6. 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

6.1 reality of multiple QTL

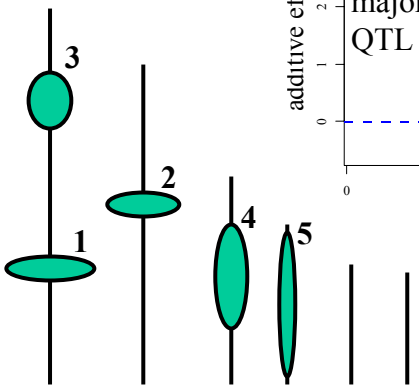
- evaluate objective
  - likelihood or posterior
- search over "space" of genetic architectures
  - number and positions of loci
  - gene action: additive, dominance, epistasis
  - how to efficiently search the model space?
- select "best" or "better" model(s)
  - what criteria to use? where to draw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

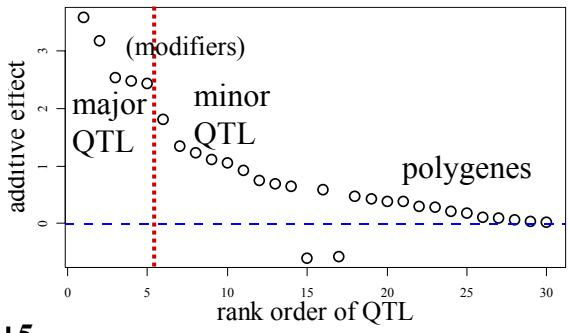
advantages of multiple QTL approach

- 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





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

6.2 selecting a class of QTL models

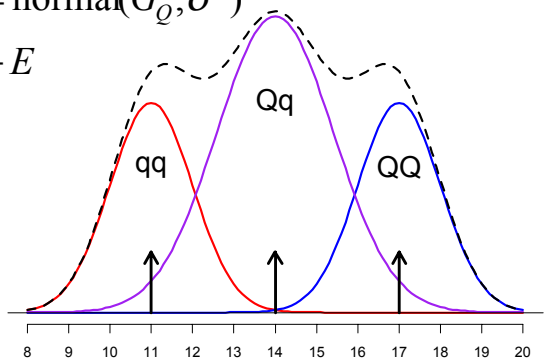
- number of QTL
  - single QTL
  - multiple QTL: known or unknown number
- location of QTL
  - known locations
  - widely spaced (no 2 in marker interval) or arbitrarily close
- gene action
  - additive (A) and/or dominance (D) effects
  - epistatic effects
    - statistical hierarchy (AA, AD, DA, DD)
    - 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

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

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

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

F2 intercross phenotype model

- here assume only one QTL
- genotypes QQ, Qq, qq
- genotypic values G_{QQ}, G_{Qq}, G_{qq}
- decompose as additive, dominance effects

genotype: $Q=$	QQ	Qq	qq
general form	$\mu+\theta_{QQ}$	$\mu+\theta_{Qq}$	$\mu+\theta_{qq}$
Mather-Jinx: $G_Q=$	$\mu+\alpha$	$\mu+\delta$	$\mu-\alpha$
Fisher-Cockerham: $G_Q=$	$\mu+\alpha-\frac{\delta}{2}$	$\mu+\frac{\delta}{2}$	$\mu-\alpha-\frac{\delta}{2}$

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

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

6.3 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 = $n - p$

classical linear models criteria

- mean squared error = MSE
  - MSE = RSS/dfe = $(\text{bias})^2$ + variance
  - bias/variance tradeoff is key issue!
- classical linear models criteria
  - Mallow's $C_p = \text{RSS}(p)/\text{MSE}(\text{full}) - (n - 2p)$
    - balances bias with increased variance
    - sensitive to estimate of MSE for "fullest" model
  - adjusted $R^2 = 1 - (1 - R^2)(n - 1)(n - p)$
    - common practice to adjust for optimistic explained variation
  - both may yield too large a model

