NCSU Summer Institute 2003
QTL II
Brian S. Yandell
University of Wisconsin-Madison

- model selection for multiple QTL
- extending the phenotype model
- Bayesian interval mapping
- multiple traits & microarrays

contact information & resources

- Email: byandell@wisc.edu
- Web: www.stat.wisc.edu/~yandell
- www.stat.wisc.edu/~yandell/statgen
  - QTL & microarray resources
  - references with hot links to papers
  - software links
  - people links

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

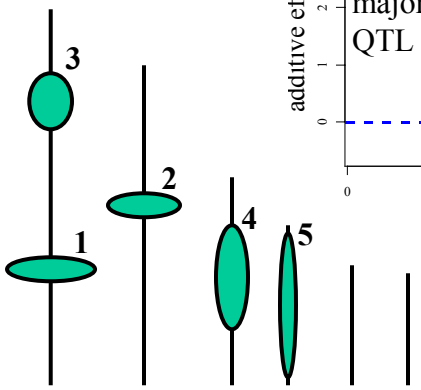
- evaluate objective
  - likelihood or posterior
- search over "space" of genetic architectures
  - number and positions of loci
  - gene action: additive, dominance, epistasis
  - how to efficiently search the model space?
- select "best" or "better" model(s)
  - what criteria to use? where to draw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

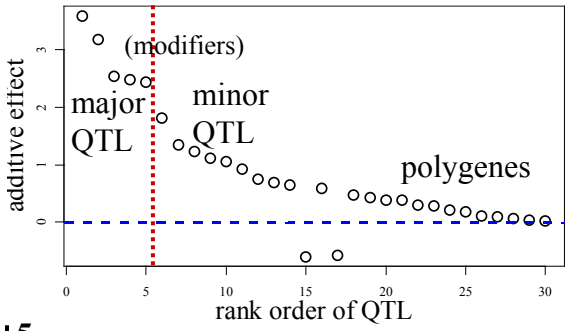
advantages of multiple QTL approach

- 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

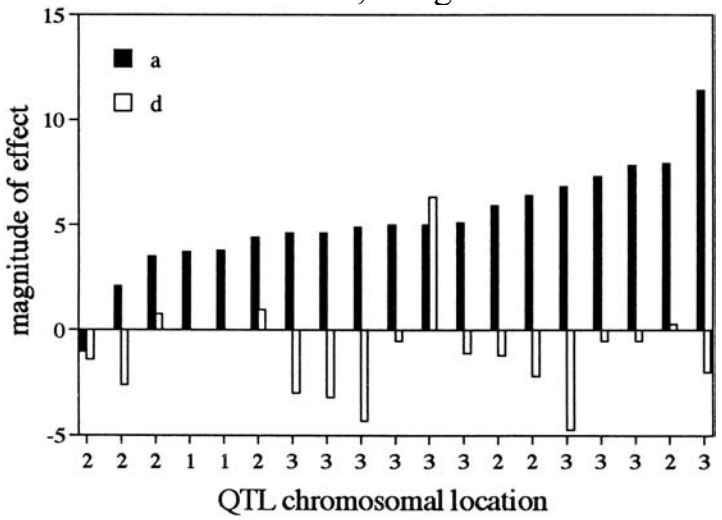
major QTL on linkage map





MIM effects for gonad shape

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



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

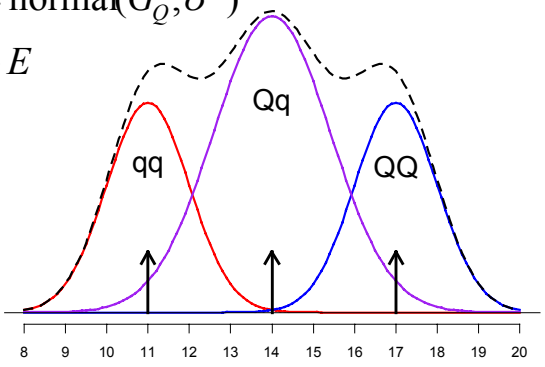
- number of QTL
  - single QTL
  - multiple QTL: known or unknown number
- location of QTL
  - known locations
  - widely spaced (no 2 in marker interval) or arbitrarily close
- gene action
  - additive (A) and/or dominance (D) effects
  - epistatic effects
    - statistical hierarchy (AA, AD, DA, DD)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

