7 multiple QTL implementation

- maximum likelihood using EM
  - composite interval mapping (Zeng)
  - multiple interval mapping (Kao Zeng)
  - sequential testing to search model space
- multiple imputation (Sen Churchill)
  - Bayesian log posterior odds
  - sequential testing and pairwise plots to search
- MCMC (Satagopan Yandell Newton)
  - Bayes factors and marginal posteriors
  - Markov chain sampling to search model space

multiple QTL likelihood

- likelihood is mixture over unknown QTL
  - likelihood = product of sum of products
  - now have multiple QTL
    - $Q = (Q_1, Q_2, \ldots, Q_m)$
    - $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$
    - $\theta = (\mu, \theta_1, \theta_2, \ldots, \theta_m, \sigma^2)$ plus interactions…

$$
\begin{aligned}
L(\theta, \lambda \mid Y, X) &= \mathrm{pr}(Y \mid X, \theta, \lambda) \\
&= \mathrm{prod}_i \ \mathrm{pr}(Y_i \mid X_i, \theta, \lambda) \\
&= \mathrm{prod}_i \ \mathrm{sum}_Q \ \mathrm{pr}(Q \mid X_i, \lambda)\mathrm{pr}(Y_i \mid Q, \theta)
\end{aligned}
$$

maximum likelihood (ML) idea

- pick QTL loci λ
  - scan whole genome for single QTL with others fixed
  - jointly scan small part of genome for two QTL
- find ML estimates of gene action θ given λ
  - maximum likelihood at peak of likelihood
  - slope is weighted average using posteriors for Q
  - estimate depends on θ -- need to iterate (again!)

$$\frac{dL(\theta,\lambda \mid Y,X)}{d\theta} = \text{sum}_{i,Q}\ \text{pr}(Q \mid Y_i, X_i, \theta, \lambda) \frac{d\log(\text{pr}(Y_i \mid Q,\theta))}{d\theta}$$

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

EM method for interval mapping

- E-step: estimate posterior recombination
  - P_{Qi} = pr(Q | Y_i, X_i, θ, λ)
  - individual i, genotype $Q = (Q_1, Q_2, \ldots, Q_m)$
  - depends on multiple QTL loci λ and effects θ
- M-steps: maximize likelihood for θ
  - technical point: caution on parallel updates
    - ECM: conditional maximization depending
    - update in series, based on what has been done already
  - depends on P_{Qi}
  - see Kao Zeng papers for details (hard problem!)

$$0 = \text{sum}_{i,Q}\ P_{Qi} \frac{d\log(\text{pr}(Y_i \mid Q,\theta))}{d\theta}$$

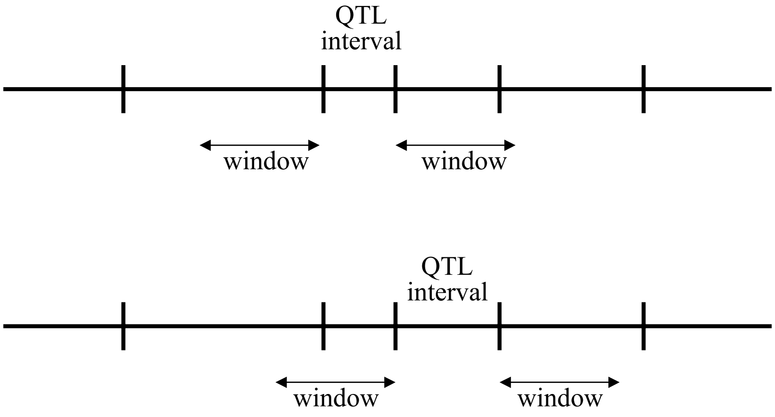
MIM search for QTL

- initial model: use CIM or marker regression
- refine model: simultaneous search
  - iteratively scan QTL to improve each locus
  - fix all other loci temporarily
- add QTL or epistatic interactions
  - forward selection: scan genome for new QTL
  - check all pairs of significant QTL for epistasis
  - scan genome for interactions with significant QTL
- drop QTL that may now be non-significant
  - backward elimination

CIM model

- CIM (or MQM) approximates MIM
  - scan for one QTL while fixing all others
  - use markers as "cofactors" for other QTL
- art to selection of cofactors
  - pick by stepwise marker regression
    - with model selection criterion
  - eliminate markers "too close" to current interval
    - wide window screens out loosely linked QTL
    - narrow window washes out effect of desired QTL
    - no firm criteria--depends on effects, marker spacing

