlier work of Nengjun Yi and Ina Hoeschele)

$$G_q = \mu + \delta_1 \beta_{q1} + \delta_2 \beta_{q2}$$

Model Selection for Multiple QTL

1. what is goal of QTL study? 2-7
2. gene action and epistasis 8-15
3. comparing QTL models 16-23
4. simulations and data studies 24-28

1. what is the goal of QTL study?

- uncover underlying biochemistry
 - identify how networks function, break down
 - find useful candidates for (medical) intervention
 - epistasis may play key role
 - statistical goal: maximize number of correctly identified QTL
- basic science/evolution
 - how is the genome organized?
 - identify units of natural selection
 - additive effects may be most important (Wright/Fisher debate)
 - statistical goal: maximize number of correctly identified QTL
- select “elite” individuals
 - predict phenotype (breeding value) using suite of characteristics (phenotypes) translated into a few QTL
 - statistical goal: minimize prediction error

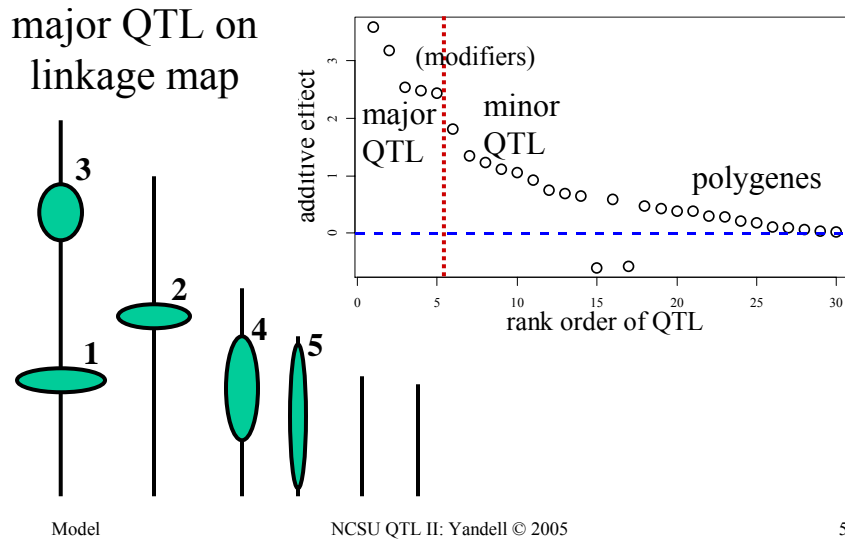
pragmatics of multiple QTL

- evaluate some objective for model given data
 - classical likelihood
 - Bayesian posterior
- search over possible genetic architectures (models)
 - number and positions of loci
 - gene action: additive, dominance, epistasis
- estimate “features” of model
 - means, variances & covariances, confidence regions
 - marginal or conditional distributions
- art of model selection
 - how select “best” or “better” model(s)?
 - how to search over useful subset of possible models?

advantages of multiple QTL approach

- improve statistical power, precision
 - increase number of QTL detected
 - better estimates of loci: less bias, smaller intervals
- improve inference of complex genetic architecture
 - patterns and individual elements of epistasis
 - appropriate estimates of means, variances, covariances
 - asymptotically unbiased, efficient
 - assess relative contributions of different QTL
- improve estimates of genotypic values
 - less bias (more accurate) and smaller variance (more precise)
 - mean squared error = $MSE = (\text{bias})^2 + \text{variance}$

Pareto diagram of QTL effects



limits of multiple QTL?

- limits of statistical inference
 - power depends on sample size, heritability, environmental variation
 - “best” model balances fit to data and complexity (model size)
 - genetic linkage = correlated estimates of gene effects
- limits of biological utility
 - sampling: only see some patterns with many QTL
 - marker assisted selection (Bernardo 2001 *Crop Sci*)
 - 10 QTL ok, 50 QTL are too many
 - phenotype better predictor than genotype when too many QTL
 - increasing sample size may not give multiple QTL any advantage
 - hard to select many QTL simultaneously
 - 3^m possible genotypes to choose from

QTL below detection level?

- problem of selection bias
 - QTL of modest effect only detected sometimes
 - their effects are biased upwards when detected
- probability that QTL detected
 - avoids sharp in/out dichotomy
 - avoid pitfalls of one “best” model
 - examine “better” models with more probable QTL
- build $m =$ number of QTL detected into QTL model
 - directly allow uncertainty in genetic architecture
 - model selection over genetic architecture

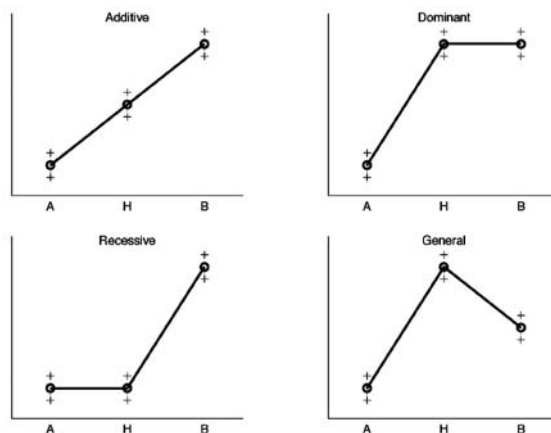
Model

NCSU QTL II: Yandell © 2005

7

2. Gene Action and Epistasis

additive, dominant, recessive, general effects
of a single QTL (Gary Churchill)

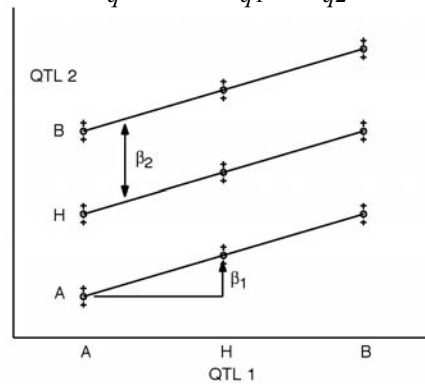


Model

8

additive effects of two QTL (Gary Churchill)

$$G_q = \mu + \beta_{q1} + \beta_{q2}$$



Model

NCSU QTL II: Yandell © 2005

9

Epistasis (Gary Churchill)

The allelic state at one locus can mask or uncover the effects of allelic variation at another.

- W. Bateson, 1907.

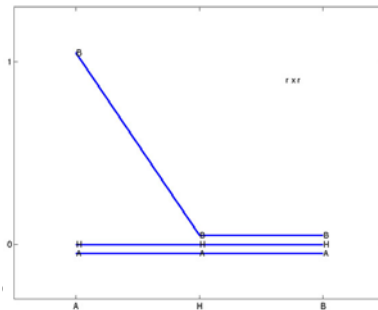
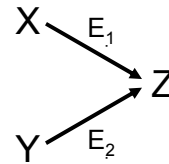
Model

NCSU QTL II: Yandell © 2005

10

epistasis in parallel pathways (GAC)

- Z keeps trait value low
- neither E_1 nor E_2 is rate limiting
- loss of function alleles are segregating from parent A at E_1 and from parent B at E_2

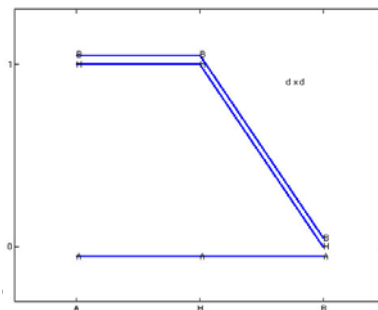


Model

NCSU QTL II: Yandell

epistasis in a serial pathway (GAC)