resampling approaches

- bootstrap
  - resample with replacement from data
- cross-validation
  - repeatedly divide data into estimation and test sets
- sequential permutation tests
  - condition on QTL already in model
  - stop when added QTL is not significant
- disadvantage for model selection
  - focus on best prediction of phenotype
  - computationally expensive

information criteria: RSS

- maximum likelihood with a penalty
  - penalize model "complexity" (many parameters)
  - balance fit (likelihood) with model complexity
  - normal data: likelihood = $(n/2)\log[\mathrm{RSS}(p)]$
- common information criteria:
  - Akaike AIC $= \log[\mathrm{RSS}(p)] + 2\,p\,/\,n$
  - Bayes/Schwartz BIC $= \log[\mathrm{RSS}(p)] + p\,\log(n)\,/\,n$
  - BIC-delta $\mathrm{BIC}_\delta = \log[\mathrm{RSS}(p)] + \delta p\,\log(n)\,/\,n$
  - Hannon-Quinn HQIC $= \log[\mathrm{RSS}(p)] + p\,\log[\log(n)]\,/\,n$
  - general form: IC $= \log[\mathrm{RSS}(p)] + p\,D(n)\,/n$

rationale/intuition

- minimizing IC ≈ using threshold on LOD
  - LOD $= (n/2)\,\log_{10}\,[\mathrm{RSS}(p_2)/\mathrm{RSS}(p_1)]$
    - p_2 = df for larger model; p_1 = df for reduced model
  - threshold $= (p_1 - p_2)\,D(n)\,/\,2\log(10)$
- Broman-Speed (2002) recommendation
  - pick $D(n)$ = threshold / 2 log(10)
    - $\delta = 2$ threshold / $\log_{10}(n)$
  - threshold ≈ 2.5 for genome-wide 5% level
    - δ = 2.56, 2.10, 1.85 for n = 100, 250, 500
    - $\mathrm{BIC}_\delta \approx \log[\mathrm{RSS}(p)] + 2.56\,p\,\log(n)\,/\,n$ when $n = 100$
    - $\mathrm{BIC}_\delta \approx \log[\mathrm{RSS}(p)] + 11.5\,p\,/\,n$

recall: from RSS to LR to LOD

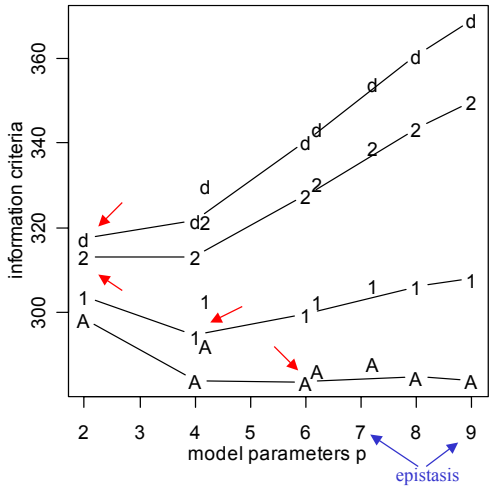
- normal data at a marker
  - RSS(p) = sum$_i$ $(Y_i - \mu - G_{Qi})^2$
    - depends on data and model with p parameters
  - LR = ratio of likelihoods for two models
    - p_2 = df for larger model; p_1 = df for reduced model
  - 2 log(LR) = n log [RSS(p_2)/RSS(p_1)]
  - LOD = $\log_{10}$ (LR) = log(LR)/log(10)
- normal data for interval mapping
  - likelihoods are more complicated
    - mixture of RSS across possible genotypes
    - same relationship of LR to LOD
- non-normal data: RSS replaced by deviance (later)

information criteria: likelihoods

- $L(p)$ = likelihood for model with p parameters
- common information criteria:
  - Akaike AIC = $-2 \log[L(p)] + 2p$
  - Bayes/Schwartz BIC = $-2 \log[L(p)] + p \log(n)$
  - BIC-delta $\text{BIC}_\delta = -2 \log[L(p)] + \delta p \log(n)$
  - Hannon-Quinn HQIC = $-2 \log[L(p)] + p \log[\log(n)]$
  - general form: IC = $-2 \log[L(p)] + p D(n)$
- comparison of models
  - $LR(p_1,p_2) = L(p_2) / L(p_1)$
  - $\text{IC}(p_1,p_2) = 2 \log[LR(p_1,p_2)] + (p_2 - p_1) D(n)$