CIM idea



QTL interval

window

window

QTL interval

window

window

examples of MIM

- Liu et al. (1996); Zeng et al. (2000)
  - gonad shape on *Drosophila* spp.
  - reciprocal backcross, 500 individuals per cross
  - 19 QTL
- Weber et al. (1999)
  - *Drosophila* wing shape
  - 700 RI individuals
  - 19 QTL plus 10 interactions

phenotype: outline of posterior lobe (Liu et al. 1996)

Shape Analysis in Drosophila 1131

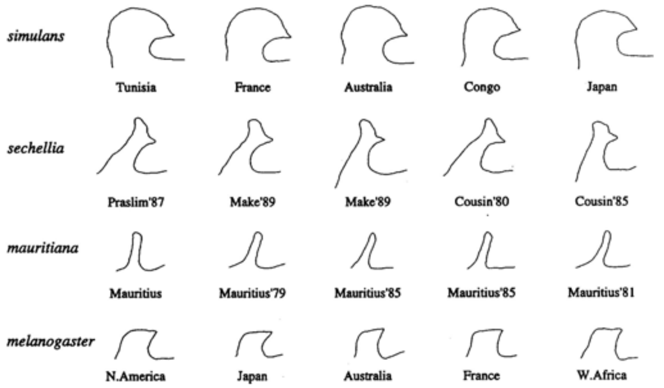
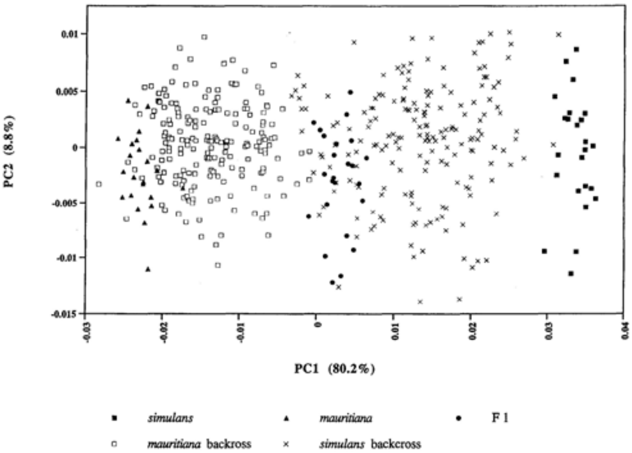


FIGURE 1.—Posterior lobe outlines from a sample of five isofemale lines from each of four Drosophila species. Additional information about the lines is provided in Table 1.

principal components



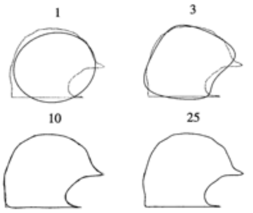


FIGURE 2.—The effect of harmonic number on the accuracy of reconstruction of a posterior lobe outline by elliptical Fourier analysis.

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*

sample shapes for PC1 in parents, F1 and backcrosses

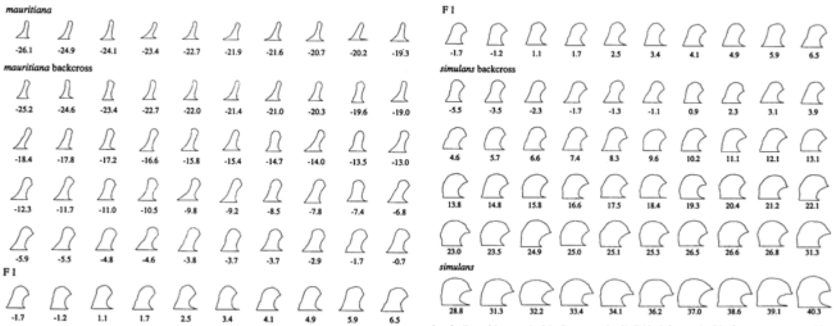
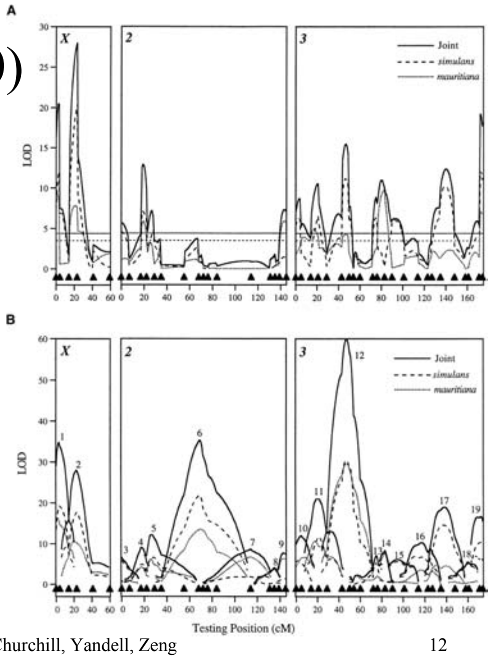