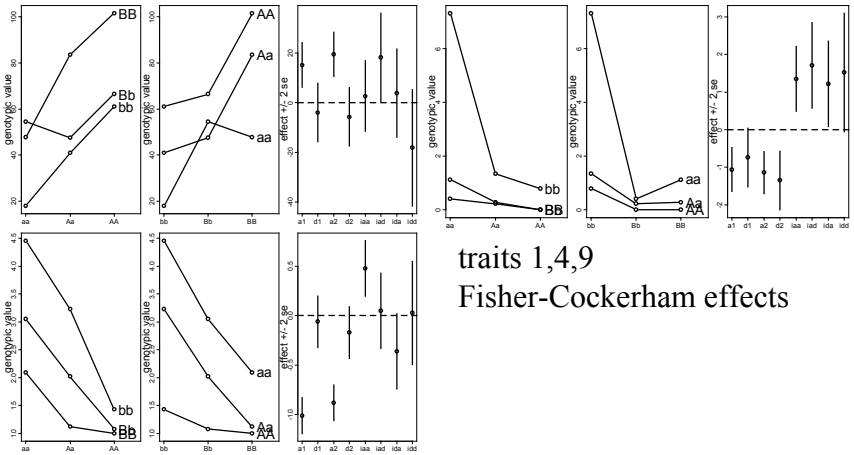
$$G_Q = \theta_{Q(1)} + \theta_{Q(2)} + \theta_{Q(1,2)}$$

- 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 *Genetics*; 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

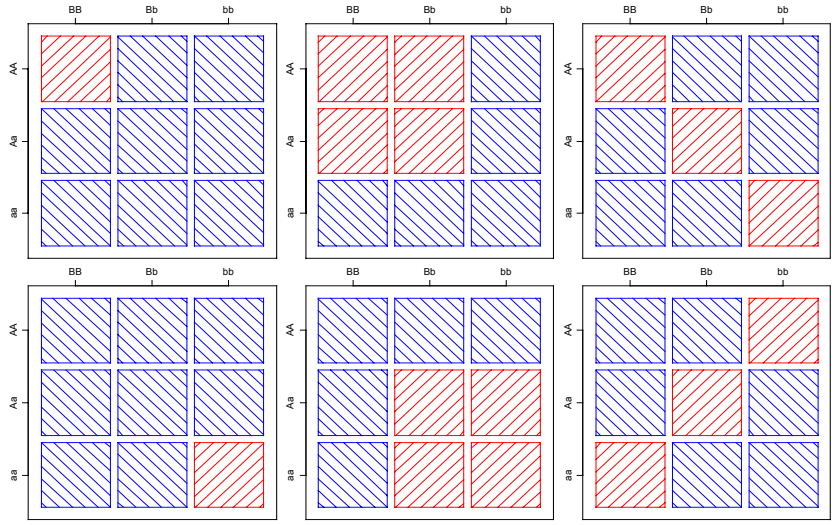
  $\sigma^2_G = \text{sum}_k \sigma^2_{kG}, \sigma^2_{kG} = \text{sum}_j \sigma^2_{kjG}, \sigma^2_{kjG} = \text{var}(\theta_{kjQ})$

