Model Selection for Multiple QTL

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

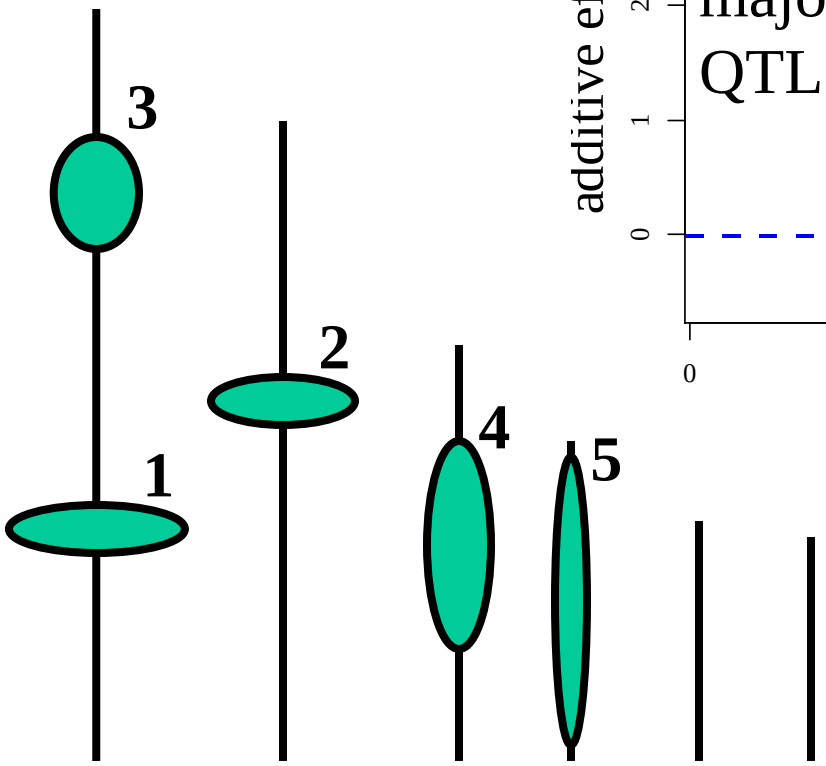
- evaluate some objective for model given data
  - classical likelihood
  - Bayesian posterior
- search over possible genetic architectures (models)
  - number and positions of loci
  - gene action: additive, dominance, epistasis
- estimate “features” of model
  - means, variances & covariances, confidence 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?

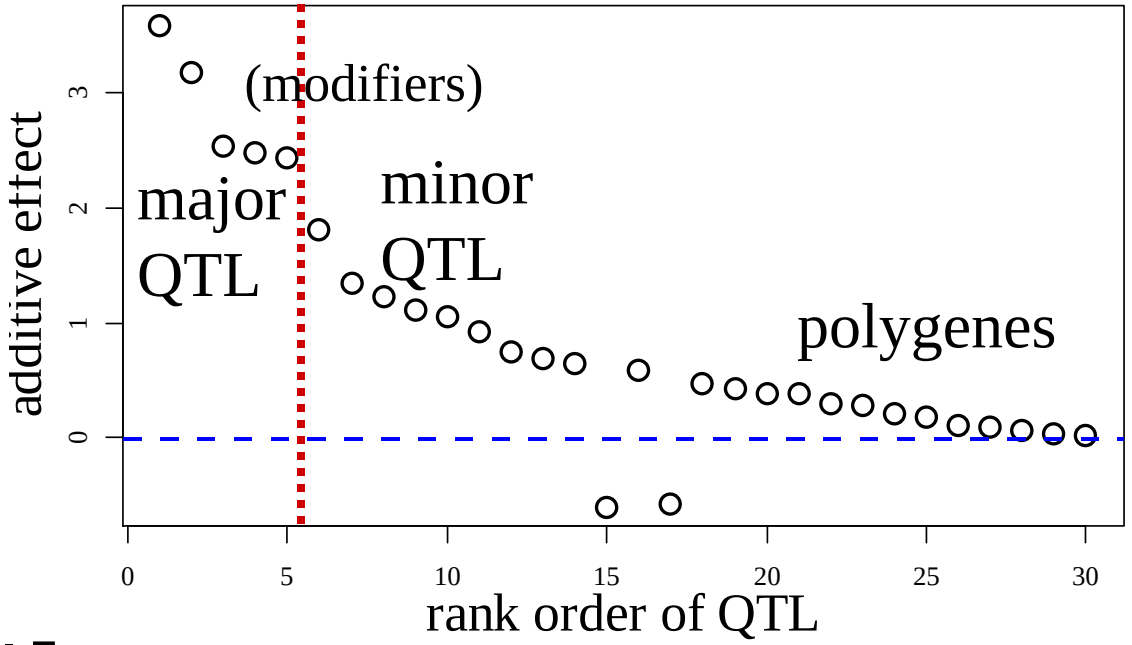
advantages of multiple QTL approach

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

Pareto diagram of QTL effects

major QTL on linkage map





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

2. selecting a class of QTL models

- phenotype distribution
  - normal (usual), binomial, Poisson, …
  - exponential family, semi-parametric, nonparametric
- θ = gene action
  - additive (A) or general (A+D) effects
  - epistatic interactions (AA, AD, ..., or other types?)
- λ = location of QTL
  - known locations?
  - widely spaced (no 2 in marker interval) or arbitrarily close?
- m = number of QTL
  - single QTL?
  - multiple QTL: known or unknown number?

normal phenotype

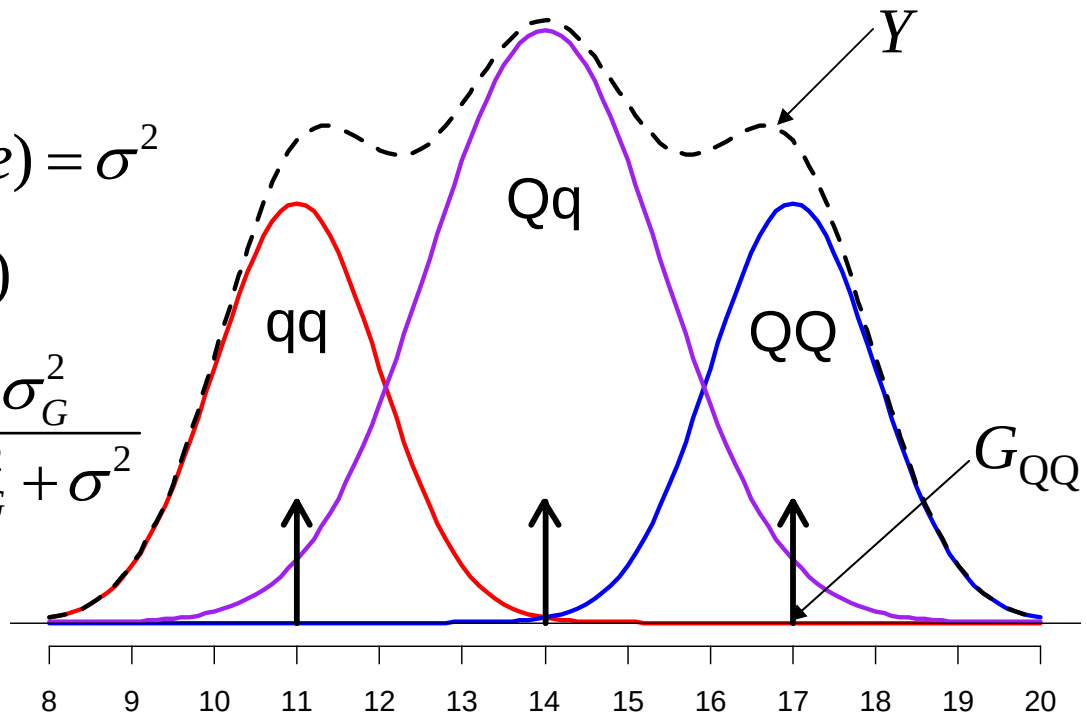
- trait = mean + genetic + environment
- genetic effect uncorrelated with environment
- pr(trait Y | genotype Q, effects θ)

$$Y = G_Q + e$$

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

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

$$\text{heritability } h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma^2}$$