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

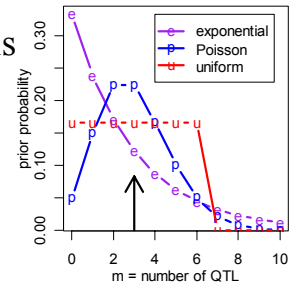
$$B_{12} = \frac{\text{pr}(\text{ model}_1 \mid Y) / \text{pr}(\text{ model}_2 \mid Y)}{\text{pr}(\text{ model}_1) / \text{pr}(\text{ model}_2)} = \frac{\text{pr}(Y \mid \text{model}_1)}{\text{pr}(Y \mid \text{model}_2)}$$

- 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



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

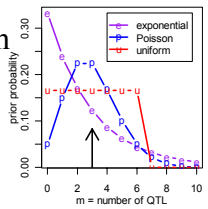
computing marginal means

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

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

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

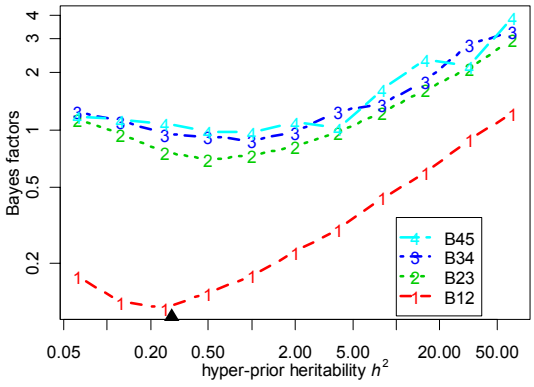
$$\approx N(\hat{G}_Q, C_Q \sigma^2)$$

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

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

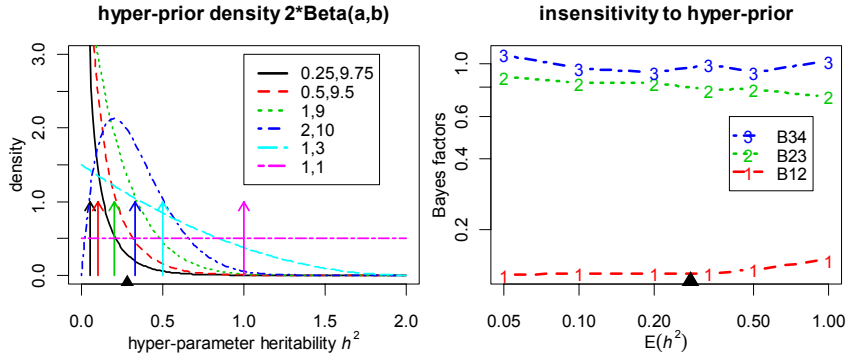
shrinkage $B_Q = \kappa/(\kappa + C_Q) \to 1$

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



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

B. napus 8-week vernalization whole genome study

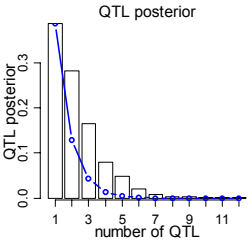
- 108 plants from double haploid
  - similar genetics to backcross: follow 1 gamete
  - parents are Major (biennial) and Stellar (annual)
- 300 markers across genome
  - 19 chromosomes
  - average 6cM between markers
    - median 3.8cM, max 34cM
  - 83% markers genotyped
- phenotype is days to flowering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