- 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

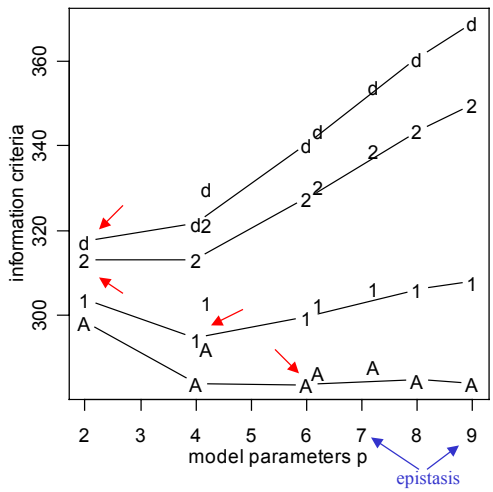
- normal data at a marker
  - likelihood $L(p) = (n/2)\log[\text{RSS}(p)]$
  - LR = ratio of likelihoods for two models
    - p_2 = df for larger model
    - p_1 = df for reduced model
  - $2\log(LR) = L(p_2) - L(p_1) = n \log [\text{RSS}(p_2)/\text{RSS}(p_1)]$
  - $LOD = \log_{10}(LR) = \log(LR)/\log(10)$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

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

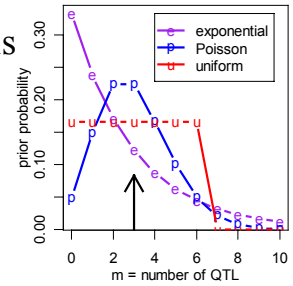
$$B_{12} = \frac{\text{pr}(\text{model}_1 \mid Y) / \text{pr}(\text{model}_2 \mid Y)}{\text{pr}(\text{model}_1) / \text{pr}(\text{model}_2)} = \frac{\text{pr}(Y \mid \text{model}_1)}{\text{pr}(Y \mid \text{model}_2)}$$

- 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



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

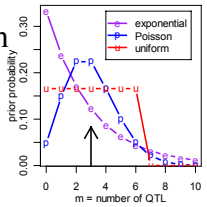
computing marginal means

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

- 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

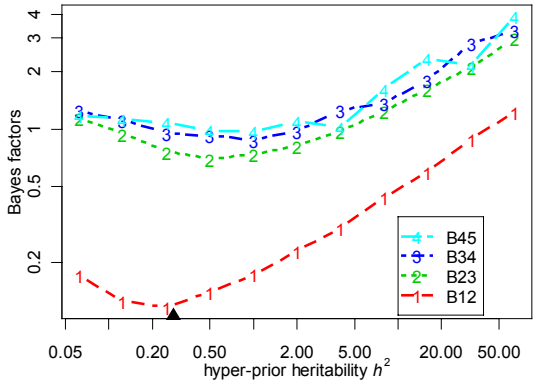
$$\mu_0 \approx \overline{Y} \text{ and } \kappa_1 \approx \frac{h^2}{1-h^2} = \frac{\sigma_G^2}{\sigma^2}$$

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

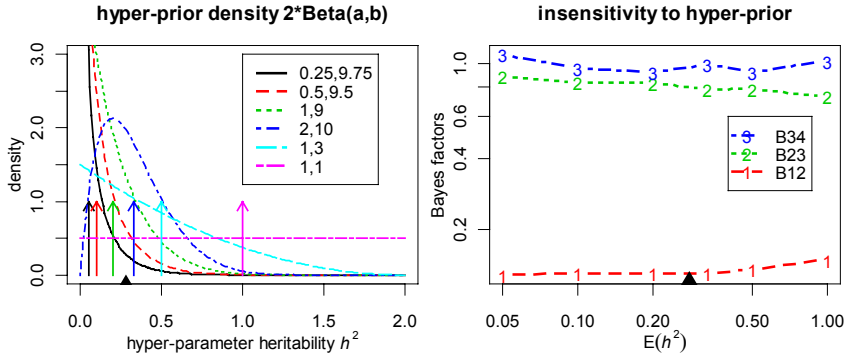
$$G_Q \mid Y, m \sim N(B_Q\hat{G}_Q, B_Q C_Q \sigma^2)$$
$$\approx N(\hat{G}_Q, C_Q\sigma^2)$$
$$\text{LS estimate } \hat{G}_Q = \sum_i \sum_j \hat{\theta}_{ijQ} = \sum_i w_{iQ} Y_i$$
$$\text{variance } V(\hat{G}_Q) = \sum_i w_{iQ}^2 \sigma^2 = C_Q\sigma^2$$
$$\text{shrinkage } B_Q = \kappa/(\kappa + C_Q) \to 1$$