two QTL with epistasis

- same phenotype model overview

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

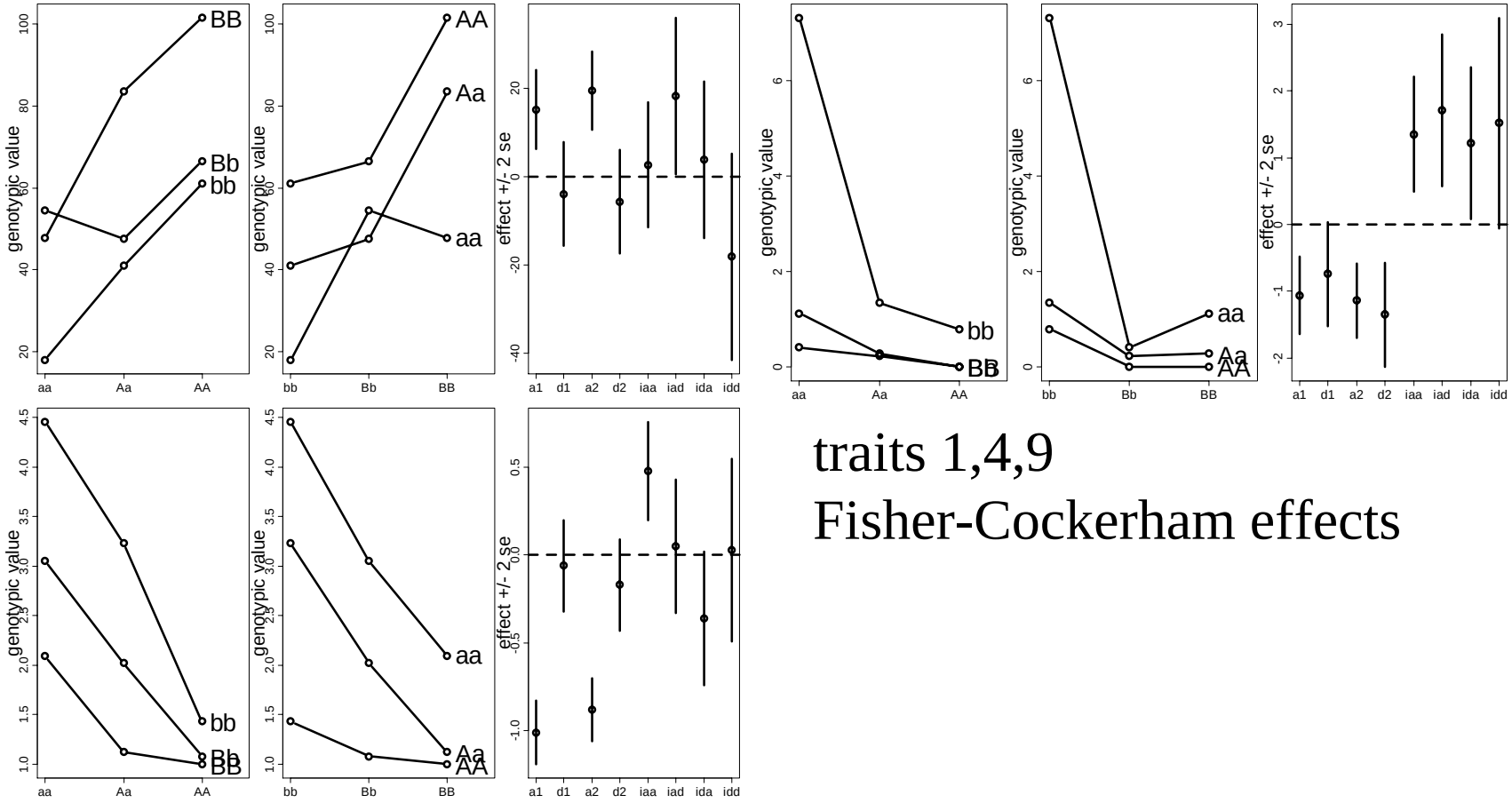
- partition of genotypic value with epistasis

$$G_Q = \mu + \beta_1(Q) + \beta_2(Q) + \beta_{12}(Q)$$

- partition of genetic variance

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

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



traits 1,4,9
Fisher-Cockerham effects

multiple QTL with epistasis

- same overview model

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

- sum over multiple QTL in model $M = \{1,2,12,\ldots\}$

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

- partition genetic variance in same manner

$$\text{var}(G_Q) = \sigma_G^2 = \text{sum}_{\{j \in M\}} \sigma_j^2$$

- could restrict attention to 2-QTL interactions

model selection with epistasis

- additive by additive 2-QTL interaction
  - adds only 1 model degree of freedom (df) per pair
  - but could miss important kinds of interaction
- full epistasis adds many model df
  - 2 QTL in BC: 1 df (one interaction)
  - 2 QTL in F2: 4 df (AA, AD, DA, DD)
  - 3 QTL in F2: 20 df (3×4 d.f. 2-QTL, 8 d.f. 3-QTL)
- data-driven interactions (tree-structured)
  - contrasts comparing subsets of genotypes
  - double recessive or double dominant vs other genotypes
  - discriminant analysis based contrasts (Gilbert and Le Roy 2003, 2004)

- some issues in model search
  - 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 (2004)

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

QTL likelihoods and parameters

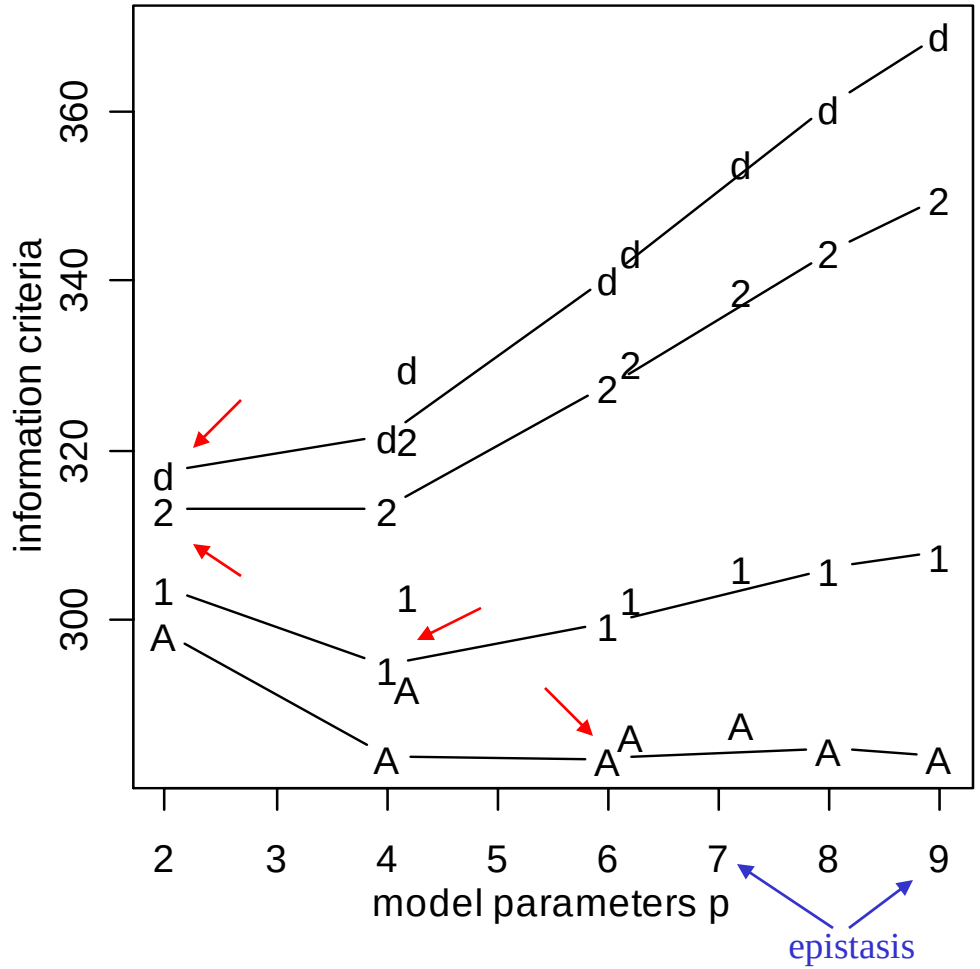
- LOD or likelihood ratio compares model
  - $L(p)$ = log likelihood for a particular model with p parameters
  - $\log(LR) = L(p_2) - L(p_1)$
  - $LOD = \log_{10}(LR) = \log(LR)/\log(10)$
- p = number of model degrees of freedom
  - consider models with m QTL and all 2-QTL epistasis terms
  - BC: $p = 1 + m + m(m\text{-}1)$
  - F2: $p = 1 + 2m + 4m(m\text{-}1)$
- Bayesian information criterion balances complexity
  - $\mathrm{BIC}(\delta) = -2 \log[L(p)] + \delta\, p \log(n)$
  - n = number of individuals in study
  - δ = Broman's BIC adjustment

