Model Selection for Multiple QTL

- reality of multiple QTL
- selecting a class of QTL models
- comparing QTL models
  - QTL model selection criteria
- assessing performance of model selection
- issues of detecting epistasis
- searching through QTL models: ch 7

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: mimimize prediction error

1 reality of multiple QTL

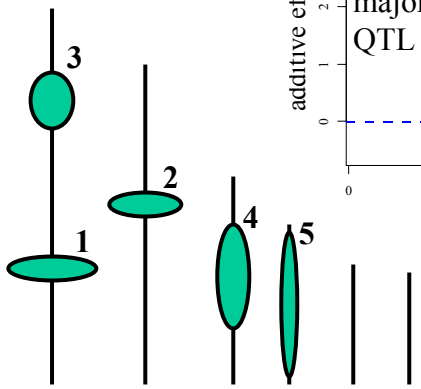
- evaluate objective
  - likelihood or posterior
- search over "space" of genetic architectures
  - number and positions of loci
  - gene action: additive, dominance, epistasis
  - how to efficiently search the model space?
- select "best" or "better" model(s)
  - what criteria to use? where to draw the line?
- estimate "features" of model
  - means, variances & covariances, confidence regions
  - marginal or conditional distributions

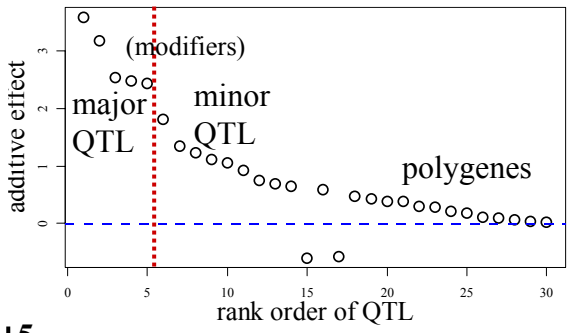
advantages of multiple QTL approach

- 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 = $(\text{bias})^2$ + variance

Pareto diagram of QTL effects

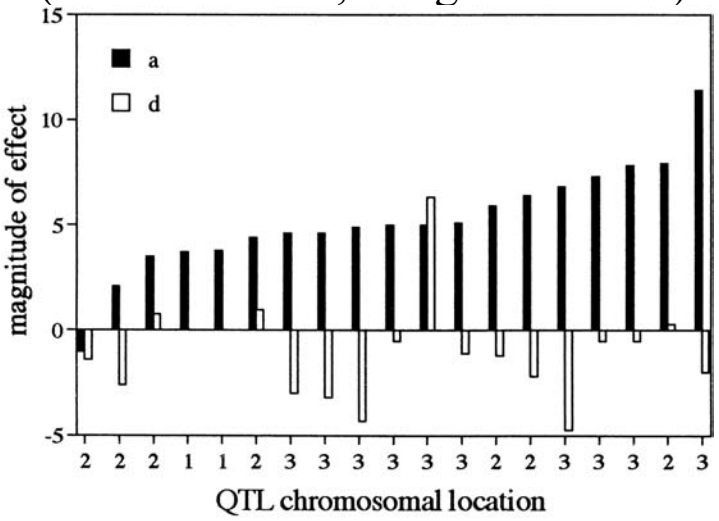
major QTL on linkage map





MIM effects for gonad shape

(Liu et al. 1996; Zeng et al. 2000)



limits of estimation for 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 does not give multiple QTL any advantage
  - hard to select many QTL simultaneously
    - 3^m possible genotypes to choose from
  - sampling & chance variation: only see some patterns
- genetic linkage = multi-collinearity (multiple regression)
  - collinearity leads to correlated estimates of gene effects
  - precision of each effect drops as more predictors are added
- want to balance bias and variance
  - a few QTL can dramatically reduce bias
  - many predictors (QTL) can increase variance
- depends on sample size, heritability, environmental variation

QTL below limits of detection?

- problem of selection bias
  - QTL of modest effect detected sometimes
  - their effects are biased upwards when detected
- how can we avoid sharp in/out dichotomy?
  - caution about only examining the "best" model
  - consider probability that a QTL is in the model
- build m = number of QTL detected into QTL model
  - directly allow uncertainty in genetic architecture
  - model selection over number of QTL, architecture