BF sensitivity to fixed prior for effects



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

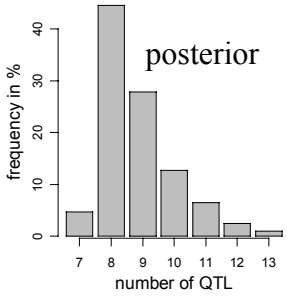
BF insensitivity to random effects prior



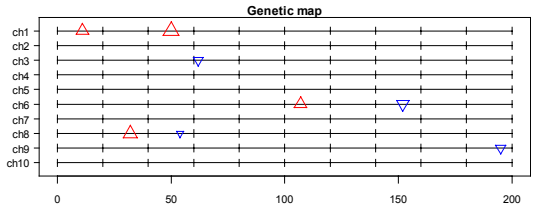
$$\theta_{jQ} \sim \mathrm{N}\left(0, \sigma_G^2 / m\right), \sigma_G^2 = h^2 \sigma_{\text{total}}^2, \tfrac{1}{2} h^2 \sim \mathrm{Beta}(a, b)$$

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

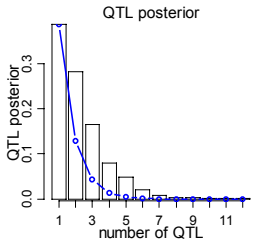
- 108 plants from double haploid
  - similar genetics to backcross: follow 1 gamete
  - parents are Major (biennial) and Stellar (annual)
- 300 markers across genome
  - 19 chromosomes
  - average 6cM between markers
    - median 3.8cM, max 34cM
  - 83% markers genotyped
- phenotype is days to flowering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

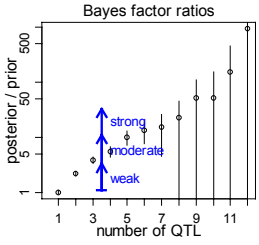
Bayesian model assessment

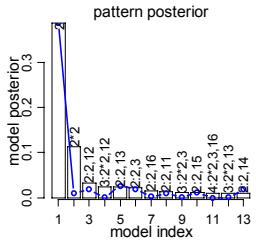
row 1: # QTL
row 2: pattern

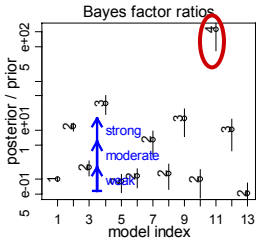
col 1: posterior
col 2: Bayes factor
note error bars on bf

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



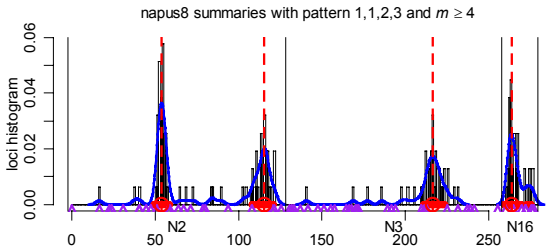




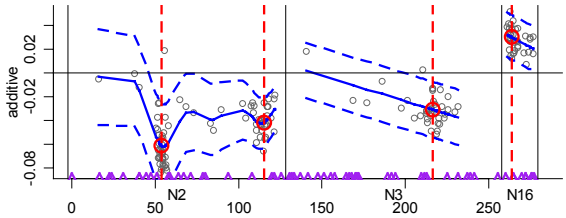


Bayesian estimates of loci & effects

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

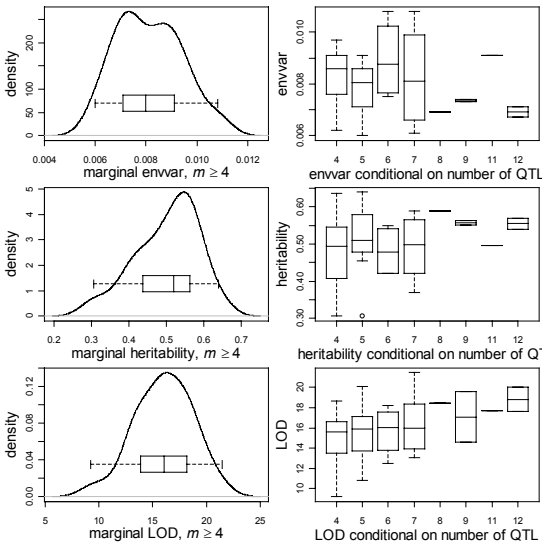


Bayesian model diagnostics

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)