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 $\text{BIC}_\delta = -2 \log[L(p)] + \delta p \log(n)$
  - general form: 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
    - $\text{IC}(p_1,p_2) = 2 \log[LR(p_1,p_2)] + (p_2 - p_1) D(n)$

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 & BIC

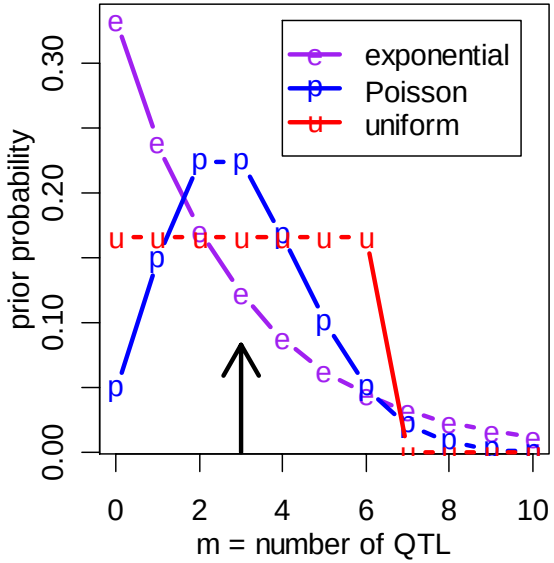
$$B_{12} = \frac{\text{pr}(\text{ model}_1 \mid Y) / \text{pr}(\text{ model}_2 \mid Y)}{\text{pr}(\text{ model}_1) / \text{pr}(\text{ model}_2)} = \frac{\text{pr}(Y \mid \text{model}_1)}{\text{pr}(Y \mid \text{model}_2)}$$

- 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)

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

QTL Bayes factors

- m = number of QTL
  - prior pr(m) chosen by user
  - posterior pr($m|Y,X$)
    - sampled marginal histogram
    - shape affected by prior pr(m)
- pattern of QTL across genome
  - more complicated prior
  - posterior easily sampled



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

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 pr($m|Y,X$) is marginal histogram

multiple QTL priors

- phenotype influenced by genotype & environment

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

- partition genotype-specific mean into QTL effects

  G_Q = mean + main effects + epistatic interactions

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

- priors on mean and effects

μ	$\sim$	$N(\mu_0, \kappa_0\sigma^2)$	grand mean
$\beta(Q)$	$\sim$	$N(0, \kappa_1\sigma^2)$	model-independent genotypic effect
$\beta_j(Q)$	$\sim$	$N(0, \kappa_1\sigma^2/\lvert M\rvert)$	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
  pr($Y|Q,\theta$) ~ $N(G_Q, \sigma^2)$, or $Y = \mu + G_Q$ + environment
- relation of posterior mean to LS estimate

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

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

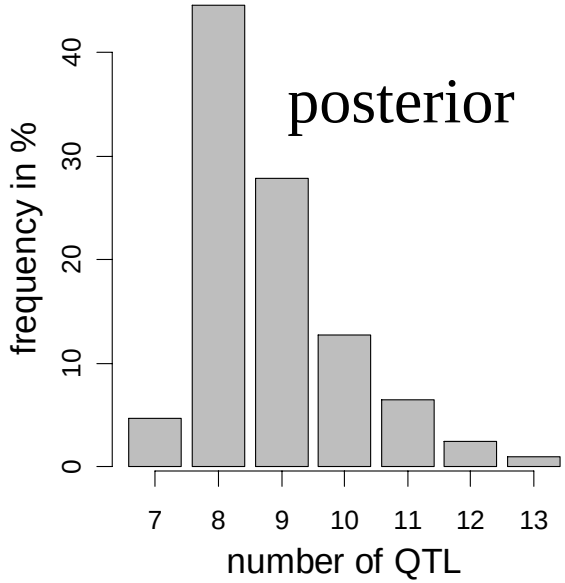
$$\text{LS estimate } \hat{G}_Q = \text{sum}_i[\text{sum}_{j \in M} \hat{\beta}_j(Q_i)] = \text{sum}_i w_{iQ} Y$$

$$\text{variance} \quad V(\hat{G}_Q) = \text{sum}_i w_{iQ}^2 \sigma^2 = C_Q \sigma^2$$

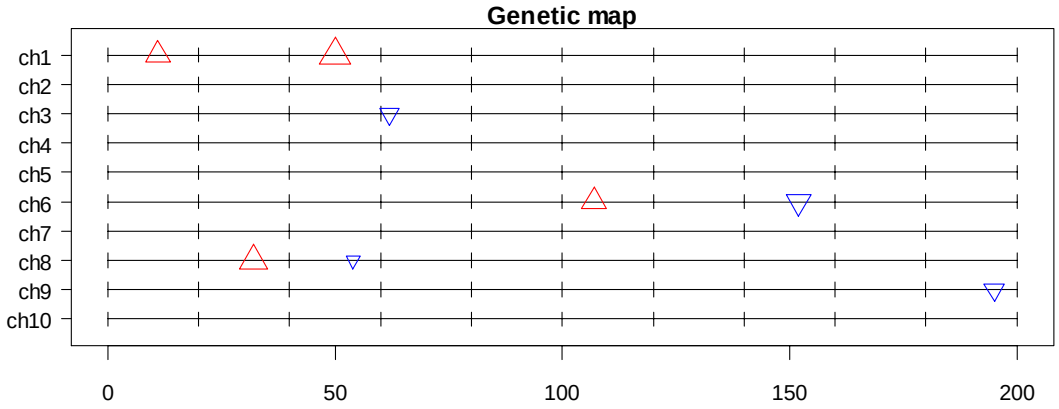
$$\text{shrinkage} \quad B_Q = \kappa/(\kappa + C_Q) \to 1$$

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

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

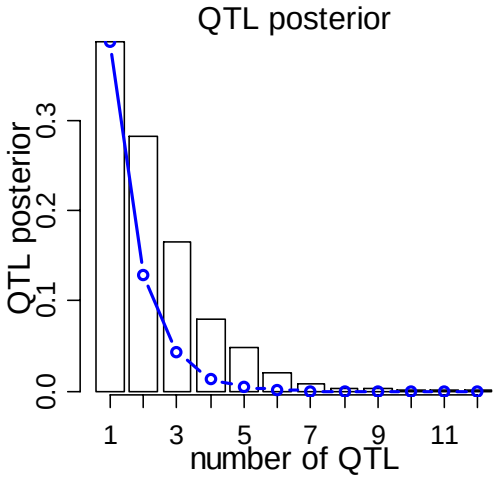
- 108 plants from double haploid
  - similar genetics to backcross: follow 1 gamete
  - parents are Major (biennial) and Stellar (annual)
- 300 markers across genome
  - 19 chromosomes
  - average 6cM between markers
    - median 3.8cM, max 34cM
  - 83% markers genotyped
- phenotype is days to flowering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

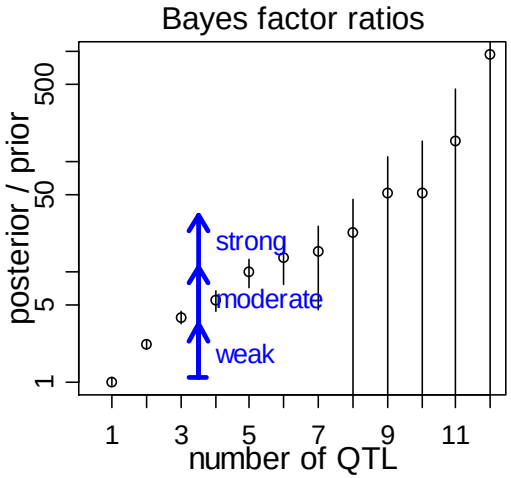
Bayesian model assessment

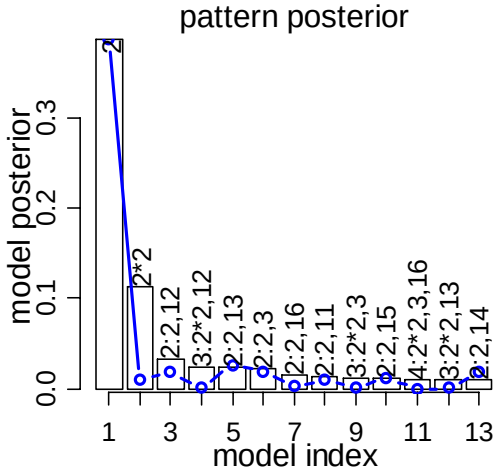
row 1: # QTL
row 2: pattern

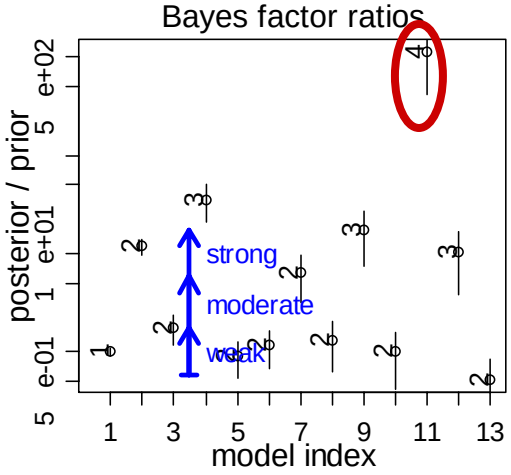
col 1: posterior
col 2: Bayes factor
note error bars on bf