2 selecting a class of QTL models

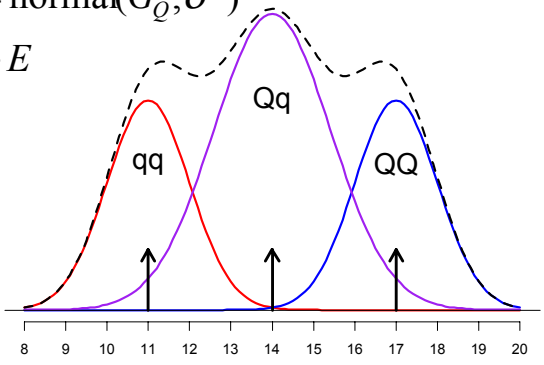
- number of QTL
  - single QTL
  - multiple QTL: known or unknown number
- location of QTL
  - known locations
  - widely spaced (no 2 in marker interval) or arbitrarily close
- gene action
  - additive (A) and/or dominance (D) effects
  - epistatic effects
    - statistical hierarchy (AA, AD, DA, DD)
    - tree-structured contrasts (qqq/qqq vs. other 8 genotypes)
  - phenotype distribution (normal, binomial, Poisson, …)

normal phenotype

- trait = mean + genetic + environment
- pr(trait Y | genotype Q, effects θ)

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

$$Y = \mu + G_Q + E$$



typical assumptions

- normal environmental variation
  - residuals *e* (not *Y*!) have bell-shaped histogram
- genetic value G_Q is composite of *m* QTL
  - $Q = (Q_1, Q_2, \ldots, Q_m)$
- genetic effect uncorrelated with environment

$$Y = \mu + G_Q + e, e \sim N(0, \sigma^2)$$

$$E(Y \mid Q, \theta) = \mu + G_Q, \text{var}(Y \mid Q, \theta) = \sigma^2$$

$$\theta = (\mu, G_Q, \sigma^2) \text{ effects}$$

partitioning multiple QTL

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

- partition of genotypic value (no epistasis)

$$G_Q = \theta_{Q(1)} + \theta_{Q(2)} + \cdots + \theta_{Q(m)} \text{ or } G_Q = \text{sum}_j\, \theta_{Q(j)}$$

- partition of genetic variance

$$\text{var}(G_Q) = \sigma^2_G = \text{sum}_j\, \sigma^2_{G(j)}, \sigma^2_{G(j)} = \text{var}(\theta_{Q(j)})$$

- partition of heritability h^2

$$h^2 = \frac{\sigma^2_G}{\sigma^2_G + \sigma^2} = \text{sum}_j \frac{\sigma^2_{G(j)}}{\sigma^2_G + \sigma^2}$$

model selection with epistasis

- epistasis adds 1-4 model degrees of freedom
  - BC: 1, F2: 4 (AA, AD, DA, DD)
- always include epistasis?
  - BC: add 1 (no epistasis) or *m*+1 (all epistasis) df
- epistasis between significant QTL
  - check all possible pairs
  - include higher order epistasis?
- epistasis with non-significant QTL
  - whole genome paired with significant QTL
  - pairs of non-significant QTL

two QTL with epistasis

- same phenotype model overview

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

- partition of genotypic value with epistasis

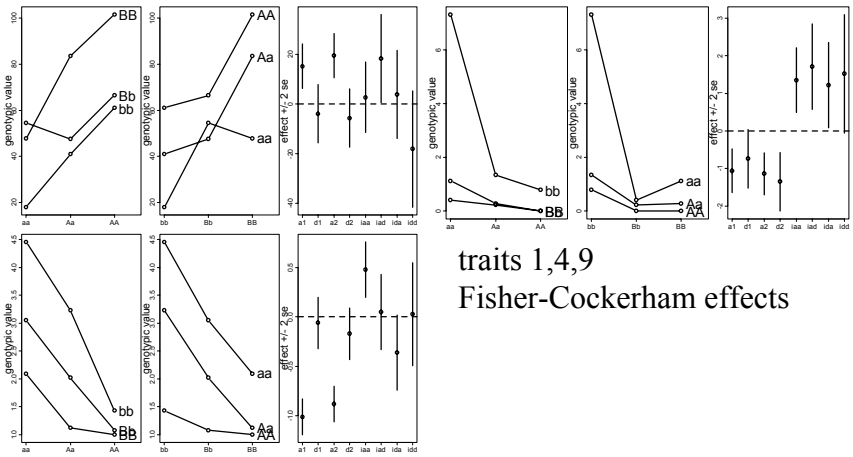
$$G_Q = \theta_{Q(1)} + \theta_{Q(2)} + \theta_{Q(1,2)}$$

- partition of genetic variance

$$\text{var}(G_Q) = \sigma^2_G = \sigma^2_{G(1)} + \sigma^2_{G(2)} + \sigma^2_{G(1,2)}$$

epistasis examples

(Doebley Stec Gustus 1995; Zeng pers. comm.)