FIGURE 6.—Outlines of the posterior lobe from a sample of individuals from each of the five groups: pure *mauritiana*, *mauritiana* backcross, F_1, *simulans* backcross, and pure *simulans*. Within each group, the outlines are presented in order of their PC1 score (sampled at even intervals from the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

Zeng et al. (2000) CIM vs. MIM

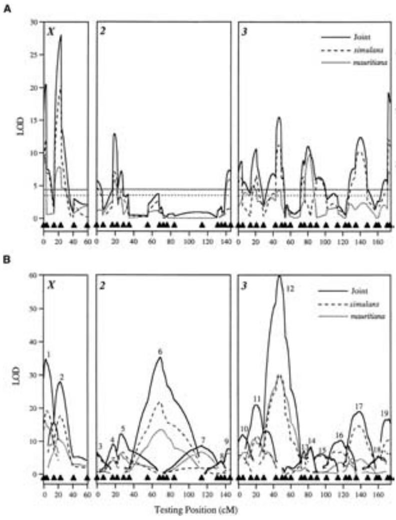


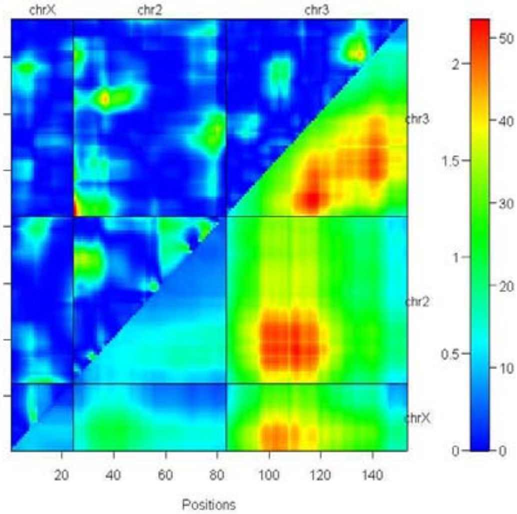
composite interval mapping
- (Liu et al. 1996)
- narrow peaks
- miss some QTL

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

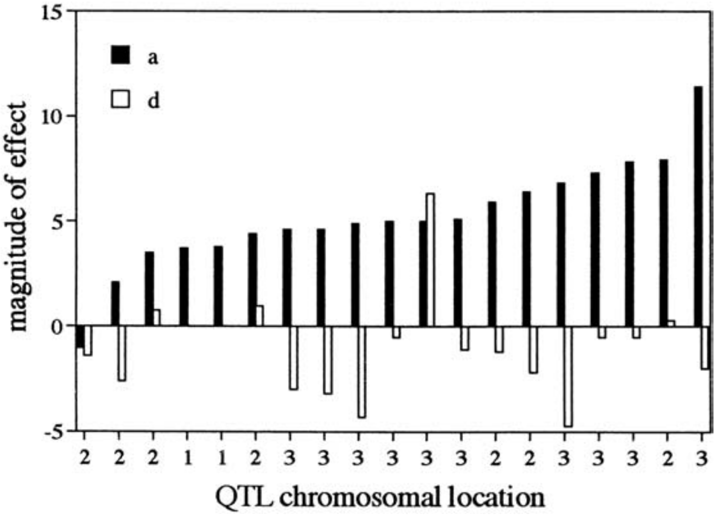
both conditional 1-D scans
- fixing all other "QTL"

pairs plot: CIM, MIM and pairscan





MIM effects for *Dm*



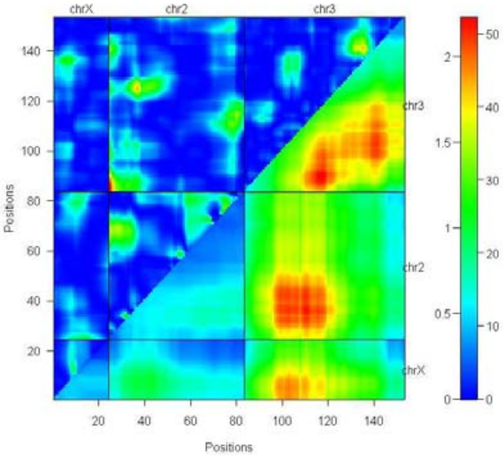
Bayesian model posterior

- augment data (*Y,X*) with unknowns *Q*
  - study unknowns (θ,λ,Q) given data (*Y,X*)
  - $Q \sim \text{pr}(Q \mid Y_i, X_i, \theta, \lambda)$
- sample genotypes *Q* for every individual at *m* QTL
  - and multiple loci λ and multiple genetic effects and epistasis θ
- study properties of posterior
  - use priors that are independent between QTL
  - draw samples from posterior in some clever way
    - multiple imputation or MCMC