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



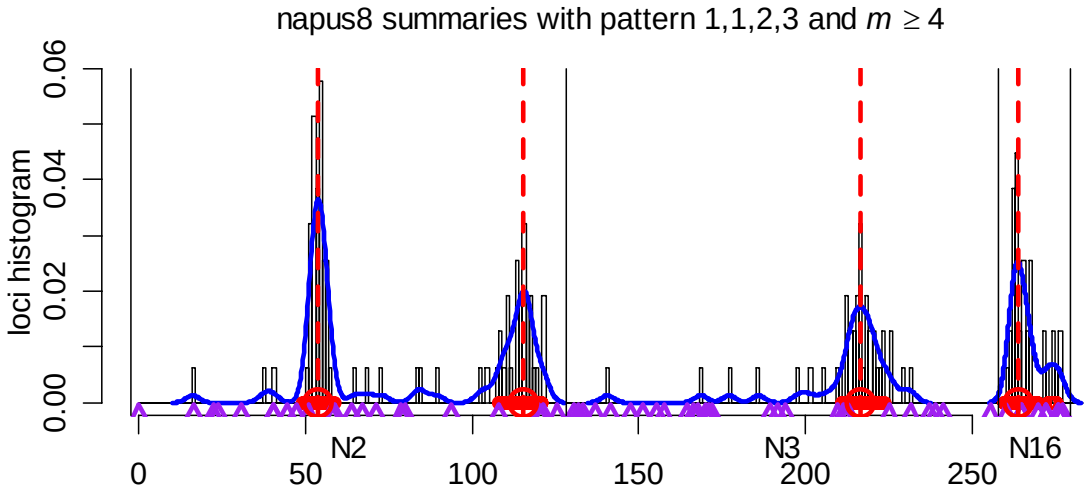




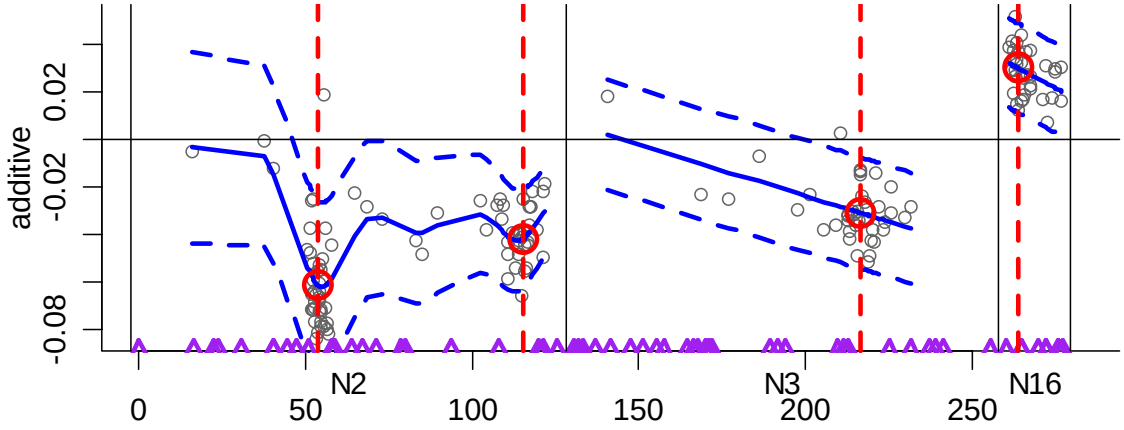


Bayesian estimates of loci & effects

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

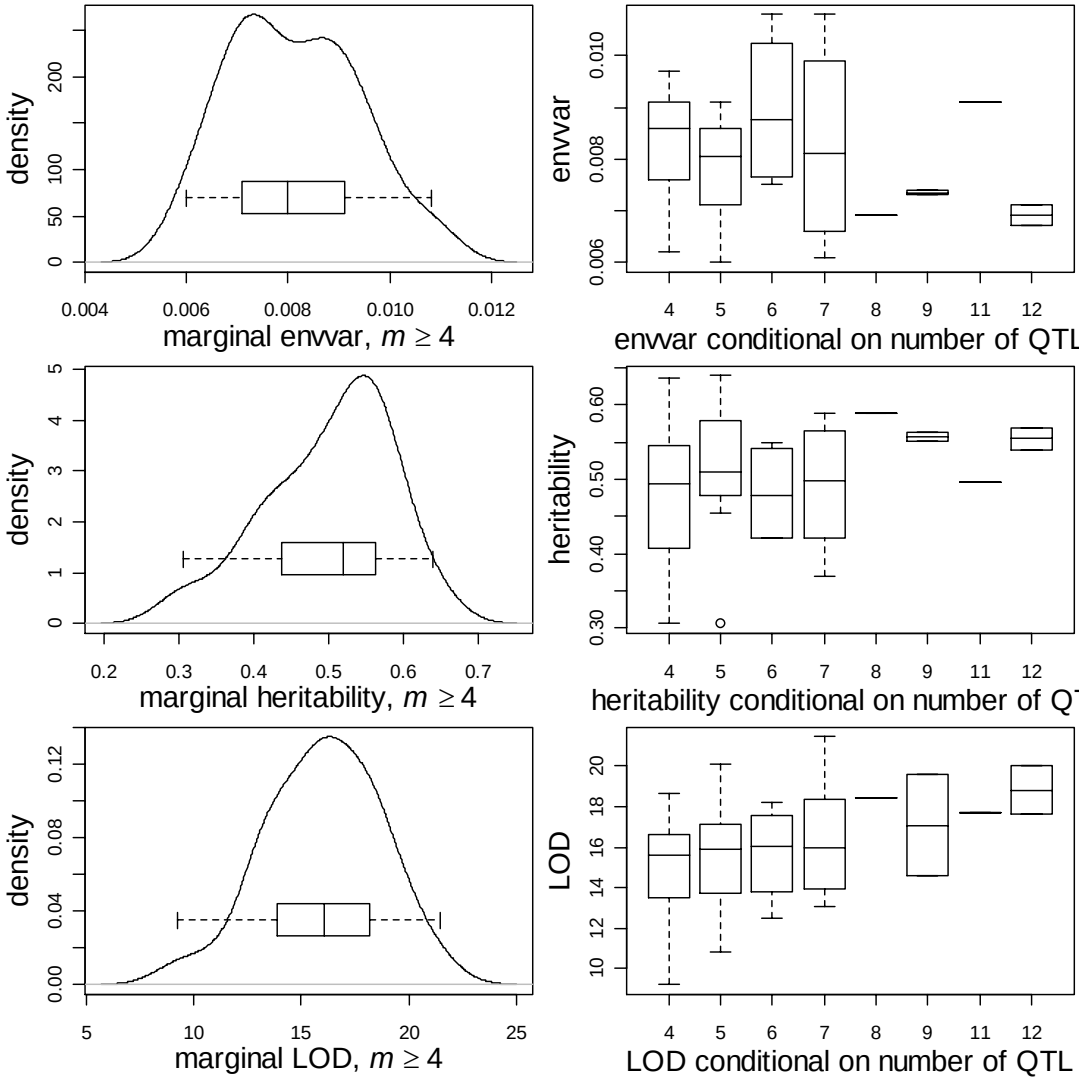


Bayesian model diagnostics

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)