traits 1,4,9
Fisher-Cockerham effects

multiple QTL with epistasis

- summation form of linear model
  $G_Q = \text{sum}_j\ \theta_{Q(j)}$
- now include 2-QTL interactions
  $G_Q = \text{sum}_j\ \theta_{1Qj} + \text{sum}_j\ \theta_{2Qj}$
- extra subscript keeps track of order of term
  $\theta_{1Qj} = \theta_{Q(j_1)}, \theta_{2Qj} = \theta_{Q(j_1,j_2)};\ j_1, j_2 = 1, \cdots, m$
- partition of genetic variance
  $\sigma^2_G = \sigma^2_{1G} + \sigma^2_{2G}, \sigma^2_{kG} = \text{sum}_j\ \sigma^2_{kGj}, \sigma^2_{kGj} = \text{var}(\theta_{kQj})$

higher order epistasis

- sum over order and over QTL index

  $G_Q = \text{sum}_k \text{ sum}_j \theta_{kjQ}$

- extra subscript keeps track of order of term

  $\theta_{kjQ} = \theta_{(j_1, j_2, \cdots, j_k)Q}$

- partition of genetic variance

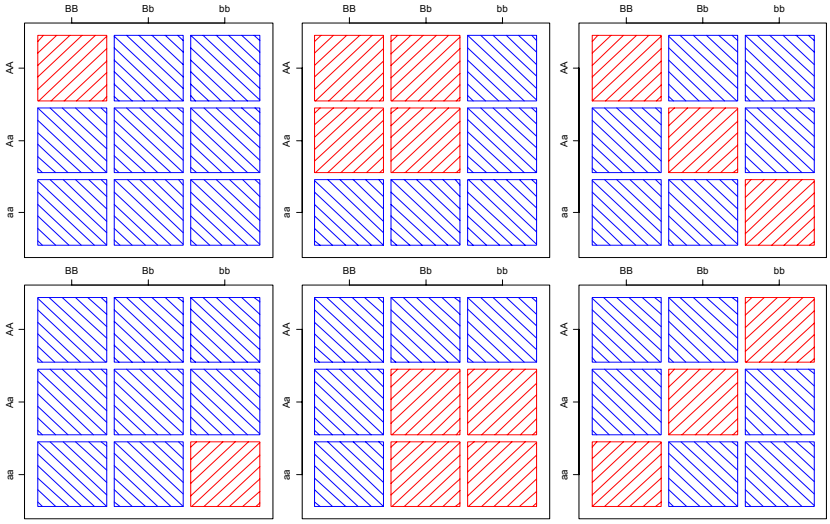
  $\sigma^2_G = \text{sum}_k \sigma^2_{kG}, \sigma^2_{kG} = \text{sum}_j \sigma^2_{kjG}, \sigma^2_{kjG} = \text{var}(\theta_{kjQ})$

- would need large sample size to estimate!

tree-structured phenotype model

- genotypic values divide into groups
  - G_{QQ}, G_{Qq} = high mean phenotype
  - G_{qq} = low mean phenotype
- extend idea to multiple QTL
  - 2 QTL in F2
    - up to 9 groups based on genotype
    - only 4 groups if full dominance
    - only 2 groups if double recessive is distinct
    - other possibilities that do not build on hierarchy

tree-structured epistasis



Bayesian model selection with epistasis

- Yi Xu (2000) *Genetics*
  - all possible pairwise epistasis
- Yi, Xu, Allison (2003) *Genetics*
  - model selection for pairwise epistasis

3 comparing QTL models

- residual sum of squares
- information criteria
  - Bayes information criteria (BIC)
- Bayes factors

residual sum of squares

- residual sum of squares = RSS
  - imagine dense marker map, or only examine markers
  - (deviation of phenotype from genotypic value)^2
  - RSS = sum$_i$ $(Y_i - \mu - G_{Qi})^2$
  - RSS never increases as model grows in size
  - goal: small RSS with "simple" model
- degrees of freedom
  - model degrees of freedom p
    - $p = m$ for backcross with m QTL
    - $p = 2m$ for F2 intercross with m QTL
    - more model df when epistasis allowed
  - error degrees of freedom dfe = $n - p$

model selection = compromise

- mean squared error = MSE
  - MSE = RSS/dfe = $(\text{bias})^2$ + variance
  - bias/variance tradeoff is key issue!
- maximum likelihood with a penalty
  - balance fit (likelihood) with model "complexity"
  - penalize model complexity
    - related to number of parameters, amount of data

recall QTL likelihoods

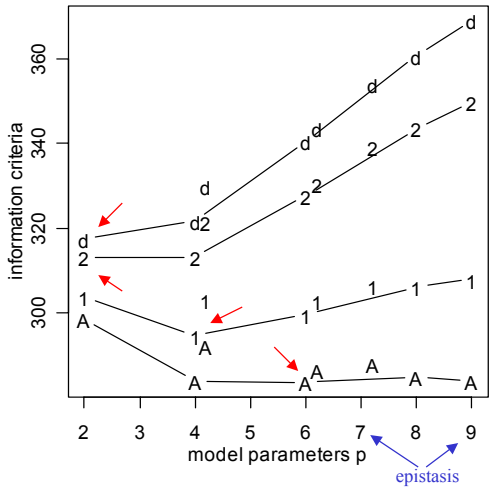
- normal data at a marker
  - likelihood $L(p) = (n/2)\log[\text{RSS}(p)]$
  - LR = ratio of likelihoods for two models
    - p_2 = df for larger model
    - p_1 = df for reduced model
  - $2\log(LR) = L(p_2) - L(p_1) = n \log [\text{RSS}(p_2)/\text{RSS}(p_1)]$
  - $LOD = \log_{10}(LR) = \log(LR)/\log(10)$
- interval mapping
  - mixture across possible genotypes
- non-normal data
  - RSS replaced by deviance