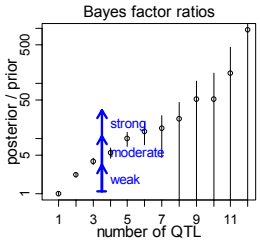
Bayesian model assessment

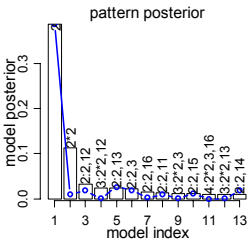
row 1: # QTL
row 2: pattern

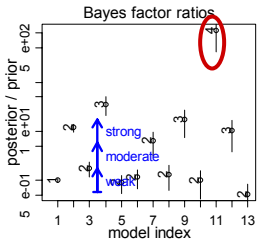
col 1: posterior
col 2: Bayes factor
note error bars on bf

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



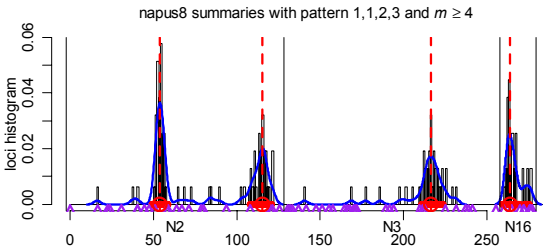




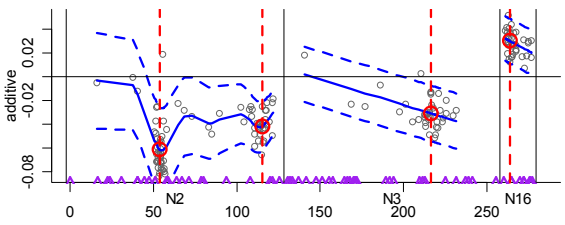


Bayesian estimates of loci & effects

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

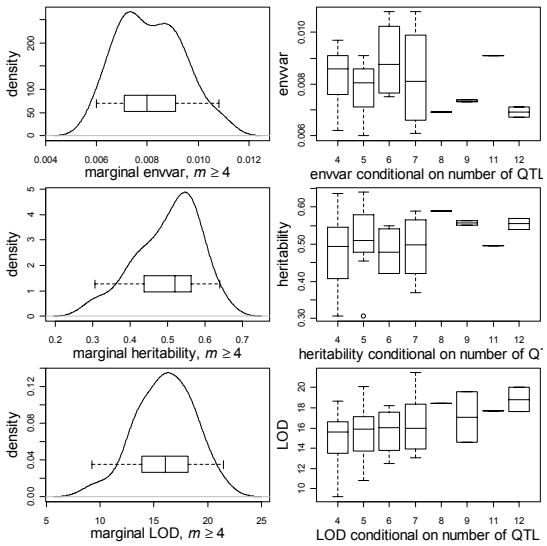


Bayesian model diagnostics

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*



6.4 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!

6.5 issues of detecting epistasis

- two QTL with epistasis
- partition into additive and dominance
  - Fisher-Cockerham model
- multiple QTL and higher order epistasis
- tree-structured phenotype models

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

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

Fisher-Cockerham interaction

- two QTL
- here show 3 of 9 genotypes (vary QTL 1)
- four interactions: aa, ad, da, aa

$$\mathrm{Q_1Q_1Q_2Q_2} \quad \mu+\alpha_1-\frac{\delta_1}{2}+\alpha_2-\frac{\delta_2}{2}+i_{aa}-\frac{i_{ad}}{2}-\frac{i_{da}}{2}+\frac{i_{dd}}{4}$$