- Z keeps trait value high
- neither E_1 nor E_2 is rate limiting
- loss of function alleles are segregating from parent B at E_1 and from parent A at E_2



Model

NCSU QTL II: Yandell

QTL with epistasis

- same phenotype model overview

$$Y = G_q + e, \text{var}(e) = \sigma^2$$

- partition of genotypic value with epistasis

$$G_q = \mu + \beta_{q1} + \beta_{q2} + \beta_{q12}$$

- partition of genetic variance & heritability

$$\text{var}(G_q) = \sigma_G^2 = \sigma_1^2 + \sigma_2^2 + \sigma_{12}^2$$

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma^2} = h_1^2 + h_2^2 + h_{12}^2$$

epistatic interactions

- model space issues
 - 2-QTL interactions only?
 - or general interactions among multiple QTL?
 - partition of effects
 - Fisher-Cockerham or tree-structured or ?
- model search issues
 - epistasis between significant QTL
 - check all possible pairs when QTL included?
 - allow higher order epistasis?
 - epistasis with non-significant QTL
 - whole genome paired with each significant QTL?
 - pairs of non-significant QTL?
- Yi Xu (2000) *Genetics*; Yi, Xu, Allison (2003) *Genetics*; Yi *et al.* (2005) *Genetics*

limits of epistatic inference

- power to detect effects
 - epistatic model size grows exponentially
 - $|M| = 3^m$ for general interactions
 - power depends on ratio of n to model size
 - want $n/|M|$ to be fairly large (say > 5)
 - $n = 100, m = 3, n/|M| \approx 4$
- empty cells mess up adjusted (Type 3) tests
 - missing q_1Q_2 / q_1Q_2 or $q_1Q_2q_3 / q_1Q_2q_3$ genotype
 - null hypotheses not what you would expect
 - can confound main effects and interactions
 - can bias AA, AD, DA, DD partition

3. comparing QTL models

- balance model fit with model "complexity"
 - want maximum likelihood
 - without too complicated a model
- information criteria quantifies the balance
 - Bayes information criteria (BIC) for likelihood
 - Bayes factors for Bayesian approach

Bayes factors & BIC

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

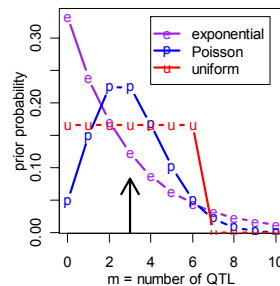
- what is a Bayes factor?
 - ratio of posterior odds to prior odds
 - ratio of model likelihoods
- BF is equivalent to *LR* statistic when
 - comparing two nested models
 - simple hypotheses (e.g. 1 vs 2 QTL)
- BF is equivalent to Bayes Information Criteria (BIC)
 - for general comparison of any models
 - want Bayes factor to be substantially larger than 1 (say 10 or more)

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

Bayes factors and genetic model M

- m = number of QTL
 - prior $\text{pr}(m)$ chosen by user
 - posterior $\text{pr}(m|Y, X)$
 - sampled marginal histogram
 - shape affected by prior $\text{pr}(m)$

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



- pattern of QTL across genome
- gene action and epistasis

choice of multiple QTL priors

- phenotype influenced by genotype & environment
 $\text{pr}(Y|Q=q, \theta) \sim N(G_q, \sigma^2)$, or $Y = G_q + \text{environment}$
- partition genotype-specific mean into QTL effects
 $G_q = \text{mean} + \text{main effects} + \text{epistatic interactions}$
 $G_q = \mu + \beta_q = \mu + \sum_{j \in M} \beta_{qj}$
- priors on mean and effects
 $\mu \sim N(\mu_0, \kappa_0 \sigma^2)$ grand mean
 $\beta_q \sim N(0, \kappa_1 \sigma^2)$ model-independent genotypic effect
 $\beta_{qj} \sim N(0, \kappa_1 \sigma^2 / |M|)$ effects down-weighted by size of M
- determine hyper-parameters via empirical Bayes

$$\mu_0 \approx \bar{Y} \text{ and } \kappa_1 \approx \frac{h^2}{1-h^2} = \frac{\sigma_G^2}{\sigma^2}$$

multiple QTL posteriors

- phenotype influenced by genotype & environment
 $Y|Q=q, \theta \sim N(G_q, \sigma^2)$
- relation of posterior mean to LS estimate

$$G_q | Y, m \sim N(B_q \hat{G}_q, B_q C_q \sigma^2)$$

$$\approx N(\hat{G}_q, C_q \sigma^2)$$

$$\text{LS estimate } \hat{G}_q = \sum_i [\sum_{j \in M} \hat{\beta}_{qji}] = \sum_i w_{qi} Y$$

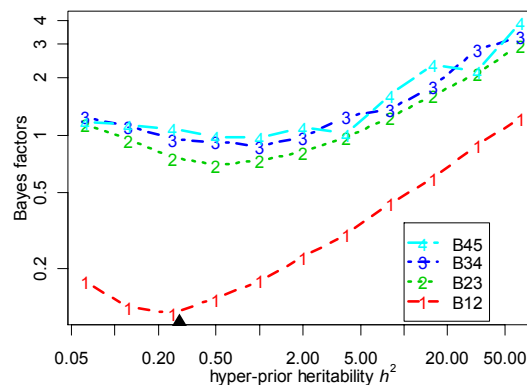
$$\text{variance } V(\hat{G}_q) = \sum_i w_{qi}^2 \sigma^2 = C_q \sigma^2$$

$$\text{shrinkage } B_q = \kappa / (\kappa + C_q) \rightarrow 1$$

issues in computing Bayes factors

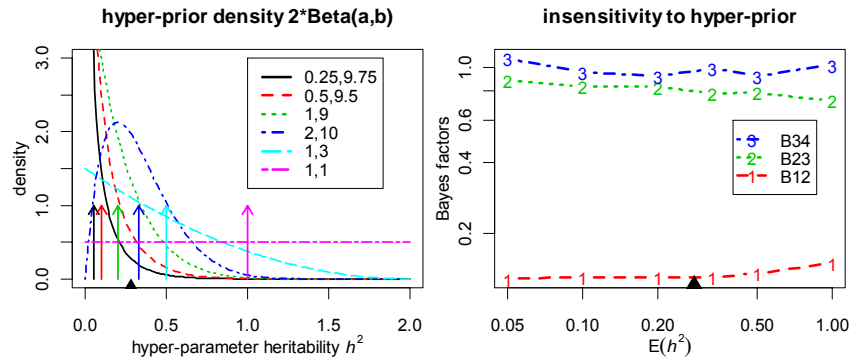
- *BF* insensitive to shape of prior on m
 - geometric, Poisson, uniform
 - precision improves when prior mimics posterior
- *BF* sensitivity to prior variance on effects θ
 - prior variance should reflect data variability
 - resolved by using hyper-priors
 - automatic algorithm; no need for user tuning
- easy to compute Bayes factors from samples
 - sample posterior using MCMC
 - posterior $\text{pr}(m/Y,X)$ is marginal histogram

BF sensitivity to fixed prior for effects



$$\beta_{qj} \sim N(0, \sigma_G^2 / m), \sigma_G^2 = h^2 \sigma_{\text{total}}^2, h^2 \text{ fixed}$$

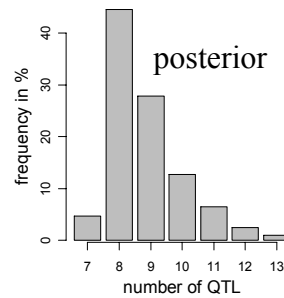
BF insensitivity to random effects prior