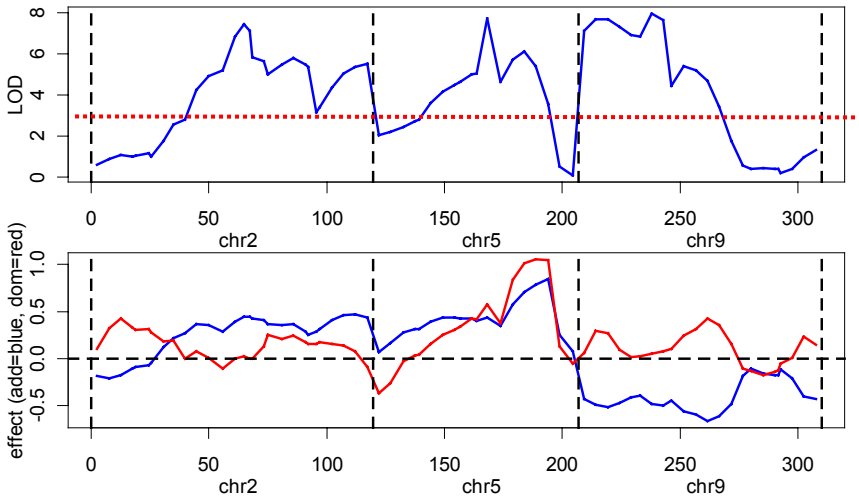
but note change with *m*



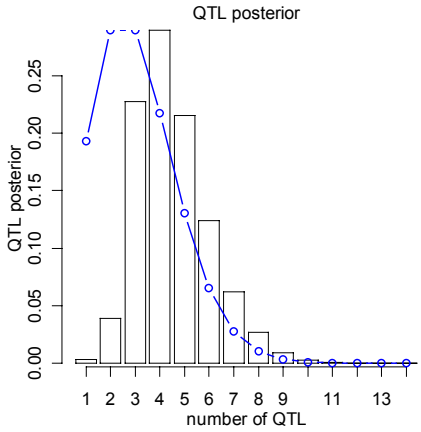
studying diabetes in an F2

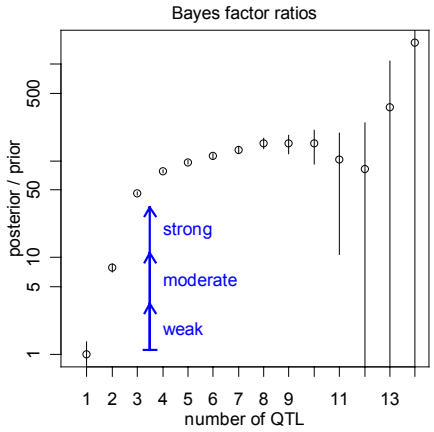
- segregating cross of inbred lines
  - B6.ob x BTBR.ob → F1 → F2
  - selected mice with ob/ob alleles at leptin gene (chr 6)
  - measured and mapped body weight, insulin, glucose at various ages (Stoehr et al. 2000 Diabetes)
  - sacrificed at 14 weeks, tissues preserved
- gene expression data
  - Affymetrix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

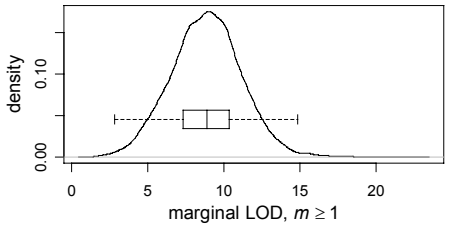


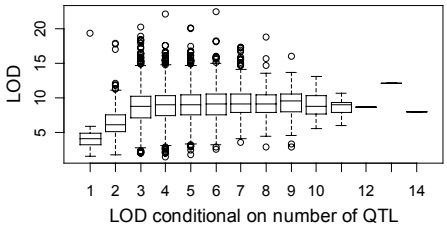
Bayesian model assessment: number of QTL for SCD1