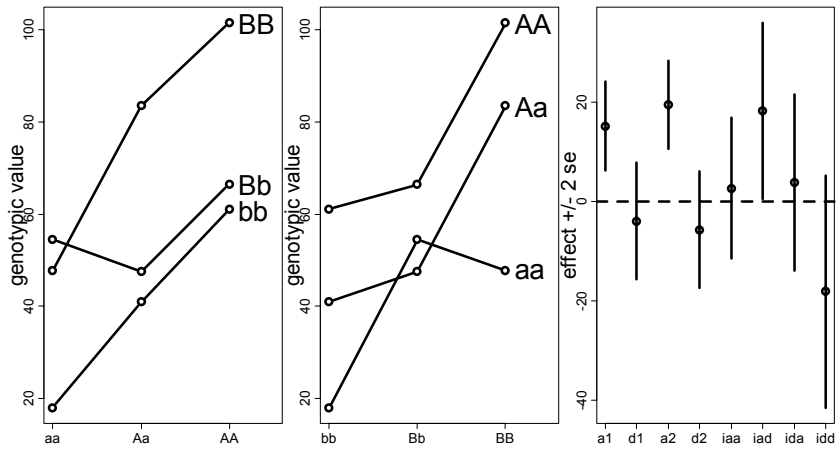
$$\mathrm{Q_1q_1Q_2Q_2} \quad \mu+\frac{\delta_1}{2}+\alpha_2-\frac{\delta_2}{2}-\frac{i_{dd}}{4}$$

$$\mathrm{q_1q_1Q_2Q_2} \quad \mu-\alpha_1-\frac{\delta_1}{2}+\alpha_2-\frac{\delta_2}{2}-i_{aa}+\frac{i_{ad}}{2}-\frac{i_{da}}{2}+\frac{i_{dd}}{4}$$

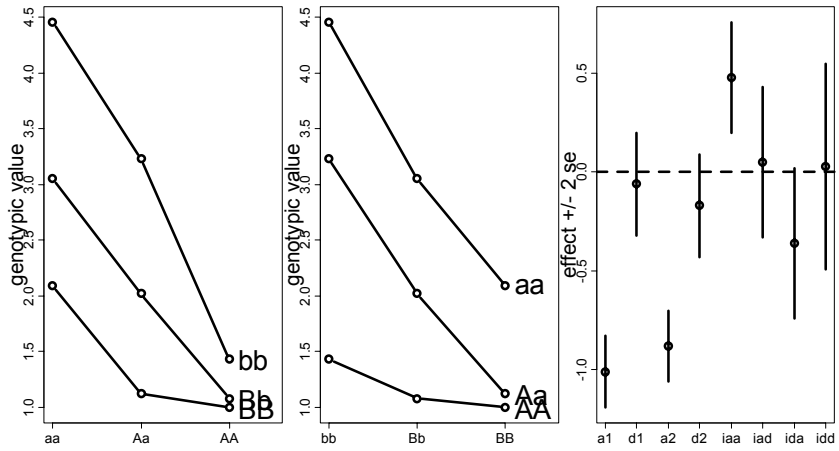
genotype effect coefficients

genotype	α_1	δ_1	α_2	δ_2	i_{aa}	i_{ad}	i_{da}	i_{dd}
QQQQ	$+1$	$-\frac{1}{2}$	$+1$	$-\frac{1}{2}$	$+1$	$-\frac{1}{2}$	$-\frac{1}{2}$	$+\frac{1}{4}$
QqQQ	0	$\frac{1}{2}$	$+1$	$-\frac{1}{2}$	0	0	$+\frac{1}{2}$	$-\frac{1}{4}$
qqQQ	-1	$-\frac{1}{2}$	$+1$	$-\frac{1}{2}$	-1	$+\frac{1}{2}$	$-\frac{1}{2}$	$+\frac{1}{4}$
QQQq	$+1$	$-\frac{1}{2}$	0	$+\frac{1}{2}$	0	$+\frac{1}{2}$	0	$-\frac{1}{4}$
QqQq	0	$+\frac{1}{2}$	0	$+\frac{1}{2}$	0	0	0	$+\frac{1}{4}$
qqQq	-1	$-\frac{1}{2}$	0	$+\frac{1}{2}$	0	$-\frac{1}{2}$	0	$-\frac{1}{4}$
QQqq	$+1$	$-\frac{1}{2}$	-1	$-\frac{1}{2}$	-1	$-\frac{1}{2}$	$+\frac{1}{2}$	$-\frac{1}{4}$
Qqqq	0	$\frac{1}{2}$	-1	$-\frac{1}{2}$	0	0	$-\frac{1}{2}$	$+\frac{1}{4}$
qqqq	-1	$-\frac{1}{2}$	-1	$-\frac{1}{2}$	$+1$	$+\frac{1}{2}$	$+\frac{1}{2}$	$+\frac{1}{4}$