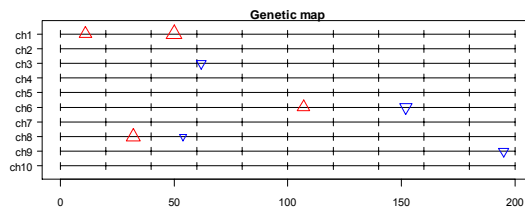
$$\beta_{qj} \sim N(0, \sigma_G^2 / m), \sigma_G^2 = h^2 \sigma_{total}^2, \frac{1}{2} h^2 \sim \text{Beta}(a, b)$$

4. simulations and data studies

- 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

<i>m</i>	<u>1</u>	2	<u>3</u>	4	5	<u>6</u>	7	<u>8</u>	<u>9</u>	10	Count of 8000
8	<u>2</u>	0	1	0	0	2	0	2	1	0	3371
9	<u>3</u>	0	1	0	0	2	0	2	1	0	751
7	2	0	1	0	0	2	0	<u>1</u>	1	0	377
9	2	0	1	0	0	2	0	2	1	0	218
9	2	0	1	0	0	<u>3</u>	0	2	1	0	218
9	2	0	1	0	0	2	0	2	<u>2</u>	0	198

B. napus 8-week vernalization whole genome study

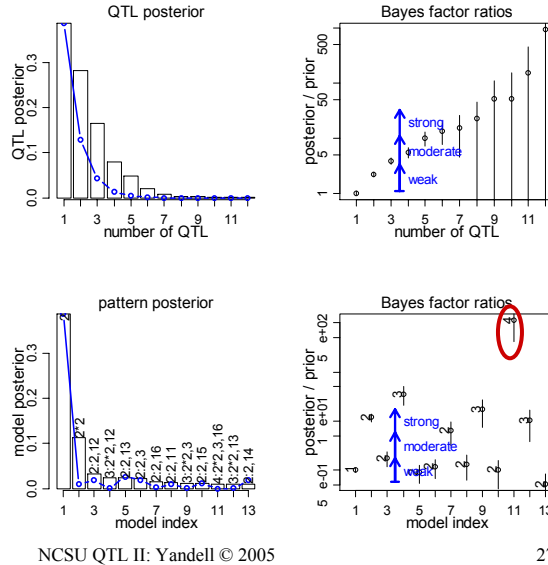
- 108 plants from double haploid
 - similar genetics to backcross: follow 1 gamete
 - parents are Major (biennial) and Stellar (annual)
- 300 markers across genome
 - 19 chromosomes
 - average 6cM between markers
 - median 3.8cM, max 34cM
 - 83% markers genotyped
- phenotype is days to flowering
 - after 8 weeks of vernalization (cooling)
 - Stellar parent requires vernalization to flower
- Ferreira et al. (1994); Kole et al. (2001); Schranz et al. (2002)

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



Model

NCSU QTL II: Yandell © 2005

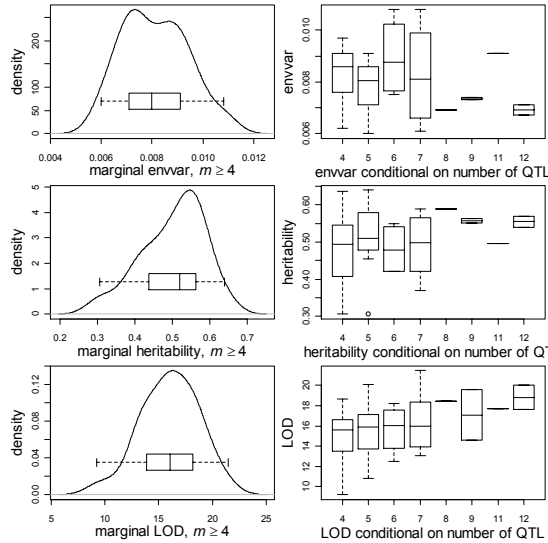
27

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



Model

NCSU QTL II: Yandell © 2005

28

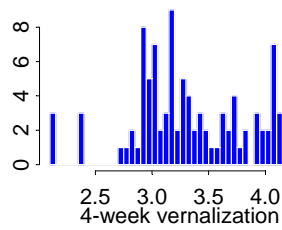
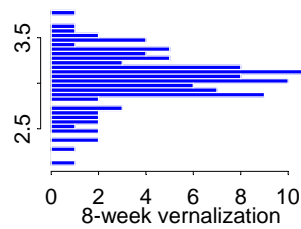
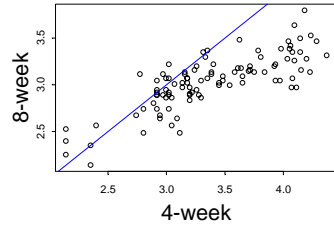
examples in detail

- days to flower for *Brassica napus* (plant) ($n = 108$)
 - single chromosome with 2 linked loci
 - whole genome
- gonad shape in *Drosophila* spp. (insect) ($n = 1000$)
 - multiple traits reduced by PC
 - many QTL and epistasis
- expression phenotype (SCD1) in mice ($n = 108$)
 - multiple QTL and epistasis
- obesity in mice ($n = 421$)
 - epistatic QTLs with no main effects

Brassica napus: 1 chromosome

- 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



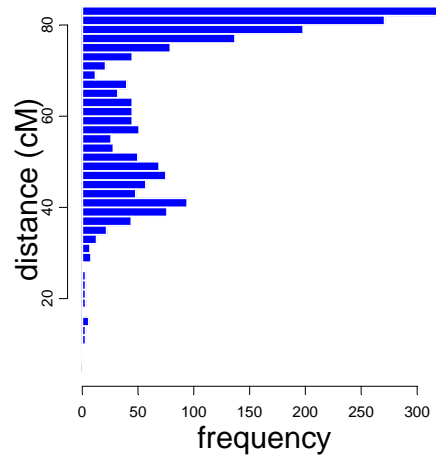
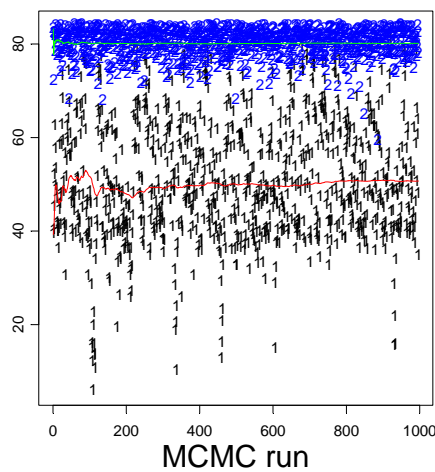
summaries of raw data
joint scatter plots
(identity line)
separate histograms