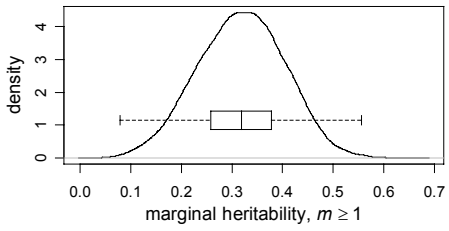


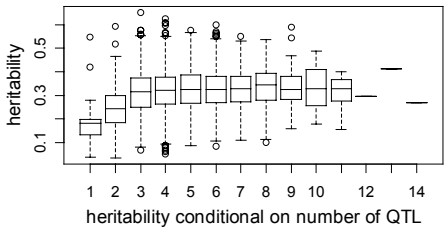


Bayesian LOD and h^2 for SCD1

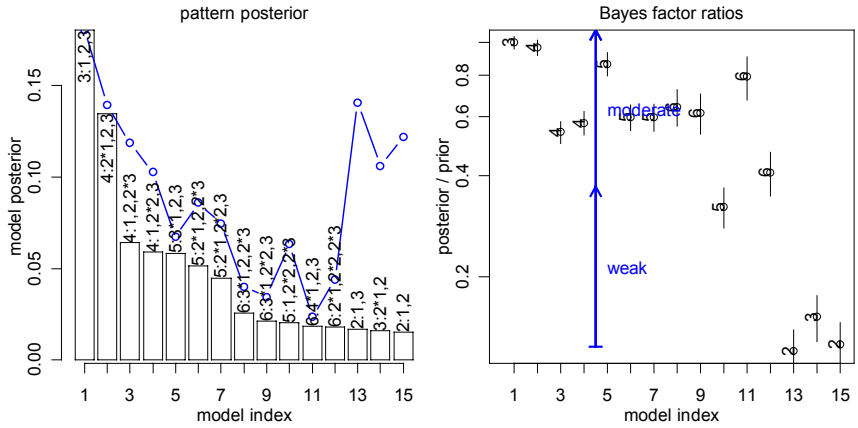






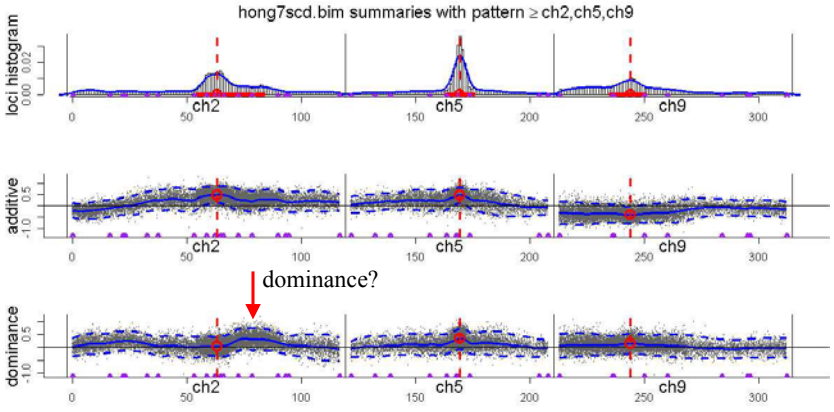


Bayesian model assessment: chromosome QTL pattern for SCD1

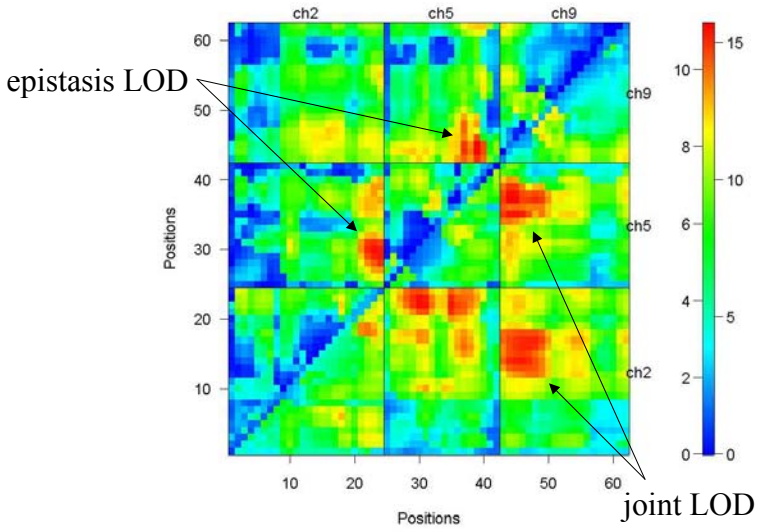


trans-acting QTL for SCD1