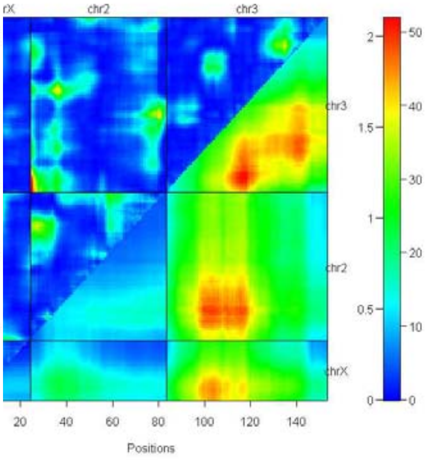
$$\text{pr}(\theta,\lambda,Q \mid Y,X) = \frac{\text{pr}(Q \mid X,\lambda)\text{pr}(Y \mid Q,\theta)\text{pr}(\lambda \mid X)\text{pr}(\theta)}{\text{pr}(Y \mid X)}$$

2 QTL: EM versus multiple imputation for *Dm*

EM method



imputation method

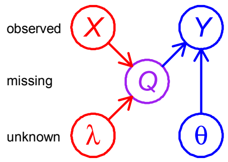


MCMC idea for QTLs

- 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

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

sample from full conditionals for model with *m* QTL



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

Bayesian interval mapping

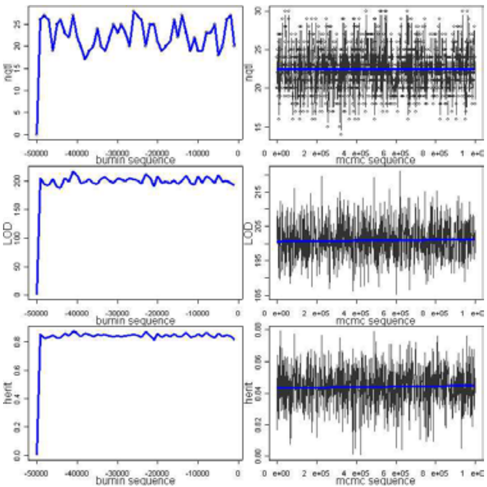
- sample missing genotypes Q
- decouple effects θ from QTL λ
- but Q depends on (θ,λ) and vice versa
- also need to specify priors

$$\begin{tikzcd}
\lambda \sim \dfrac{\mathrm{pr}(Q \mid X,\lambda)\mathrm{pr}(\lambda \mid X)}{\mathrm{pr}(Q \mid X)} \arrow[d, bend left, leftrightarrow] \arrow[dd, bend left=60, leftrightarrow] \\
Q \sim \mathrm{pr}(Q \mid Y_i, X_i, \theta, \lambda) \arrow[d, bend left, leftrightarrow] \\
\theta \sim \dfrac{\mathrm{pr}(Y \mid Q,\theta)\mathrm{pr}(\theta)}{\mathrm{pr}(Y \mid Q)}
\end{tikzcd}$$

MCMC diagnostics for *Dm* shape

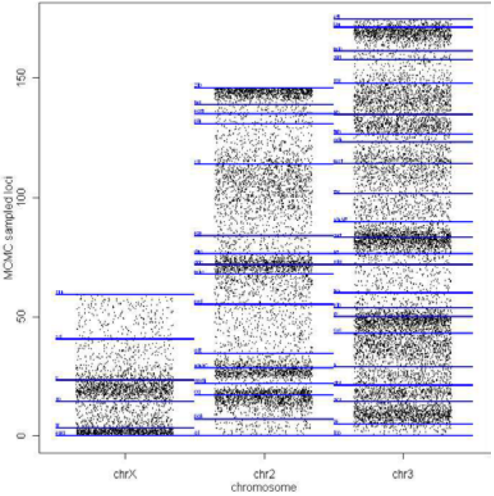
- m ~ Poisson(15) prior on number of QTL
- Bayesian LOD (log posterior density)
- Heritability