Data

NCSU QTL II: Yandell © 2005

3

Brassica 8-week data locus MCMC with $m=2$



Data

NCSU QTL II: Yandell © 2005

4

4-week vs 8-week vernalization

4-week vernalization

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

8-week vernalization

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

cM	add	cM	add
40	.30	40	.06
80	.16	80	.13

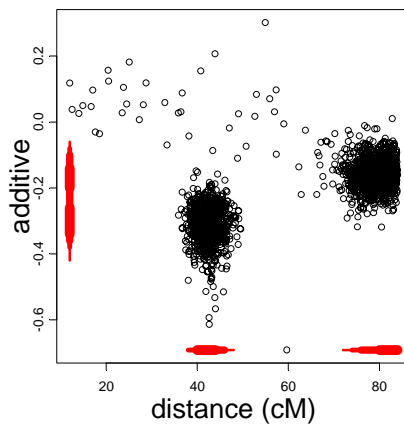
Data

NCSU QTL II: Yandell © 2005

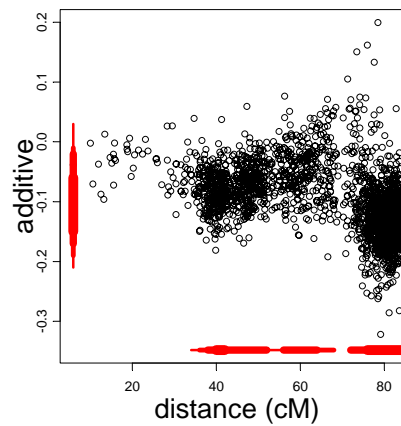
5

Brassica credible regions

4-week



8-week

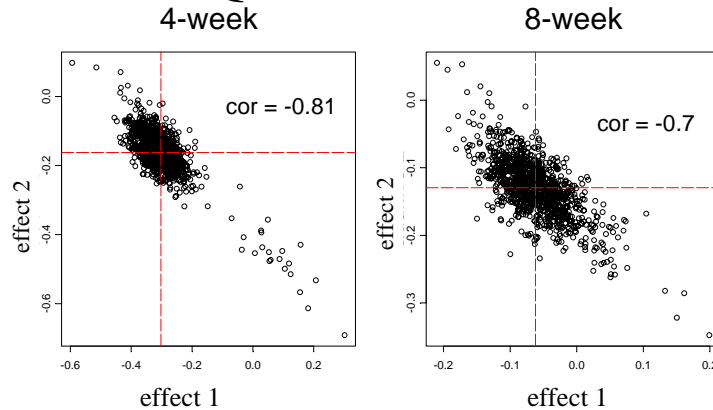


Data

NCSU QTL II: Yandell © 2005

6

collinear QTL = correlated effects



- linked QTL = collinear genotypes
 - correlated estimates of effects (negative if in coupling phase)
 - sum of linked effects usually fairly constant

Data

NCSU QTL II: Yandell © 2005

7

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
- Ferreira et al. (1994); Kole et al. (2001); Schranz et al. (2002)

Data

NCSU QTL II: Yandell © 2005

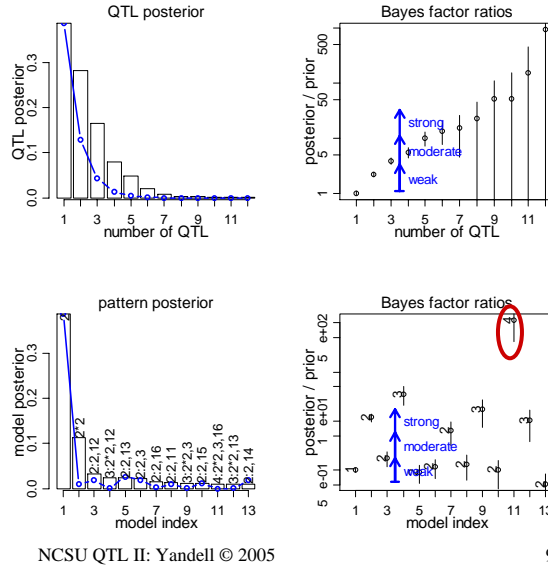
8

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



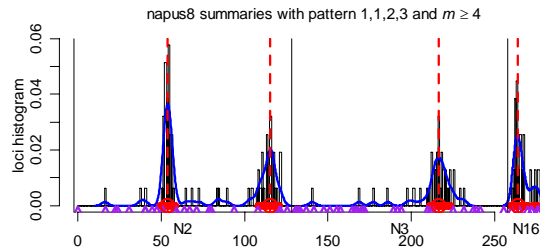
Data

NCSU QTL II: Yandell © 2005

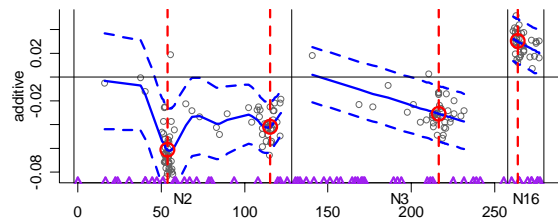
9

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



Data

NCSU QTL II: Yandell © 2005

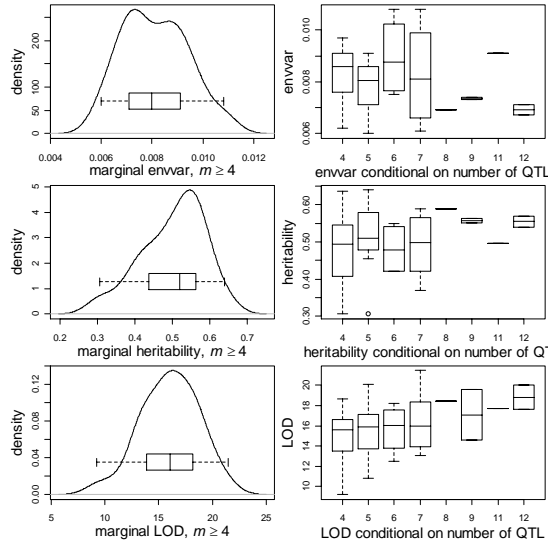
10

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



Data

NCSU QTL II: Yandell © 2005

11

shape phenotype in BC study indexed by PC1

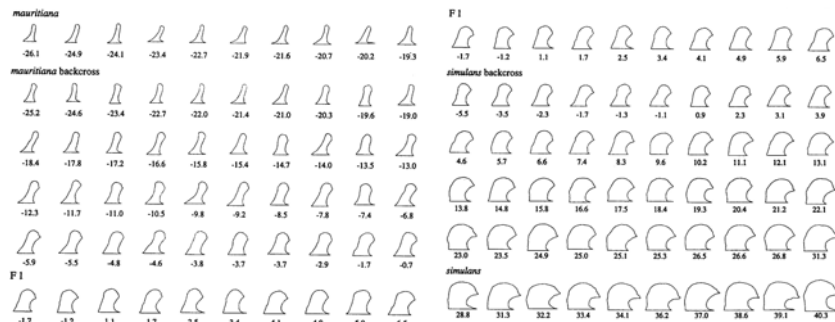


FIGURE 6.—Outlines of the posterior lobe from a sample of individuals from each of the five groups: pure *mauritaniana*, *mauritaniana* backcross, F_1 , *simonsi* backcross, and pure *simonsi*. Within each group, the outlines are presented in order of their PC1 score (sampled at even intervals from the 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*

Data

NCSU QTL II: Yandell © 2005

12

shape phenotype via PC

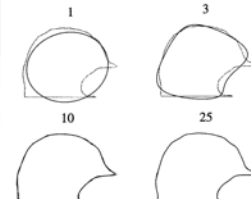
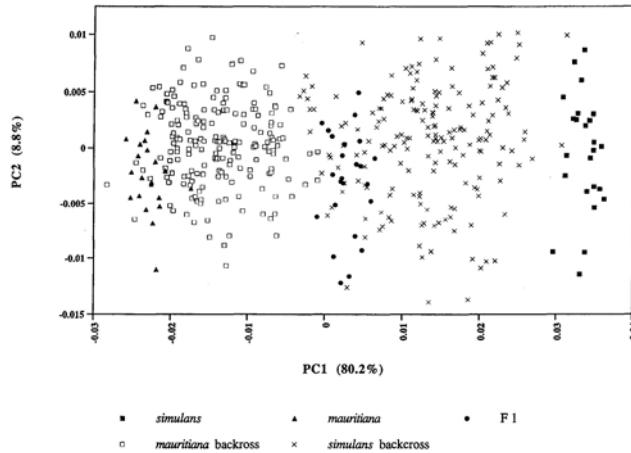


FIGURE 2.—The effect of harmonic number on the accuracy of reconstruction of a posterior lobe outline by elliptical Fourier analysis.