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

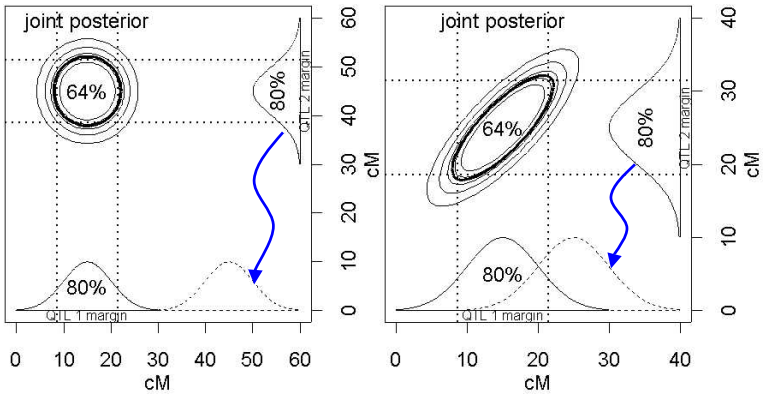


2-D scan: assumes only 2 QTL!



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

unlinked loci linked loci



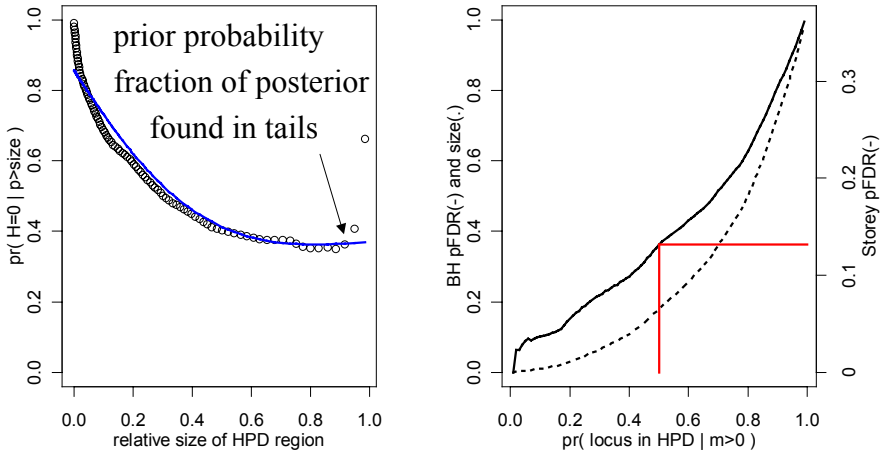
false detection rates and thresholds

- 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!