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



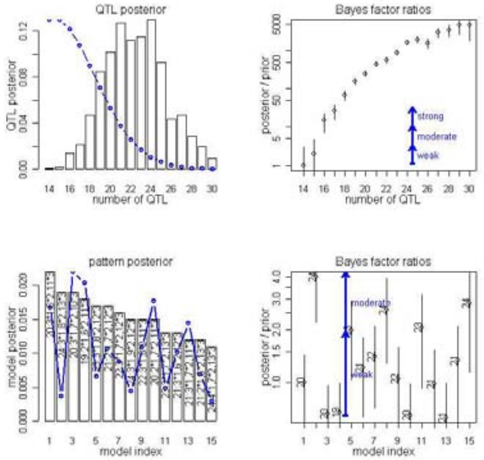
MCMC sampled loci

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



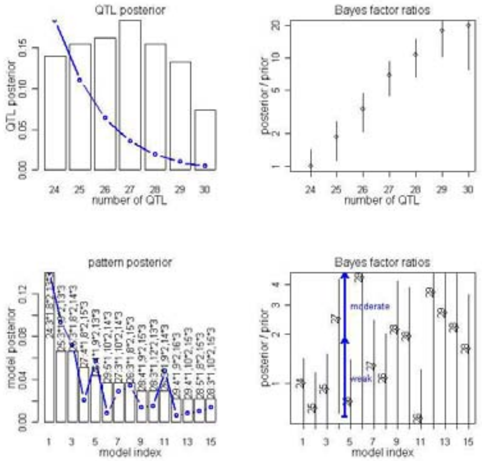
MCMC model selection

- *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



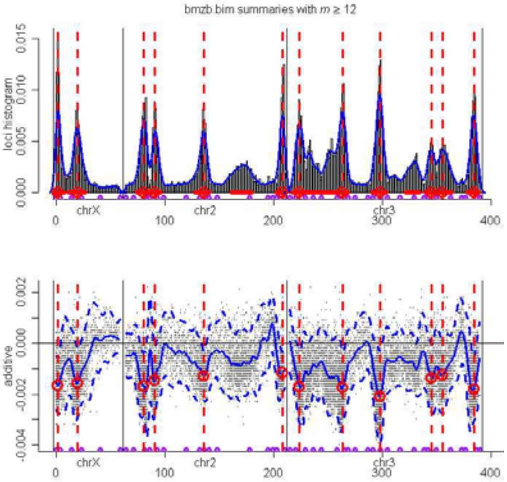
MCMC model selection restricted to “better models”

- 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

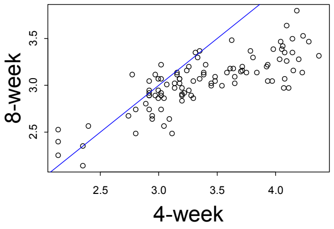
- model averaging
  - over all models
  - 1000 samples
- histogram of loci
  - marginal posteriors
  - superimposed on genome
  - 12 peaks identified
- scatterplot: loci & effects
  - smoothed mean ± 2 SE

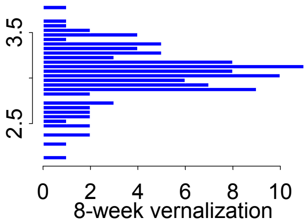


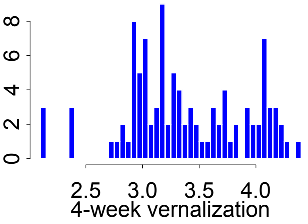
Brassica napus data

- 4-week & 8-week vernalization effect
  - log(days to flower)
- genetic cross of
  - Stellar (annual canola)
  - Major (biennial rapeseed)
- 105 F1-derived double haploid (DH) lines
  - homozygous at every locus (*QQ* or *qq*)
- 10 molecular markers (RFLPs) on LG9
  - two QTLs inferred on LG9 (now chromosome N2)
  - corroborated by Butruille (1998)
  - exploiting synteny with *Arabidopsis thaliana*