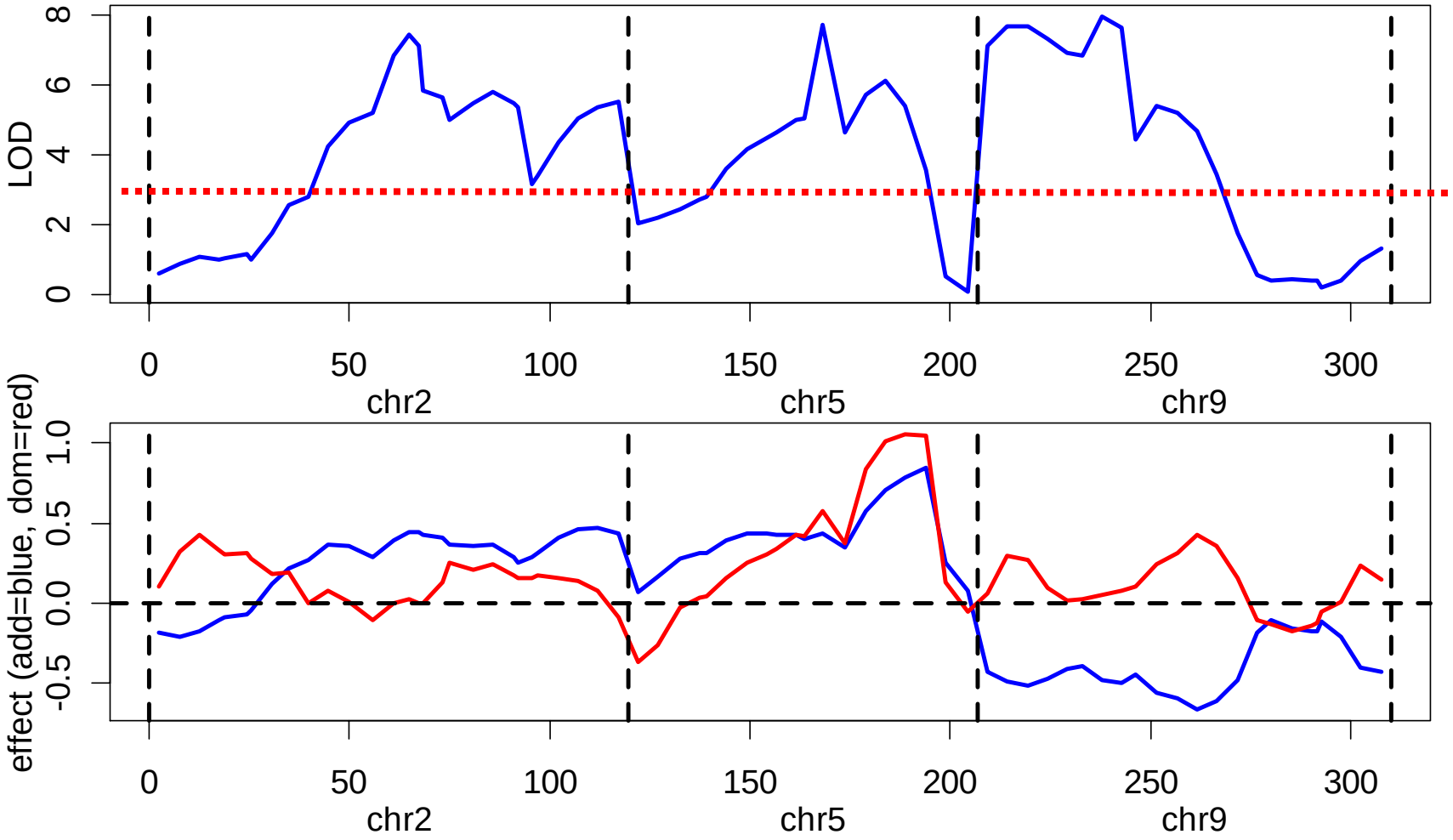
but note change with *m*



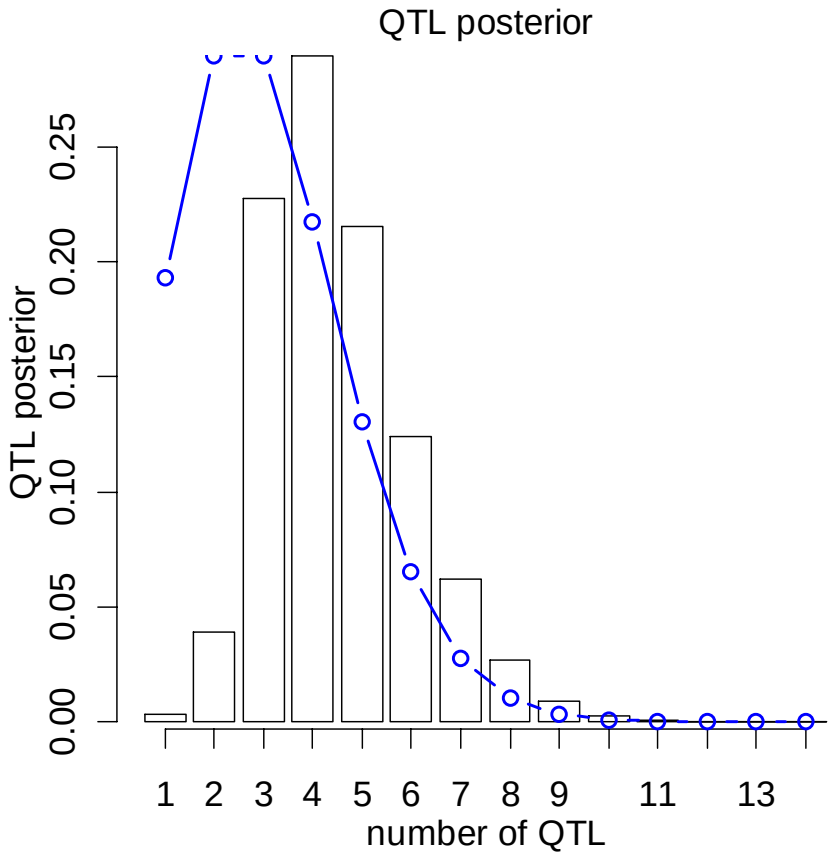
studying diabetes in an F2

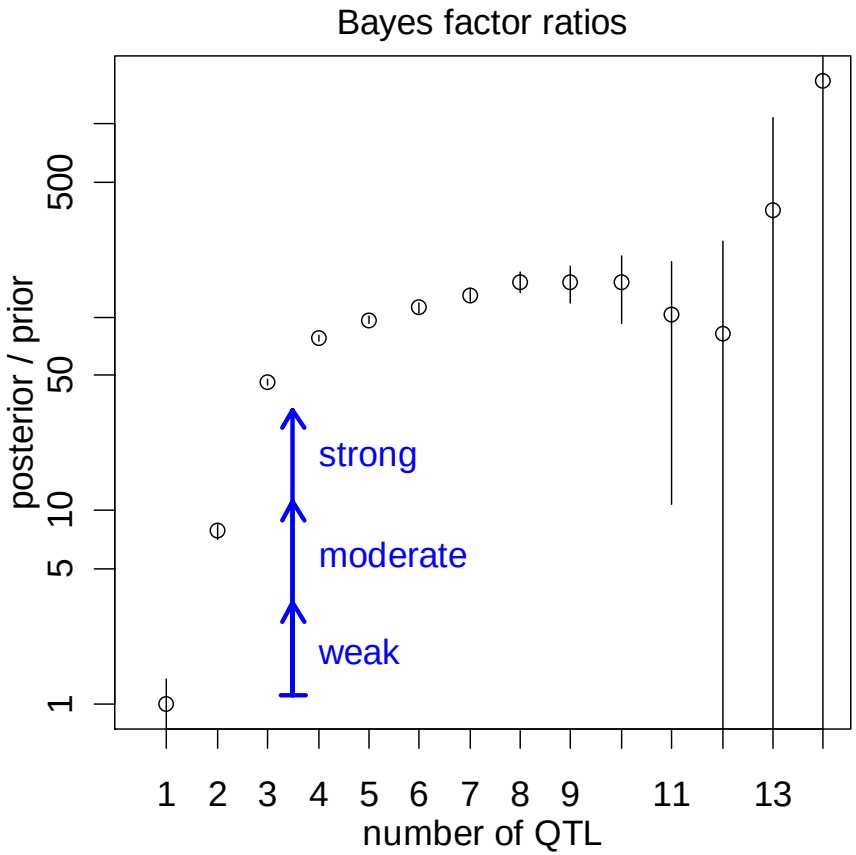
- segregating cross of inbred lines
  - B6.ob x BTBR.ob $\rightarrow$ F1 $\rightarrow$ F2
  - selected mice with ob/ob alleles at leptin gene (chr 6)
  - measured and mapped body weight, insulin, glucose at various ages (Stoehr et al. 2000 Diabetes)
  - sacrificed at 14 weeks, tissues preserved
- gene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