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. Liu et al. (1996) *Genetics*

Data

NCSU QTL II: Yandell © 2005

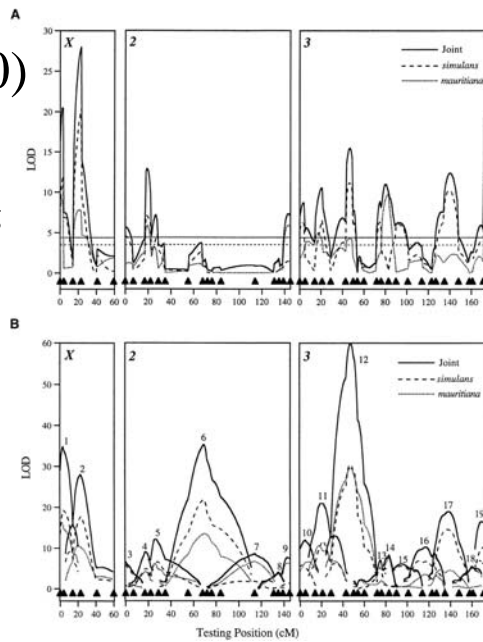
13

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"

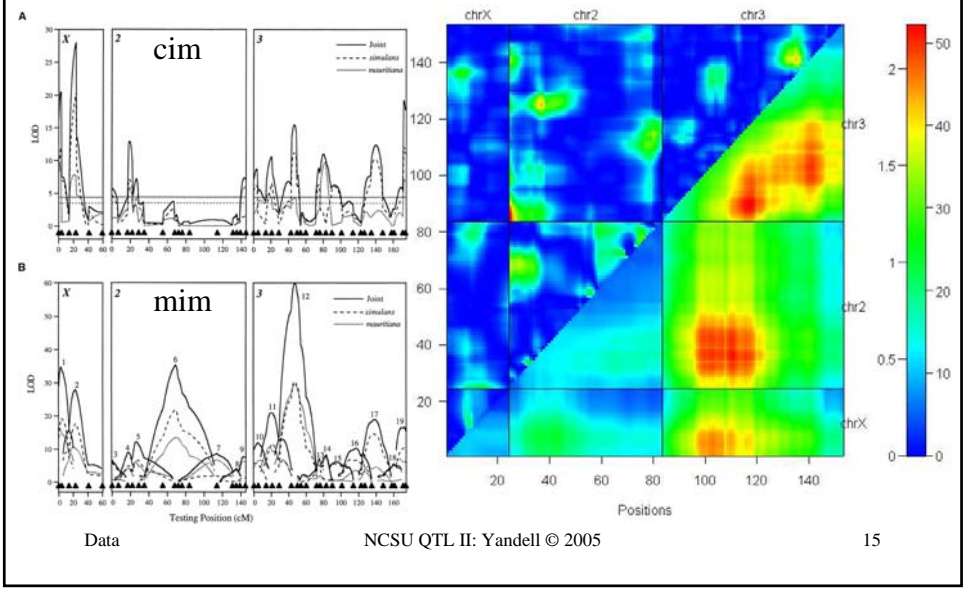


Data

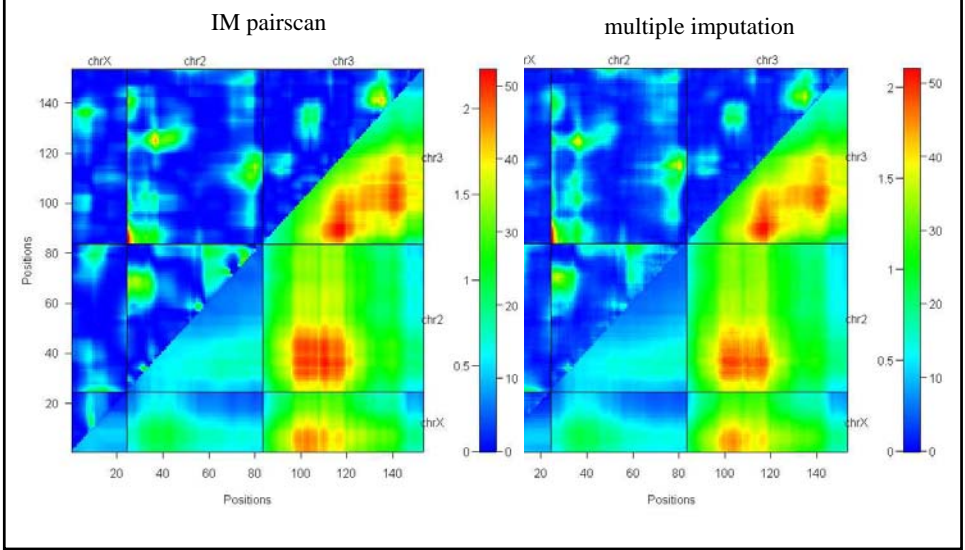
NCSU QTL II: Yandell © 2005

14

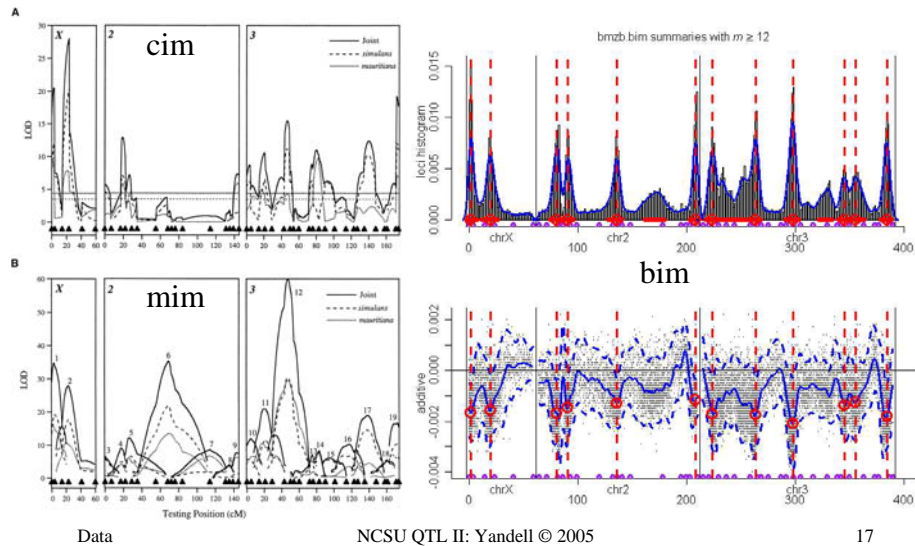
CIM, MIM and IM pairscan



2 QTL + epistasis: IM versus multiple imputation



multiple QTL: CIM, MIM and BIM



studying diabetes in an F2

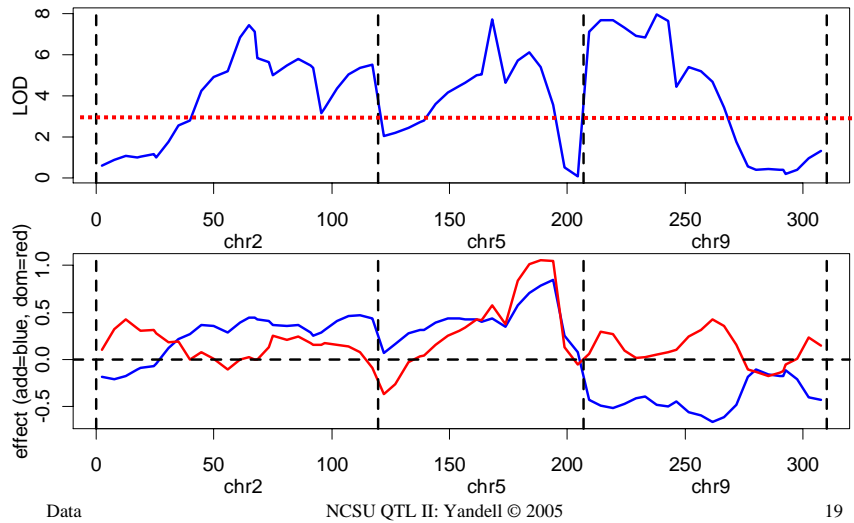
- segregating cross of inbred lines
 - B6.ob x BTBR.ob → F1 → F2
 - selected mice with ob/ob alleles at leptin gene (chr 6)
 - measured and mapped body weight, insulin, glucose at various ages (Stoehr et al. 2000 Diabetes)
 - sacrificed at 14 weeks, tissues preserved
- gene expression data
 - Affymetrix microarrays on parental strains, F1