information criteria: likelihoods

- $L(p)$ = likelihood for model with p parameters
- common information criteria:
  - Akaike $\quad AIC = -2\log[L(p)] + 2p$
  - Bayes/Schwartz $\quad BIC = -2\log[L(p)] + p\log(n)$
  - BIC-delta $\quad BIC_\delta = -2\log[L(p)] + \delta p \log(n)$
  - general form: $\quad IC = -2\log[L(p)] + p\,D(n)$
- comparison of models
  - hypothesis testing: designed for one comparison
    - $2\log[LR(p_1,p_2)] = L(p_2) - L(p_1)$
  - model selection: penalize complexity
    - $IC(p_1,p_2) = 2\log[LR(p_1,p_2)] + (p_2 - p_1)\,D(n)$

information criteria vs. model size

- WinQTL 2.0
- SCD data on F2
- A=AIC
- 1=BIC(1)
- 2=BIC(2)
- d=BIC(δ)
- models
  - 1,2,3,4 QTL
    - 2+5+9+2
  - epistasis
    - 2:2 AD



Bayes factors

Which model (1 or 2 or 3 QTLs?) has higher probability of supporting the data?

- ratio of posterior odds to prior odds
- ratio of model likelihoods

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

BF(1:2)	2log(BF)	evidence for 1st
< 1	< 0	negative
1 to 3	0 to 2	negligible
3 to 12	2 to 5	positive
12 to 150	5 to 10	strong
> 150	> 10	very strong

Bayes factors & likelihood ratio

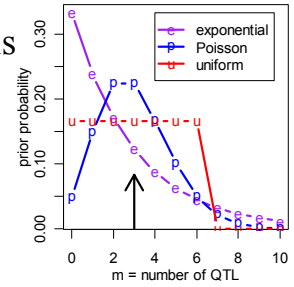
$$B_{12} = \frac{\text{pr}(\text{model}_1 \mid Y) / \text{pr}(\text{model}_2 \mid Y)}{\text{pr}(\text{model}_1) / \text{pr}(\text{model}_2)} = \frac{\text{pr}(Y \mid \text{model}_1)}{\text{pr}(Y \mid \text{model}_2)}$$

- equivalent to *LR* statistic when
  - comparing two nested models
  - simple hypotheses (e.g. 1 vs 2 QTL)
- Bayes Information Criteria (BIC) in general
  - Schwartz introduced for model selection
  - penalty for different number of parameters *p*

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

QTL Bayes factors

- compare models
  - by number of QTL *m*
  - by pattern of QTL across genome
- need prior and posterior for models
  - prior pr(*m*) chosen by user
  - posterior pr(*m*|*Y*,*X*)
    - sampled marginal histogram
    - shape affected by prior pr(*m*)
  - prior for patterns more complicate



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

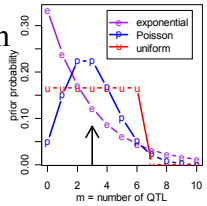
computing marginal means

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

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

computing QTL Bayes factors

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



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

partitioning multiple QTL prior

- partition of genotypic value (no epistasis)

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

- partition of genetic variance

$$G_Q = \theta_{Q(1)} + \theta_{Q(2)} + \cdots + \theta_{Q(m)}$$

- partition of heritability h^2

$$G_Q \sim N(0, \sigma_G^2), \theta_{Q(j)} \sim N(0, \sigma_G^2 / m)$$

multiple QTL phenotype model

- phenotype influenced by genotype & environment
  $\text{pr}(Y|Q,\theta) \sim N(G_Q, \sigma^2)$, or $Y = \mu + G_Q +$ environment
- partition mean into separate QTL effects
  G_Q = main effects + epistatic interactions
  $G_Q = \theta_{1Q} + \ldots + \theta_{mQ} + \ldots$
- priors on mean and effects
  $\mu \sim N(\mu_0, \kappa_0\sigma^2)$ grand mean
  $G_Q \sim N(0, \kappa_1\sigma^2)$ model independent genotypic effect
  $\theta_{jQ} \sim N(0, \kappa_1\sigma^2/m)$ effects down-weighted by m
- determine hyper-parameters via Empirical Bayes

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

phenotype posterior mean

- phenotype influenced by genotype & environment
  $\text{pr}(Y|Q,\theta) \sim N(G_Q, \sigma^2)$, or $Y = \mu + G_Q +$ environment
- relation of posterior mean to LS estimate