Brassica 4- & 8-week data

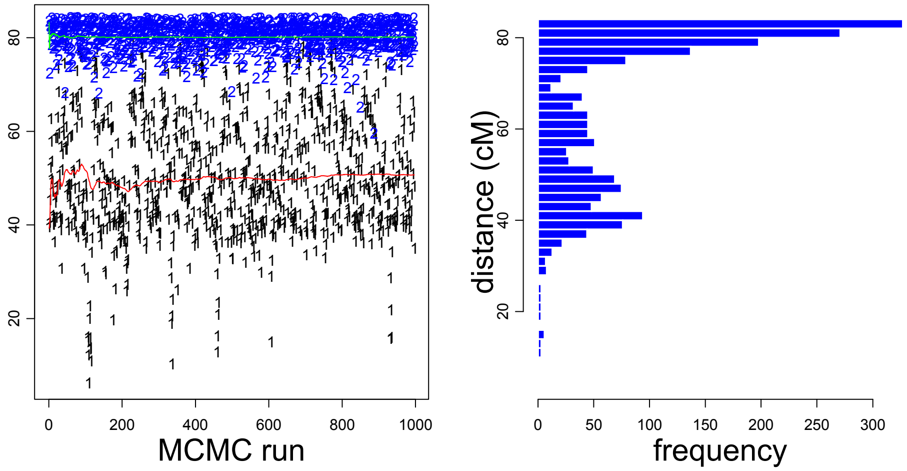






summaries of raw data
joint scatter plots
(identity line)
separate histograms

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



4-week vs 8-week vernalization

4-week vernalization

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

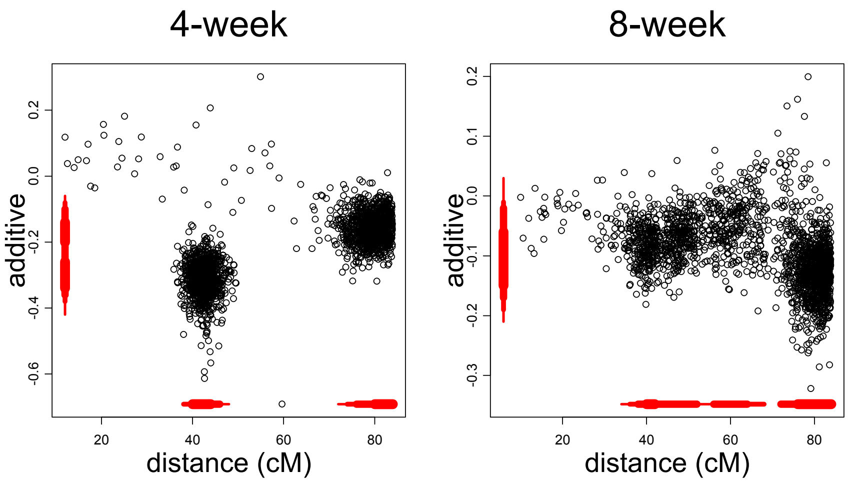
cM	add
40	.30
80	.16

8-week vernalization

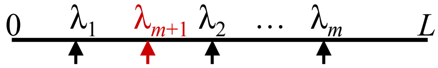
- shorter time to flower
- larger LOD at 80cM
- modest LOD at 40cM
- loci poorly determined