 - key tissues: adipose, liver, muscle, β -cells
 - novel discoveries of differential expression (Nadler et al. 2000 PNAS; Lan et al. 2002 in review; Ntambi et al. 2002 PNAS)
 - RT-PCR on 108 F2 mice liver tissues
 - 15 genes, selected as important in diabetes pathways
 - SCD1, PEPCK, ACO, FAS, GPAT, PPARgamma, PPARalpha, G6Pase, PDI,...

Data

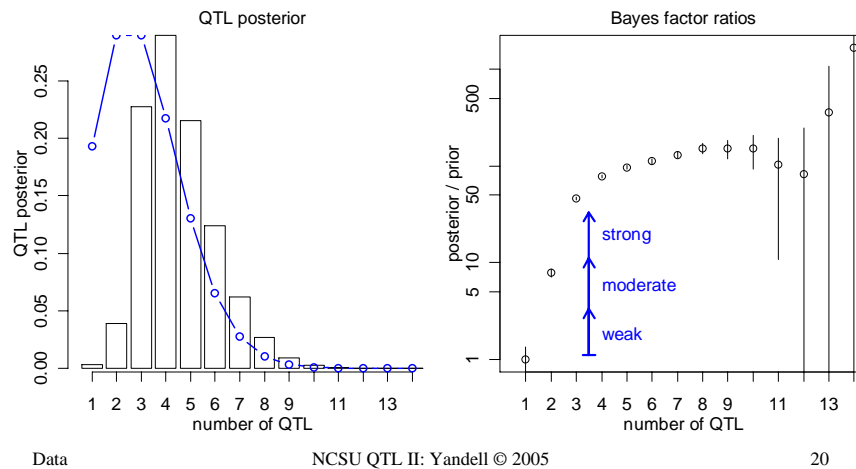
NCSU QTL II: Yandell © 2005

18

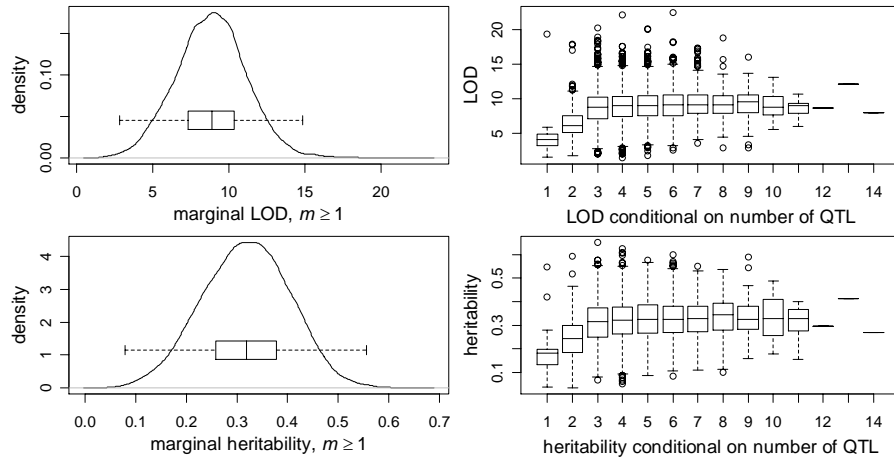
Multiple Interval Mapping (QTLCart) SCD1: multiple QTL plus epistasis!



Bayesian model assessment: number of QTL for SCD1



Bayesian LOD and h^2 for SCD1

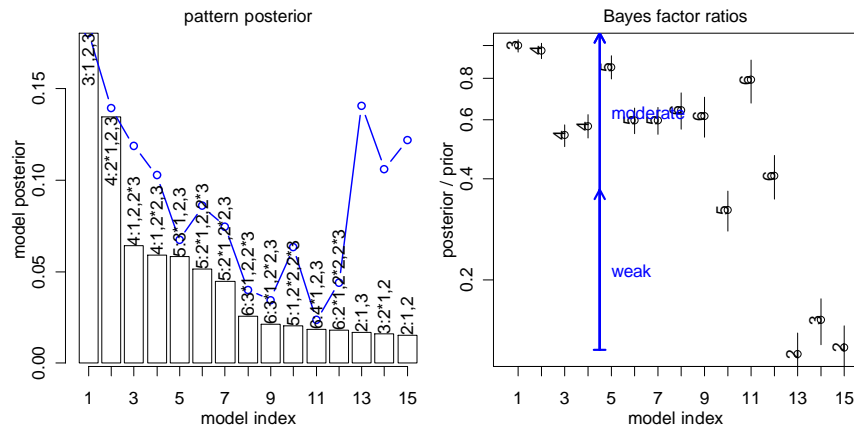


Data

NCSU QTL II: Yandell © 2005

21

Bayesian model assessment: chromosome QTL pattern for SCD1



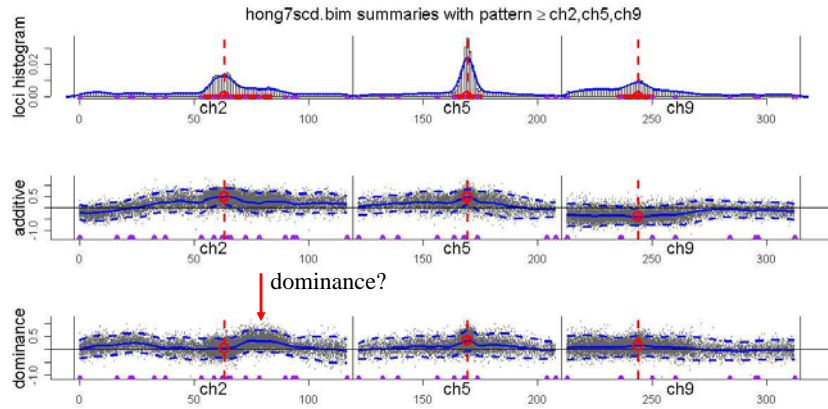
Data

NCSU QTL II: Yandell © 2005

22

trans-acting QTL for SCD1

(no epistasis yet: see Yi, Xu, Allison 2003)