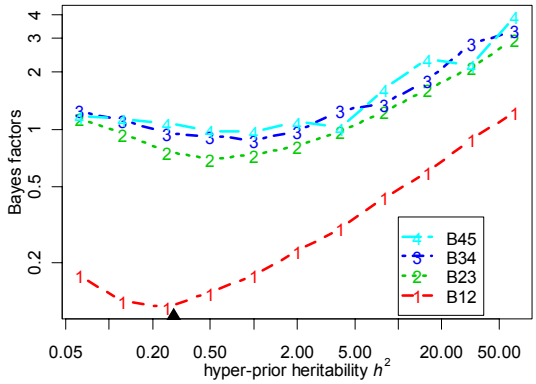
$$G_Q \mid Y, m \sim N(B_Q\hat{G}_Q, B_Q C_Q \sigma^2)$$
$$\approx N(\hat{G}_Q, C_Q\sigma^2)$$

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

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

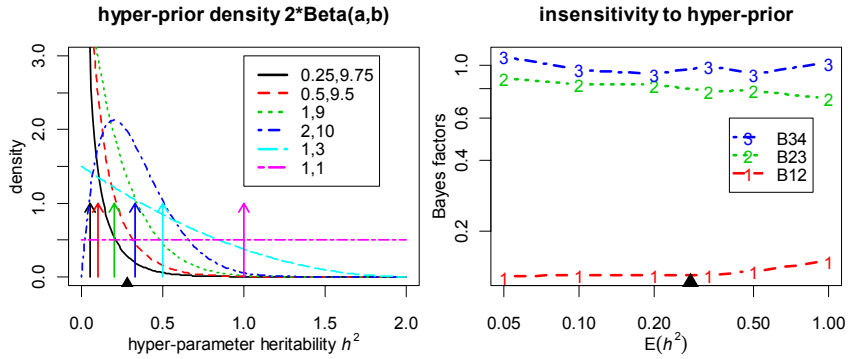
shrinkage $B_Q = \kappa/(\kappa + C_Q) \to 1$

BF sensitivity to fixed prior for effects



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

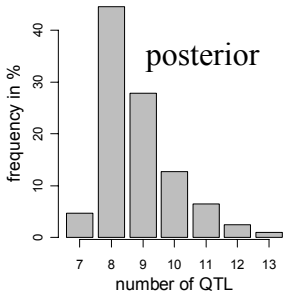
BF insensitivity to random effects prior



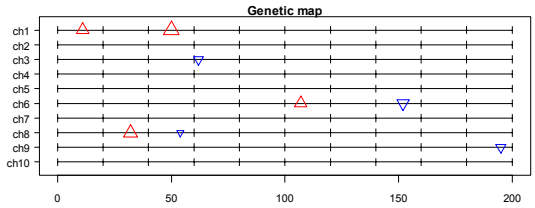
$$\theta_{jQ} \sim \mathrm{N}\left(0, \sigma_G^2 / m\right), \sigma_G^2 = h^2 \sigma_{\text{total}}^2, \tfrac{1}{2} h^2 \sim \mathrm{Beta}(a,b)$$

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

B. napus 8-week vernalization whole genome study

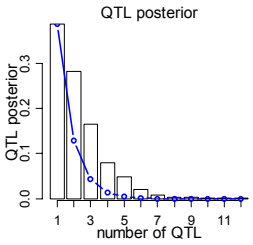
- 108 plants from double haploid
  - similar genetics to backcross: follow 1 gamete
  - parents are Major (biennial) and Stellar (annual)
- 300 markers across genome
  - 19 chromosomes
  - average 6cM between markers
    - median 3.8cM, max 34cM
  - 83% markers genotyped
- phenotype is days to flowering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

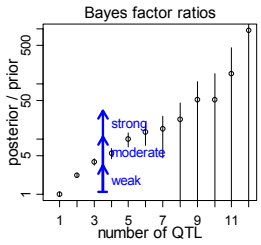
Bayesian model assessment

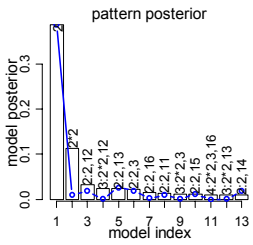
row 1: # QTL
row 2: pattern

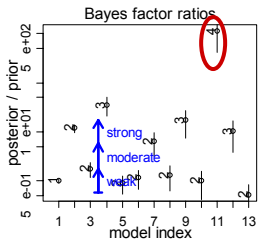
col 1: posterior
col 2: Bayes factor
note error bars on bf