cM	add
40	.06
80	.13

Brassica credible regions



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

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

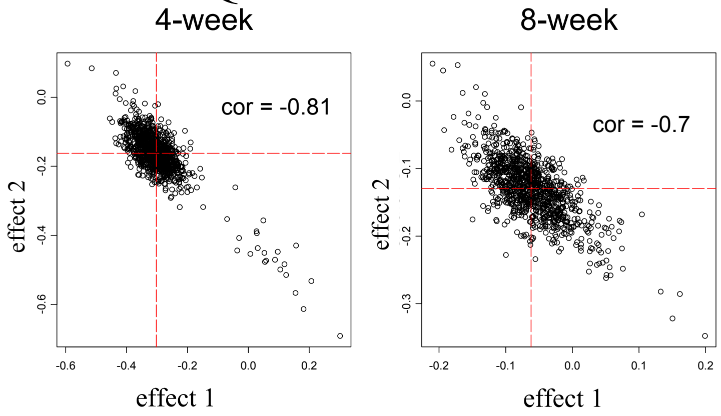
Model Selection in Regression

- 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

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

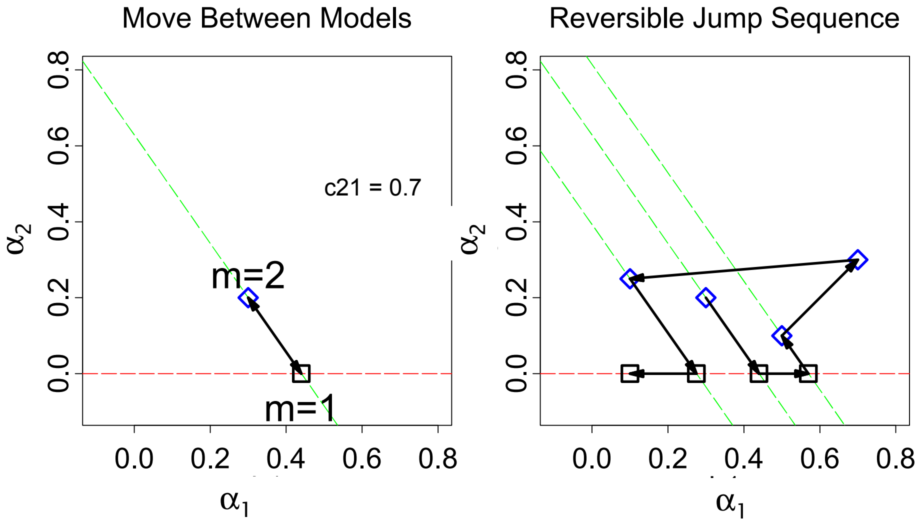
$$m = 2 : Y_i = \mu + \alpha_1 (Q_{i1} - \overline{Q}_1) + \alpha_2 (Q_{i1} - \overline{Q}_1) + e_i$$

collinear QTL = correlated effects

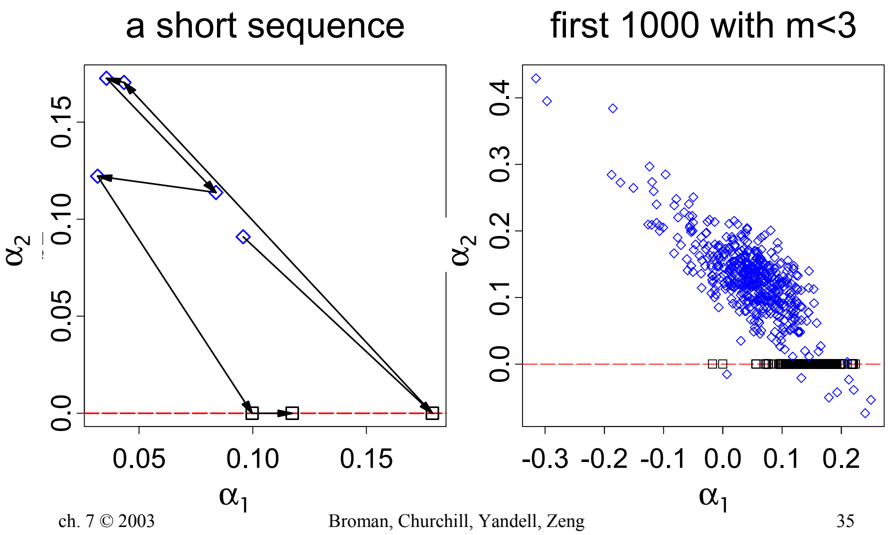


- linked QTL: collinear genotypes & correlated effect estimates
  - sum of linked effects usually well determined
- which QTL to go after in breeding, genome walking?

Geometry of Reversible Jump

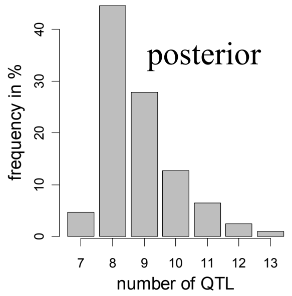


QT additive Reversible Jump

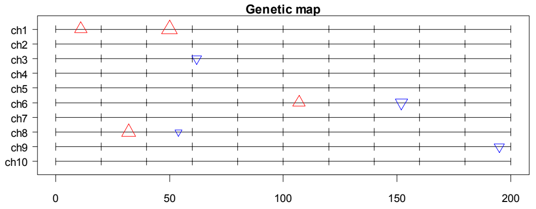


a complicated simulation

- 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

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

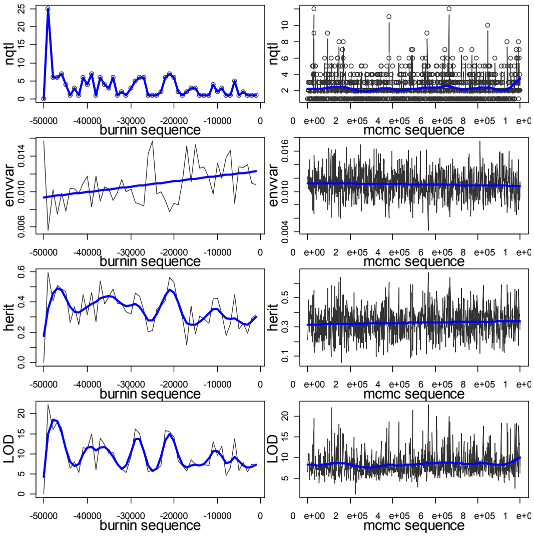
B. napus 8-week vernalization whole genome study