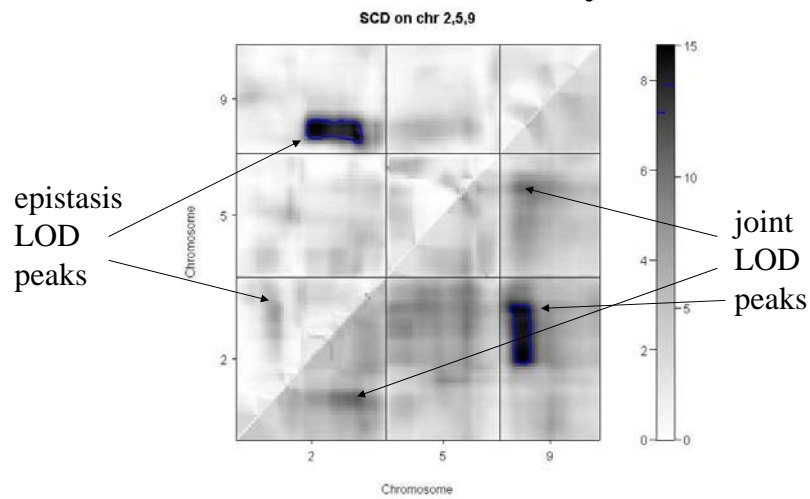


Data

NCSU QTL II: Yandell © 2005

23

2-D scan: assumes only 2 QTL!

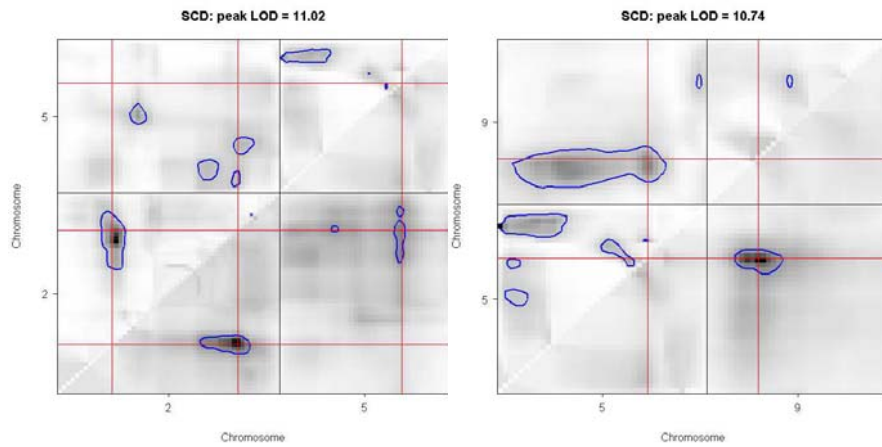


Data

NCSU QTL II: Yandell © 2005

24

sub-peaks can be easily overlooked!

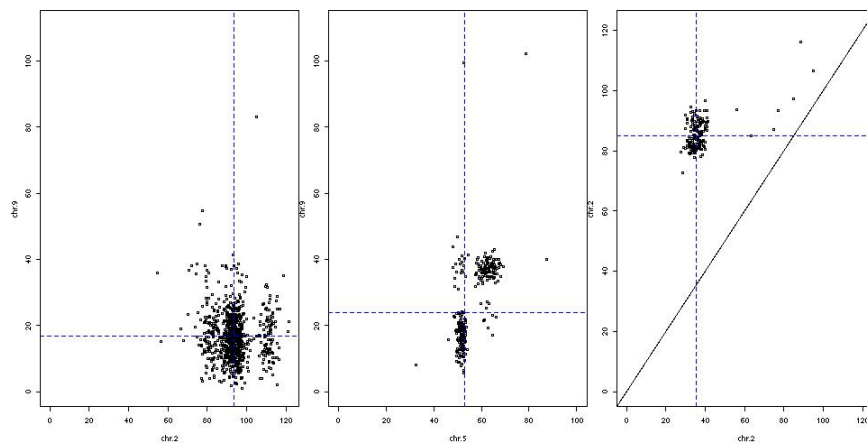


Data

NCSU QTL II: Yandell © 2005

25

epistatic model fit

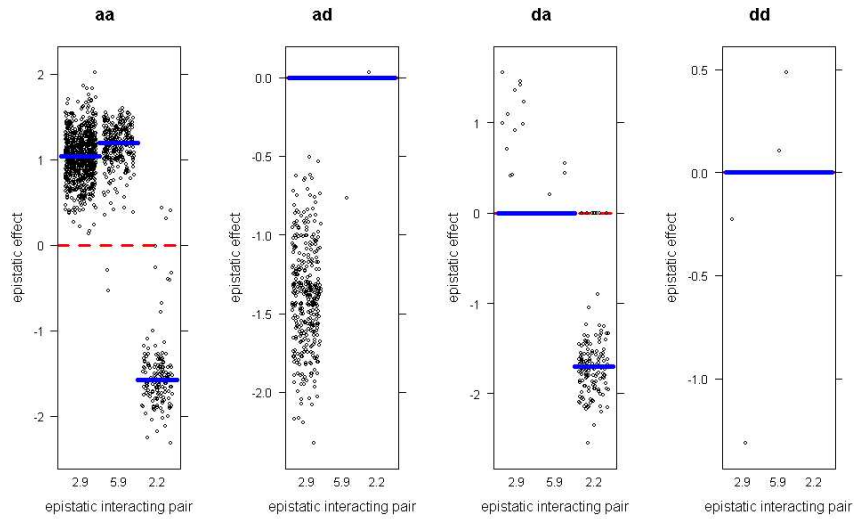


Data

NCSU QTL II: Yandell © 2005

26

Cockerham epistatic effects



Data

NCSU QTL II: Yandell © 2005

27

obesity in CAST/Ei BC onto M16i

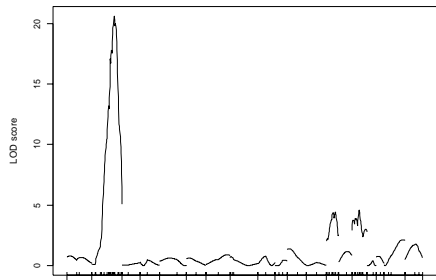
- 421 mice (Daniel Pomp)
 - (213 male, 208 female)
- 92 microsatellites on 19 chromosomes
 - 1214 cM map
- subcutaneous fat pads
 - pre-adjusted for sex and dam effects
- Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005) *Genetics* (in press)