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



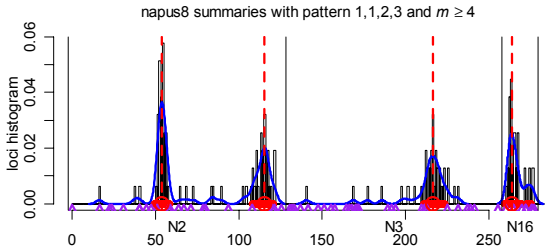




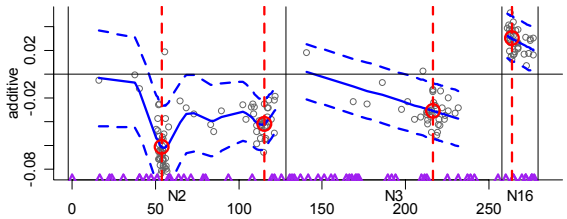


Bayesian estimates of loci & effects

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

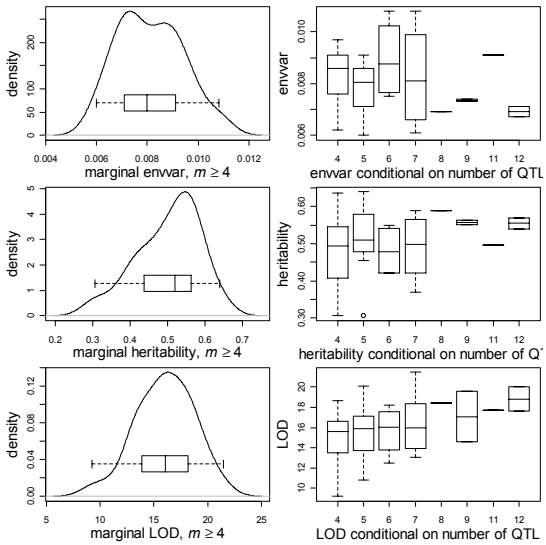


Bayesian model diagnostics

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)

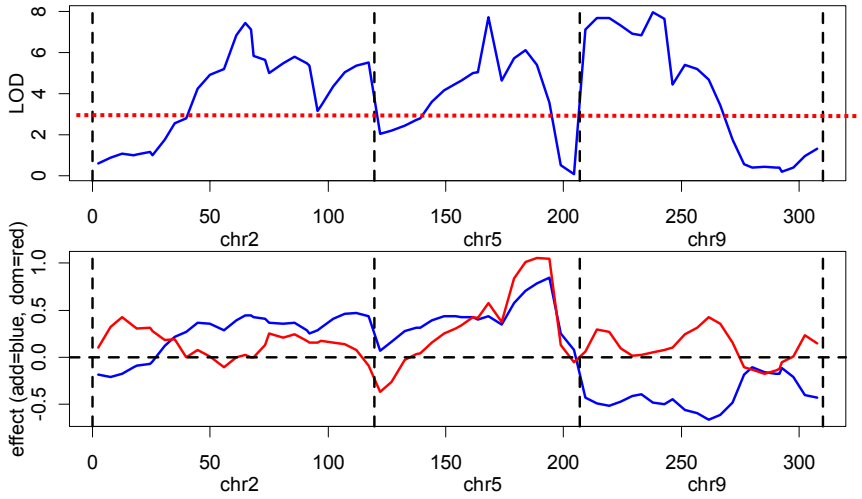
but note change with *m*



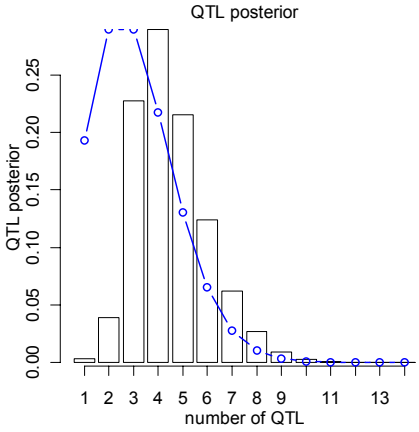
studying diabetes in an F2

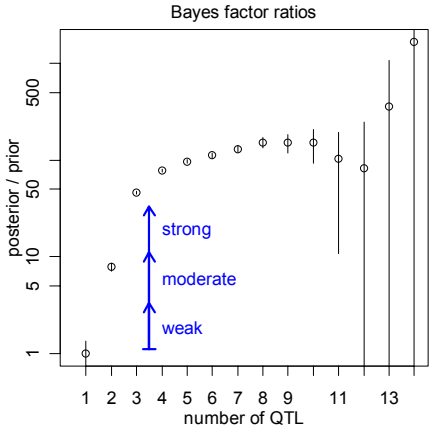
- segregating cross of inbred lines
  - B6.ob x BTBR.ob → F1 → F2
  - selected mice with ob/ob alleles at leptin gene (chr 6)
  - measured and mapped body weight, insulin, glucose at various ages (Stoehr et al. 2000 Diabetes)
  - sacrificed at 14 weeks, tissues preserved
- gene expression data
  - Affymetrix 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,…

Multiple Interval Mapping
SCD1: multiple QTL plus epistasis!