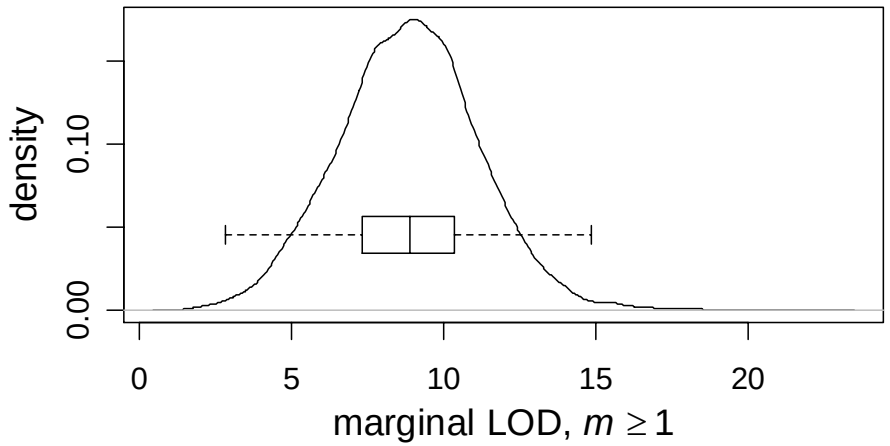


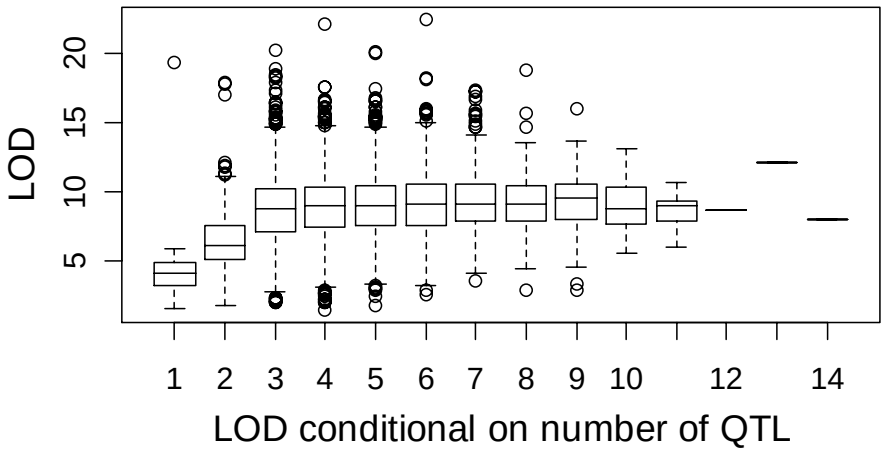
Bayesian model assessment: number of QTL for SCD1