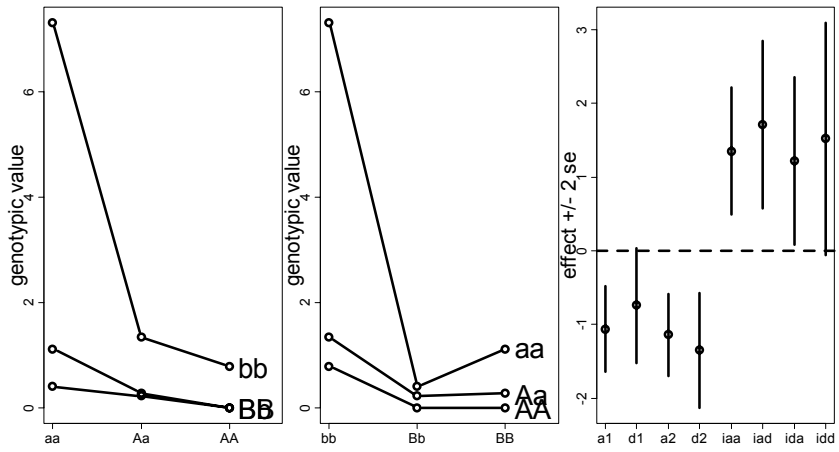
epistasis (Doebley Stec Gustus 1995) trait 1



epistasis (Doebley Stec Gustus 1995) trait 4



epistasis (Doebley Stec Gustus 1995) trait 9



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

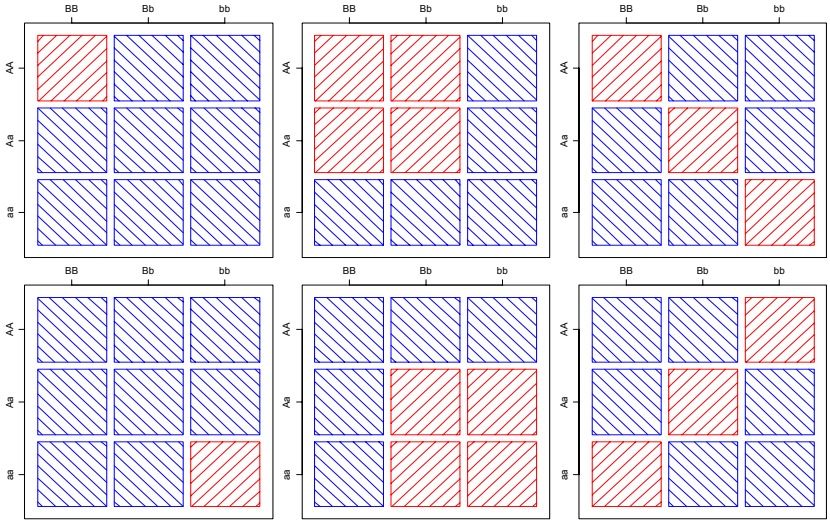
  $\sigma^2_G = \text{sum}_k \sigma^2_{kG}, \sigma^2_{kG} = \text{sum}_j \sigma^2_{kjG}, \sigma^2_{kjG} = \text{var}(\theta_{kjQ})$

- 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



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