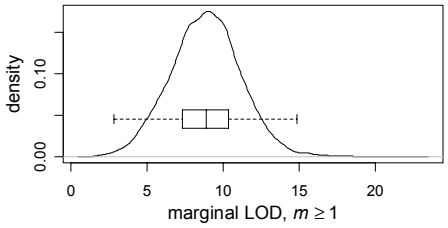


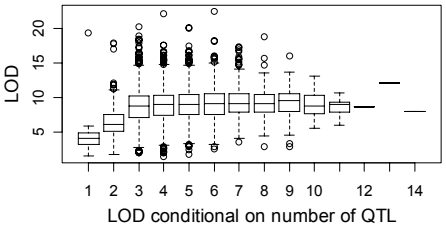
Bayesian model assessment: number of QTL for SCD1

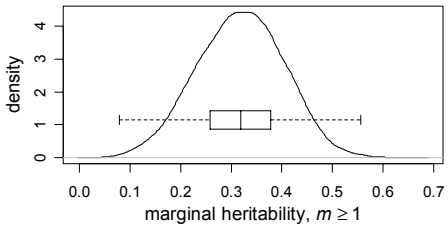


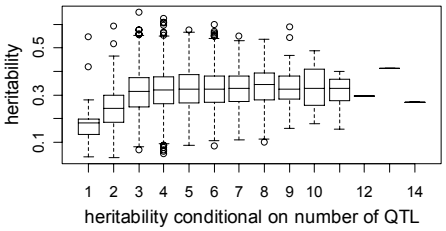


Bayesian LOD and h^2 for SCD1

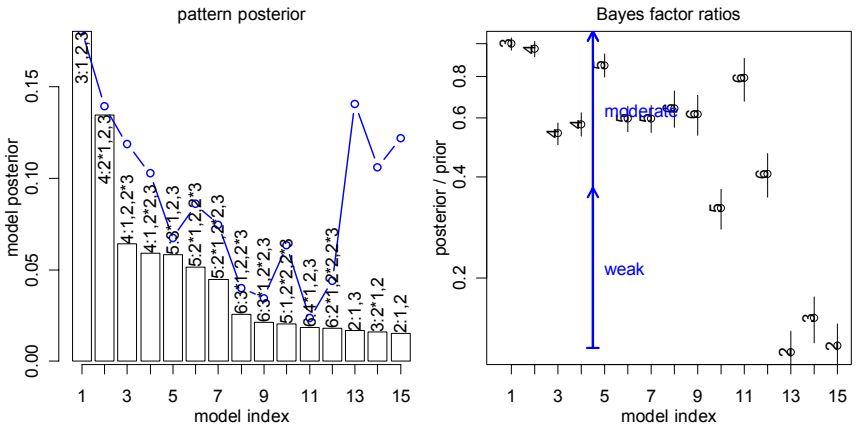






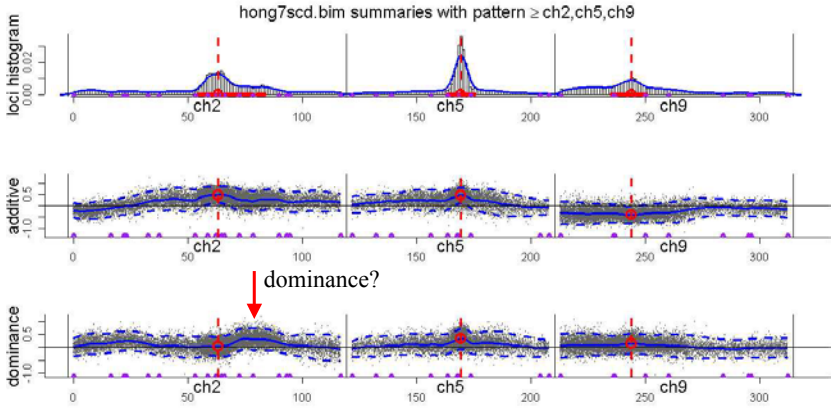


Bayesian model assessment: chromosome QTL pattern for SCD1

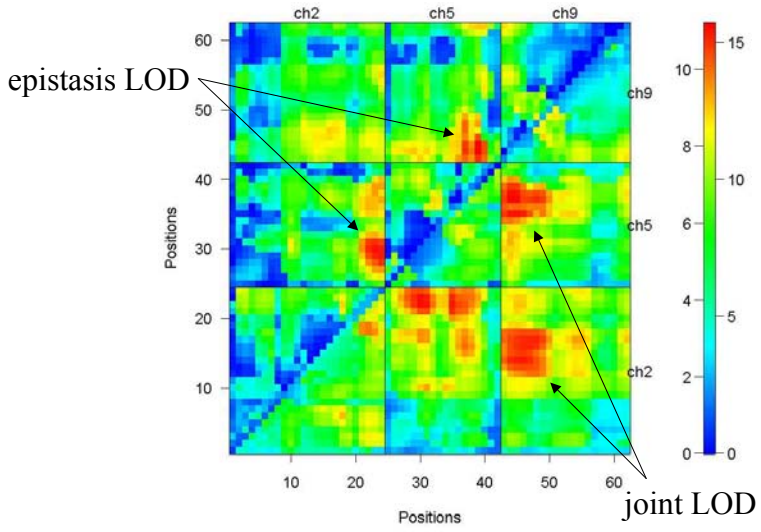


trans-acting QTL for SCD1

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

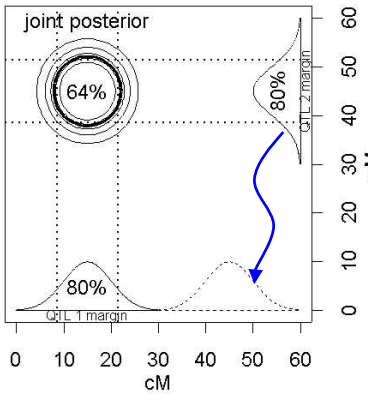


2-D scan: assumes only 2 QTL!

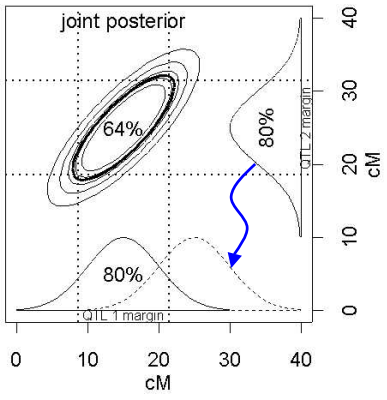


1-D and 2-D marginals
pr(QTL at λ | *Y,X, m*)

unlinked loci



linked loci



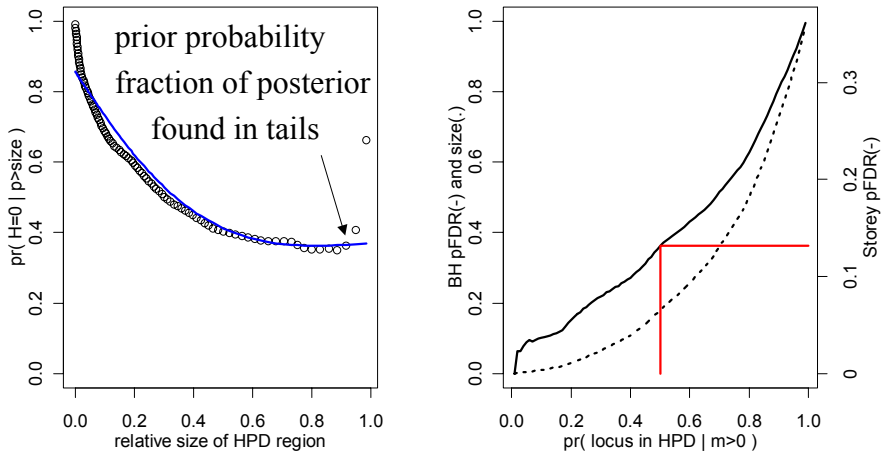
false detection rates and thresholds

- multiple comparisons: test QTL across genome
  - size = pr(LOD(λ) > threshold | no QTL at λ)
  - threshold guards against a single false detection
    - very conservative on genome-wide basis
  - difficult to extend to multiple QTL
- positive false discovery rate (Storey 2001)
  - pFDR = pr(no QTL at λ | LOD(λ) > threshold)