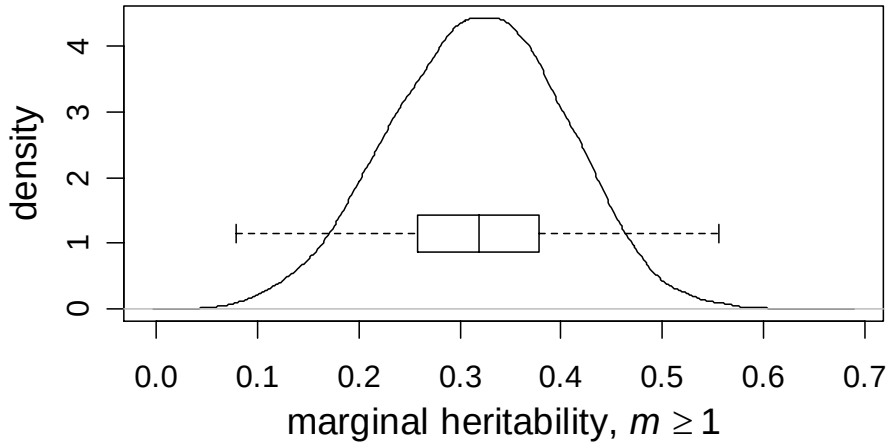


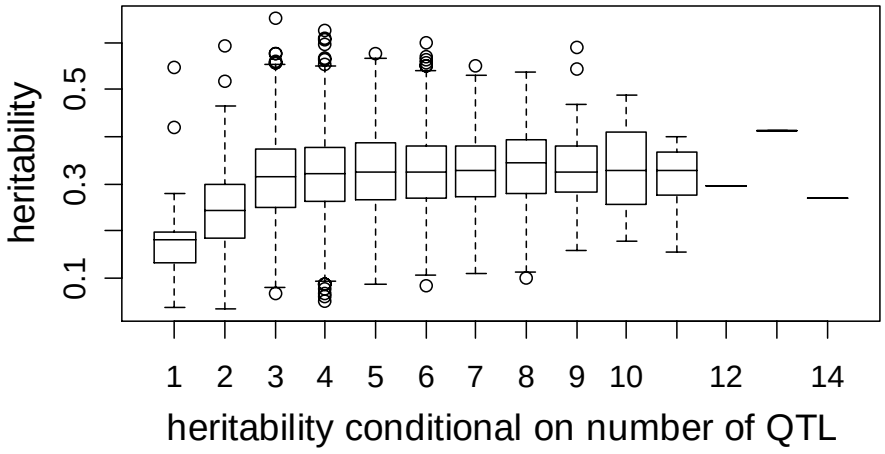


Bayesian LOD and h^2 for SCD1

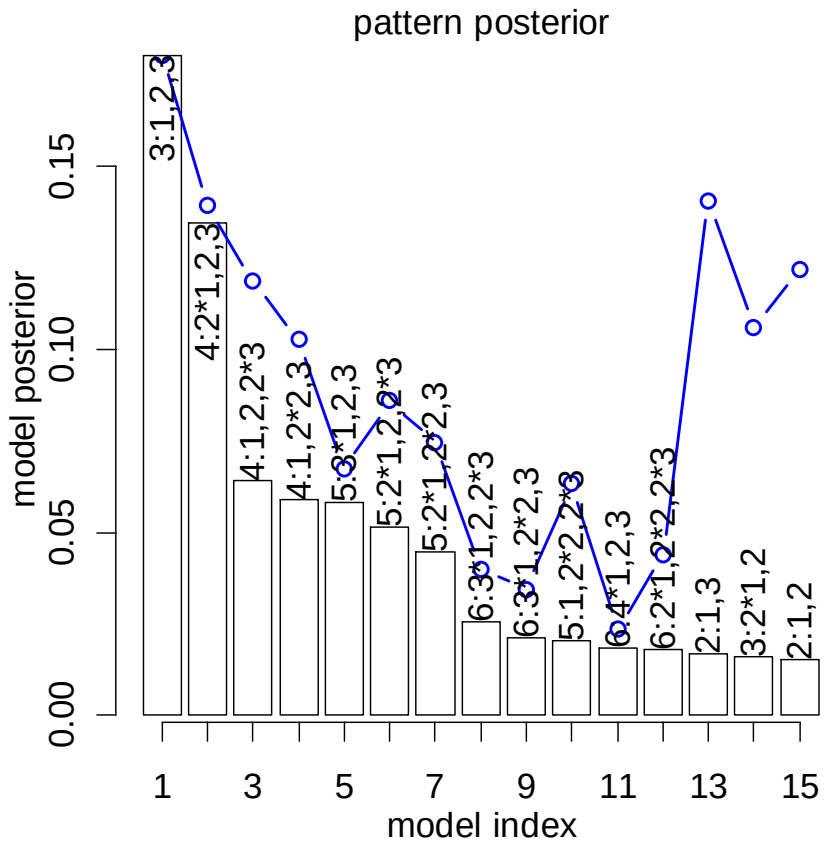


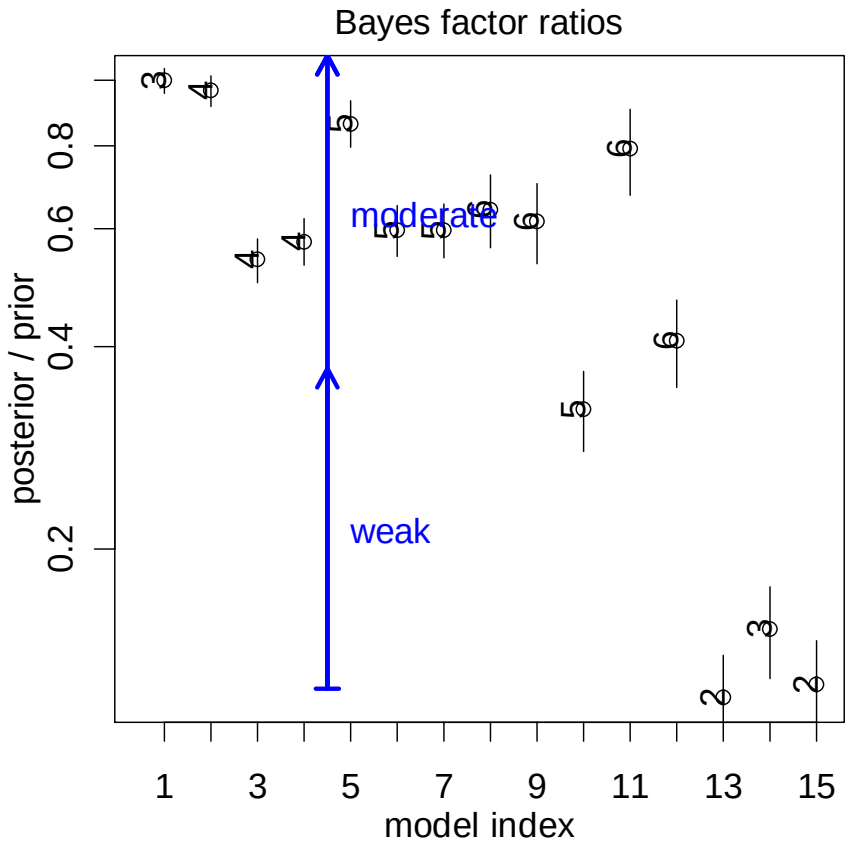






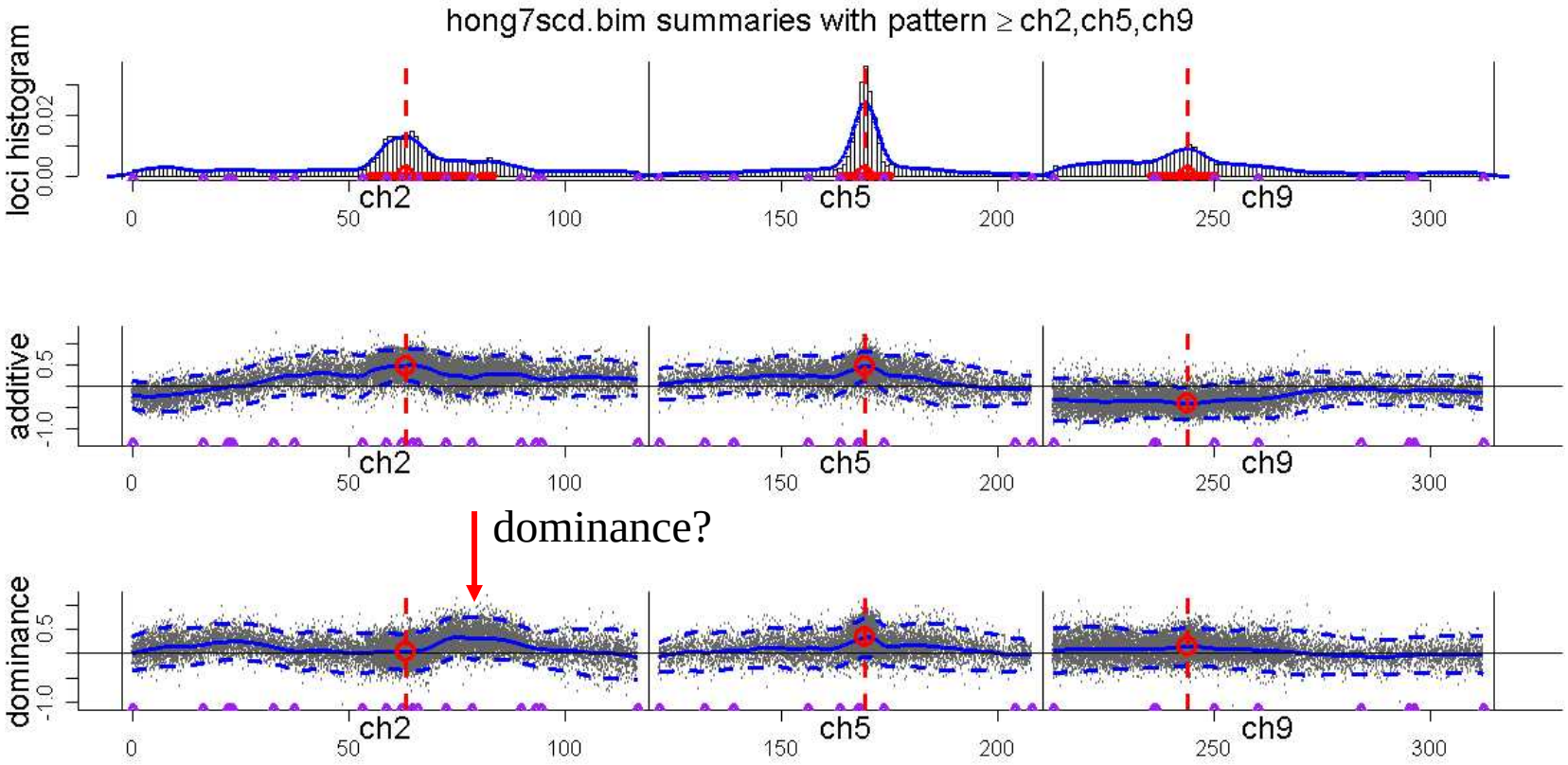
Bayesian model assessment: chromosome QTL pattern for SCD1