- 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

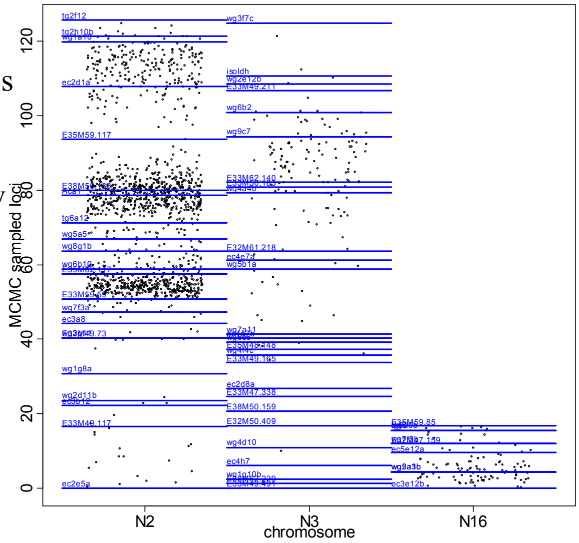


MCMC sampled loci

subset of chromosomes
N2, N3, N16

points jittered for view
blue lines at markers

note concentration
on chromosome N2

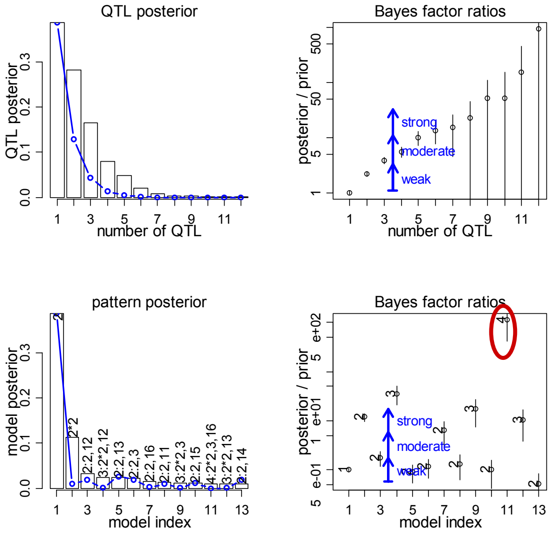


Bayesian model assessment

row 1: # QTL
row 2: pattern

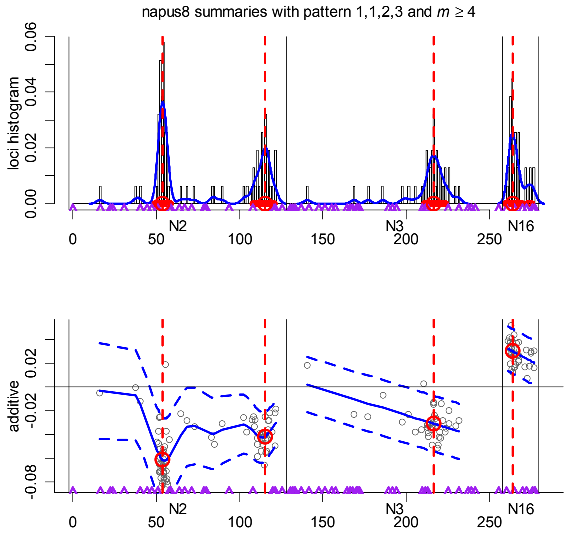
col 1: posterior
col 2: Bayes factor
note error bars on bf

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



Bayesian estimates of loci & effects

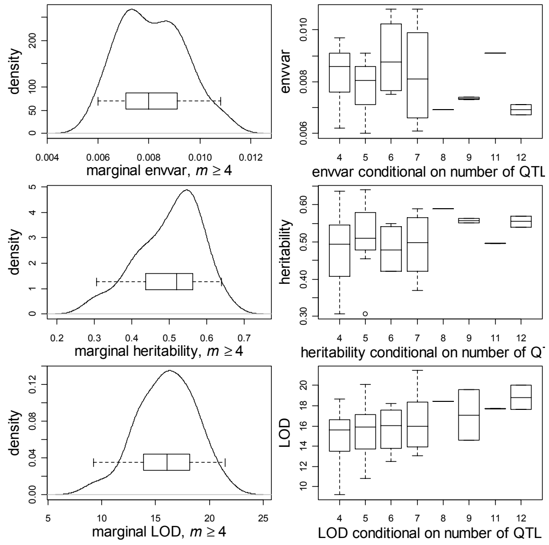
histogram of loci
blue line is density
red lines at estimates



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*