  - Bayesian posterior HPD region based on threshold
    - $\Lambda = \{\lambda \mid \text{LOD}(\lambda) > \text{threshold}\} \approx \{\lambda \mid \text{pr}(\lambda \mid Y,X,m) \text{ large}\}$
  - extends naturally to multiple QTL

pFDR and QTL posterior

- positive false detection rate
  - pFDR = pr(no QTL at λ | $Y,X,\ \lambda$ in Λ)
  - pFDR = $\dfrac{\text{pr}(H=0)*\text{size}}{\text{pr}(m=0)*\text{size}+\text{pr}(m>0)*\text{power}}$
  - power = posterior = pr(QTL in Λ | $Y,X,\ m>0$)
  - size = (length of Λ) / (length of genome)
- extends to other model comparisons
  - m = 1 vs. m = 2 or more QTL
  - pattern = ch1,ch2,ch3 vs. pattern > 2*ch1,ch2,ch3

pFDR for SCD1 analysis



4 assessing performance of model selection procedures

- Broman Speed (2002) article
  - http://www.biostat.jhsph.edu/~kbroman/presentations/rss_ho.pdf
  - focuses on sparse marker map, no missing data
  - marker-based MCMC is different!
    - include/exclude markers in model
- model selection on "continuous" genome
  - infinity of possible predictors
  - uncertainty in position now more important
  - backward elimination requires some care
    - cannot include everything!