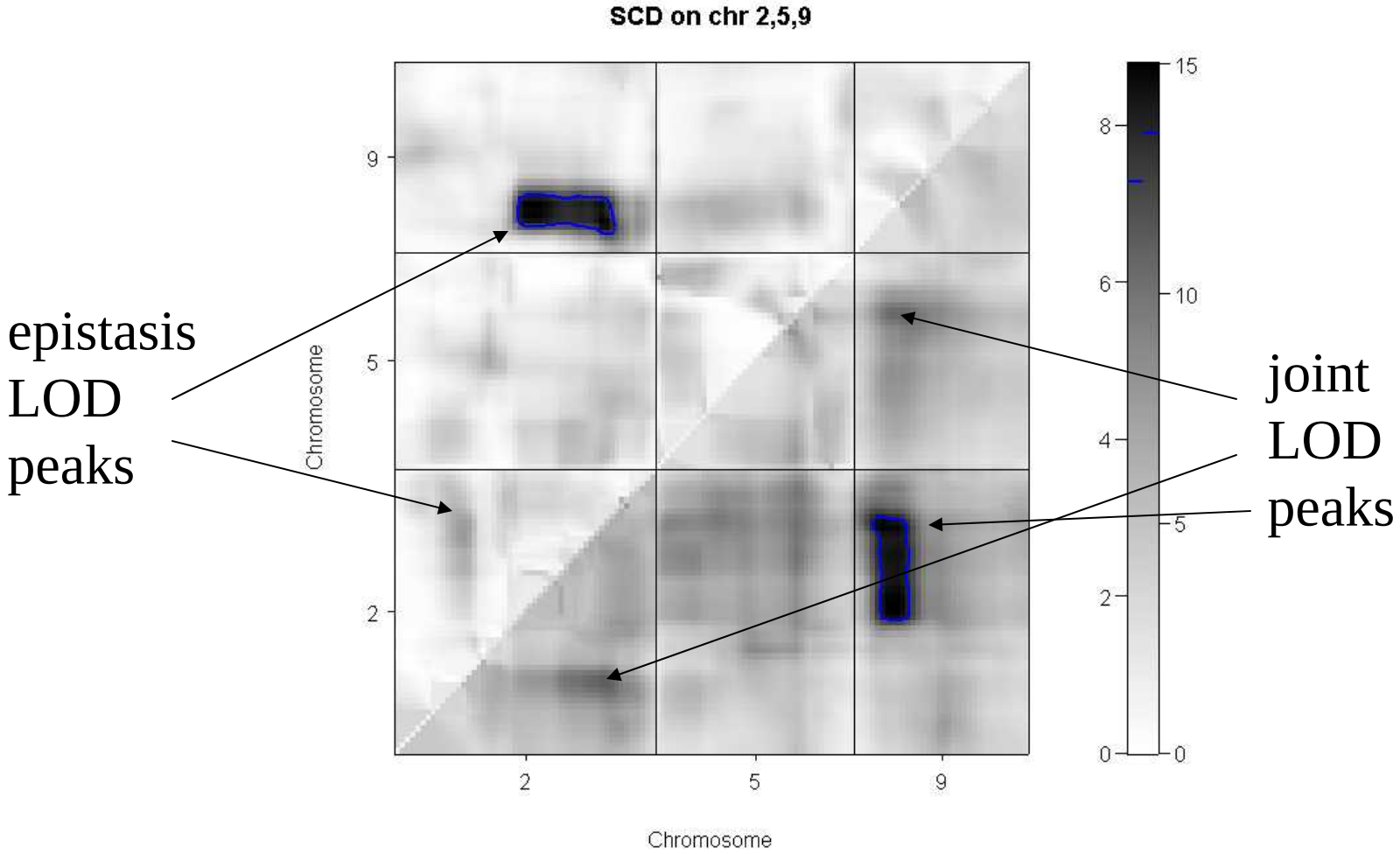


trans-acting QTL for SCD1

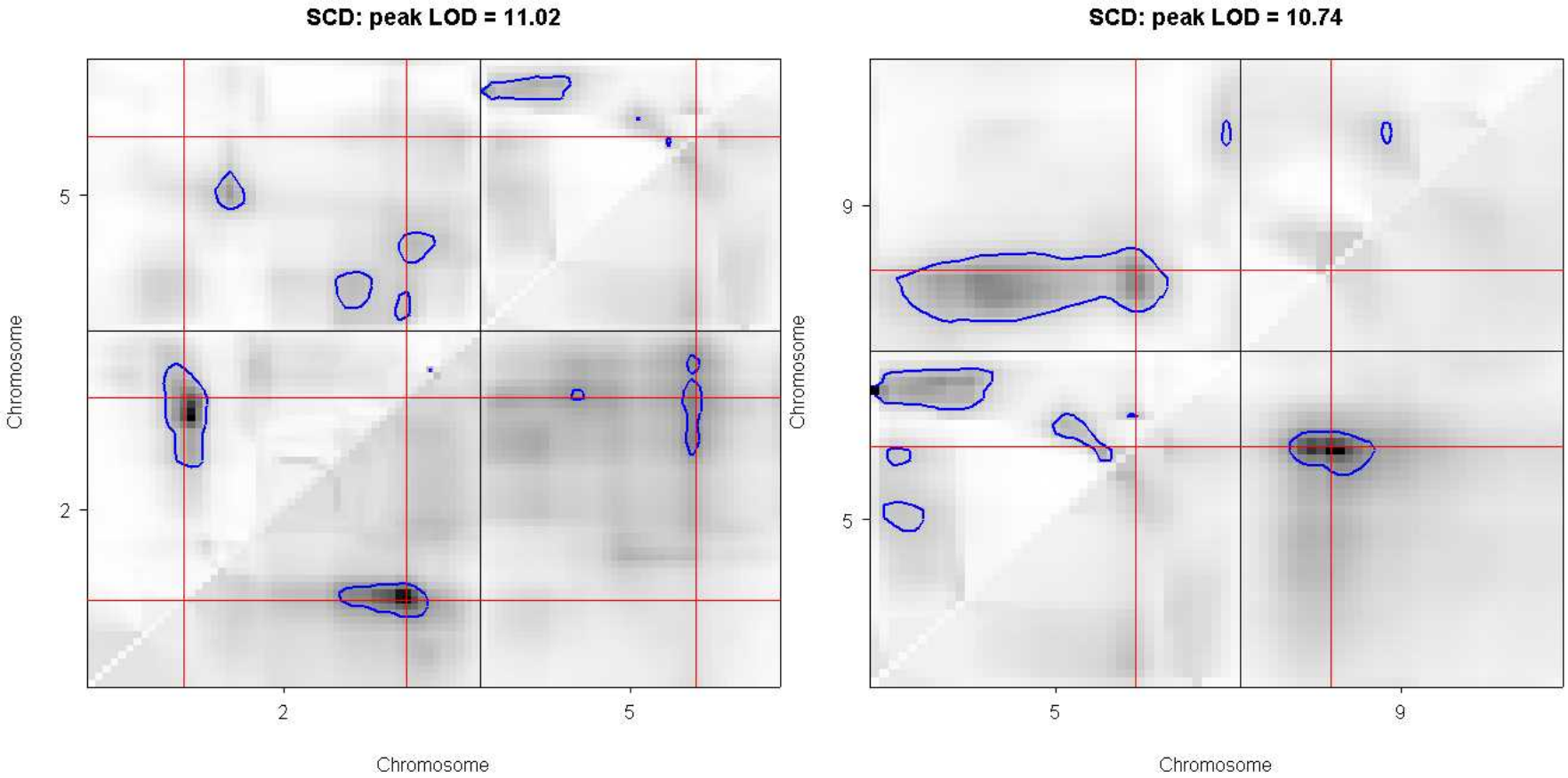
(no epistasis yet: see Yi, Xu, Allison 2003)



2-D scan: assumes only 2 QTL!

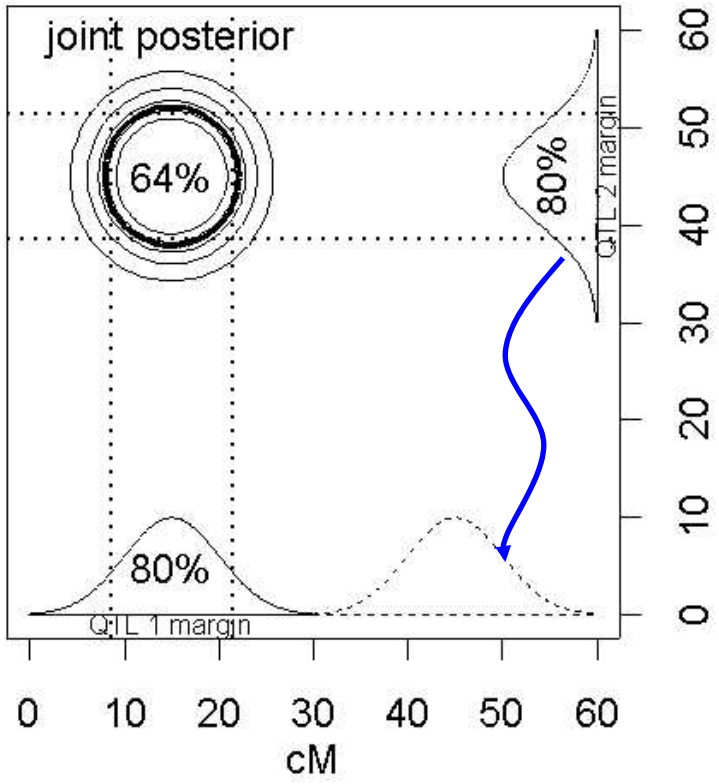


sub-peaks can be easily overlooked!

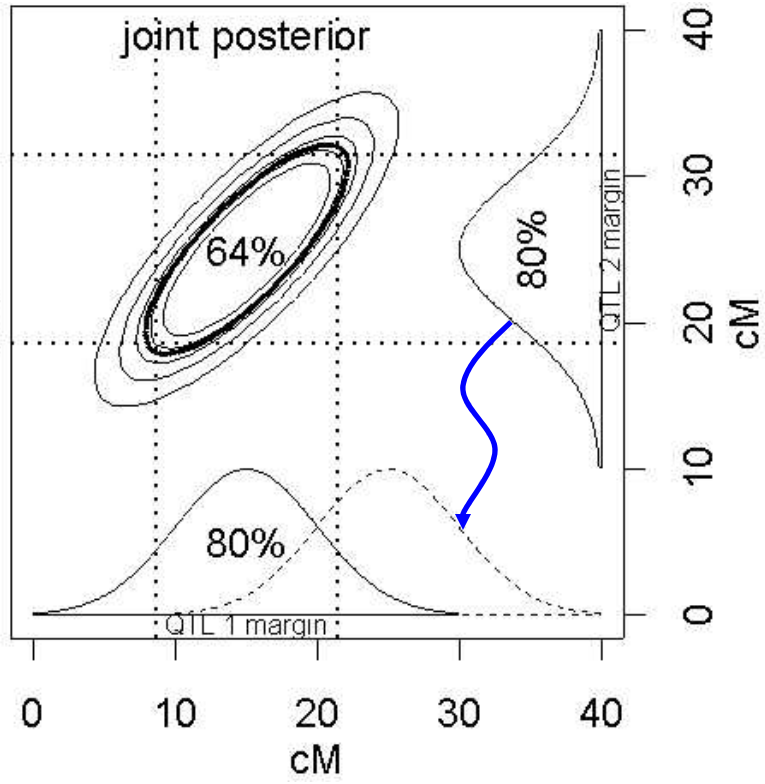


1-D and 2-D marginals
pr(QTL at λ | *Y*,*X*, *m*)

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, λ 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