Data

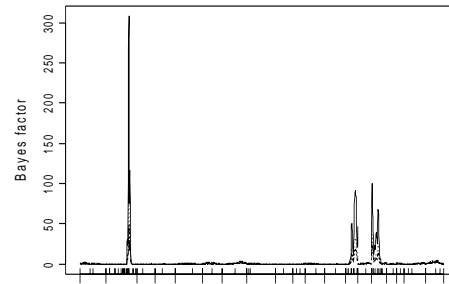
NCSU QTL II: Yandell © 2005

28

non-epistatic analysis



single QTL LOD profile



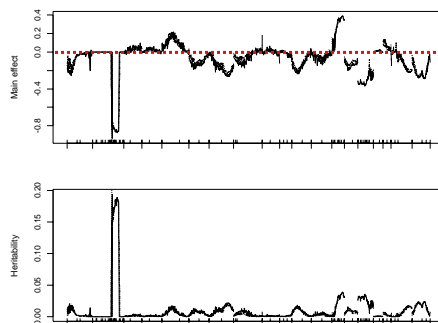
multiple QTL
Bayes factor profile

Data

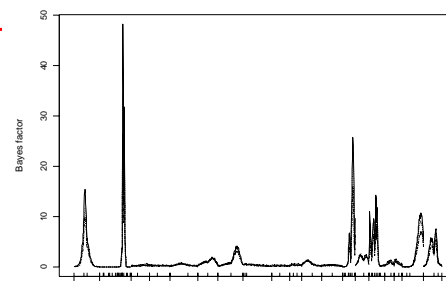
NCSU QTL II: Yandell © 2005

29

posterior profile of main effects in epistatic analysis



main effects & heritability profile



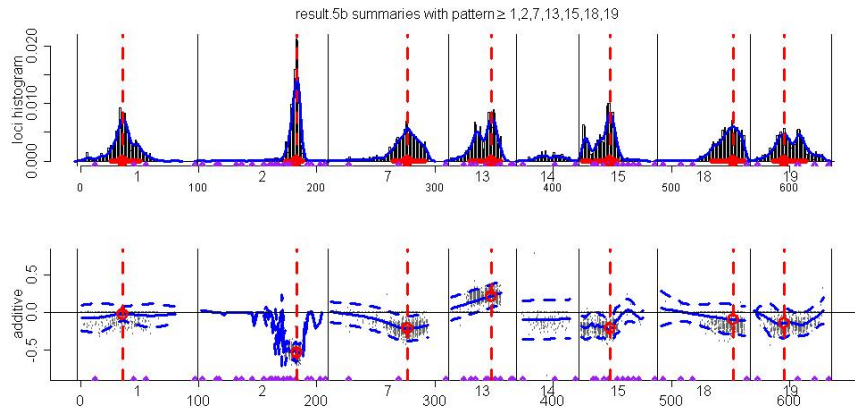
Bayes factor profile

Data

NCSU QTL II: Yandell © 2005

30

posterior profile of main effects in epistatic analysis

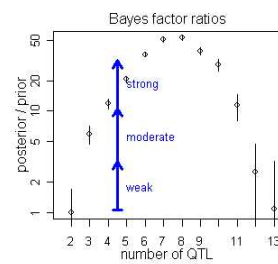
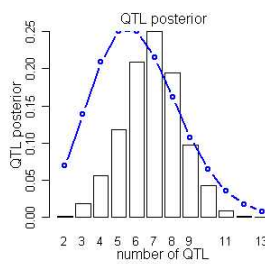


Data

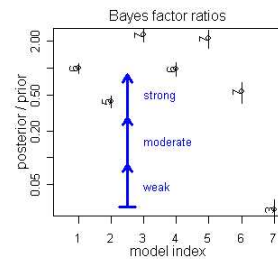
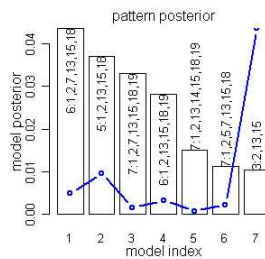
NCSU QTL II: Yandell © 2005

31

model selection
via
Bayes factors
for
epistatic model



number of QTL
QTL pattern

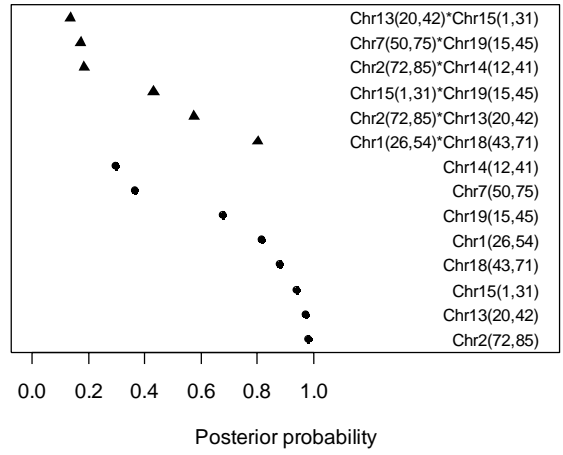


Data

NCSU QTL II: Yandell © 2005

32

posterior probability of effects

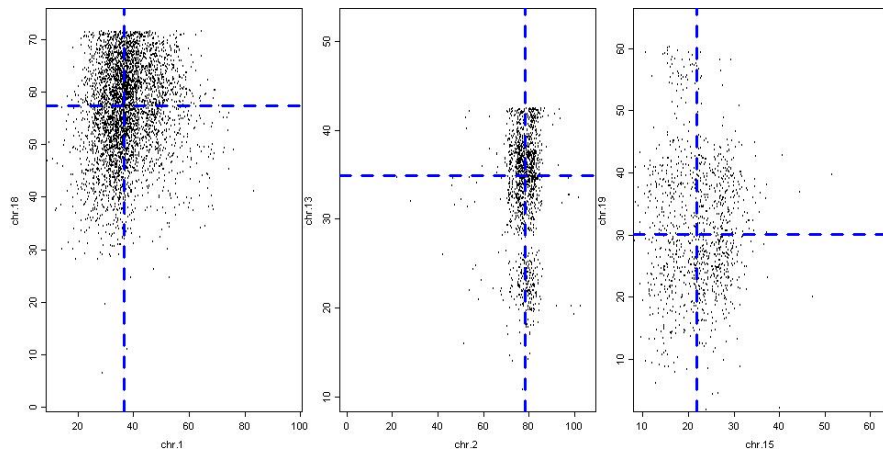


Data

NCSU QTL II: Yandell © 2005

33

scatterplot estimates of epistatic loci

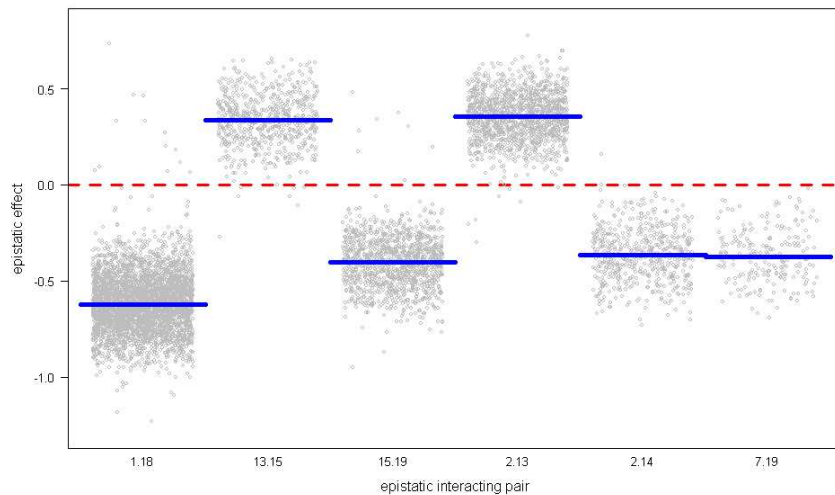


Data

NCSU QTL II: Yandell © 2005

34

stronger epistatic effects



Data

NCSU QTL II: Yandell © 2005

35

our RJ-MCMC software

- R: www.r-project.org
 - freely available statistical computing application R
 - library(bim) builds on Broman's library(qtl)
- QTLCart: statgen.ncsu.edu/qtlcart
 - Bmapqtl incorporated into QTLCart (S Wang 2003)
- www.stat.wisc.edu/~yandell/qtl/software/bmqtl
- R/bim
 - initially designed by JM Satagopan (1996)
 - major revision and extension by PJ Gaffney (2001)
 - whole genome, multivariate and long range updates
 - speed improvements, pre-burnin
 - built as official R library (H Wu, Yandell, Gaffney, CF Jin 2003)
- R/bmqtl
 - collaboration with N Yi, H Wu, GA Churchill
 - initial working module: Winter 2005
 - improved module and official release: Summer/Fall 2005
 - major NIH grant (PI: Yi)

Data

NCSU QTL II: Yandell © 2005

36