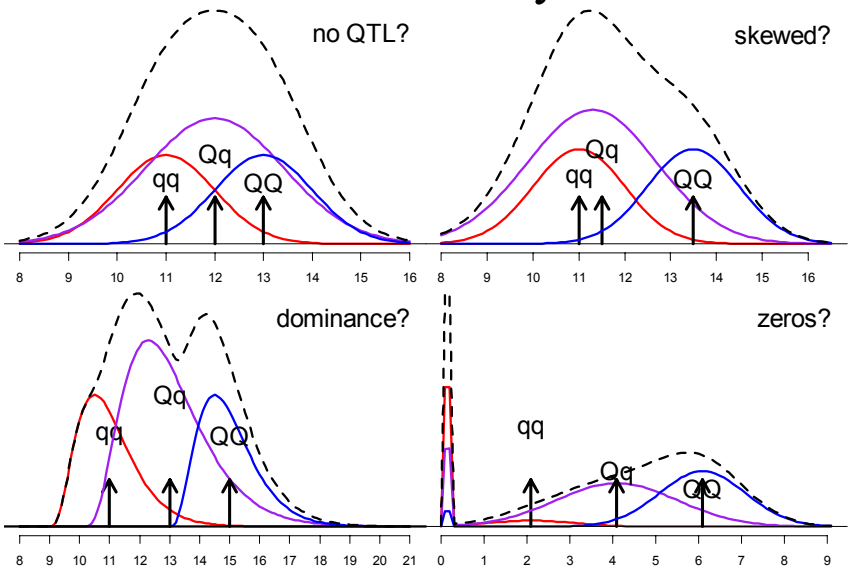
Extending the Phenotype Model

- limitations of parametric models
- diagnostic tools for QTL analysis
- QTL mapping with other parametric "families"
- quick fixes via data transformations
- semi-parametric approaches
- non-parametric approaches
- bottom line:
  - normal phenotype model works well to pick up loci, but may be poor at estimating effects if data not normal

limitations of parametric models

- measurements not normal
  - categorical traits: counts (*e.g.* number of tumors)
    - use methods specific for counts
    - binomial, Poisson, negative binomial
  - traits measured over time and/or space
    - survival time (*e.g.* days to flowering)
    - developmental process; signal transduction between cells
    - TP Speed (pers. comm.); Ma, Casella, Wu (2002)
- false positives due to miss-specified model
  - how to check model assumptions?
- want more robust estimates of effects
  - parametric: only center (mean), spread (SD)
  - shape of distribution may be important

what if data are far away from ideal?



diagnostic tools for QTL (Hackett 1997)

- illustrated with BC, adapt regression diagnostics
- normality & equal variance (fig. 1)
  - plot fitted values vs. residuals--football shaped?
  - normal scores plot of residuals--straight line?
- number of QTL: likelihood profile (fig. 2)
  - flat shoulders near LOD peak: evidence for 1 vs. 2 QTL
- genetic effects
  - effect estimate near QTL should be $(1–2r)a$
  - plot effect vs. location

marker density & sample size: 2 QTL

modest sample size
dense vs. sparse markers

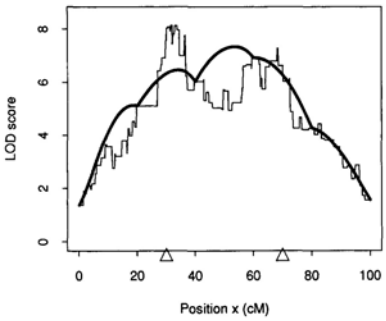


FIGURE 1.—The two-QTL true model with a QTL at 30 cM and a second QTL of somewhat smaller effect at 70 cM (true locations indicated by △). A normal single-QTL model is assumed and the LOD score for 100 simulated individuals is given for dense markers (thin curve) and markers at 20-cM intervals (bold curve).

large sample size
dense vs. sparse markers

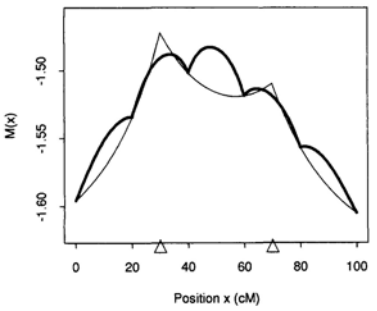


FIGURE 4.—$M(x)$ for a normal single-QTL assumed model under a two-QTL true model when both of the genes lie on the chromosome under study. This scenario was originally depicted in Figure 1. With dense markers (thin curve), $M(x)$ peaks at exactly 30 cM, the location of the QTL of stronger effect. With nondense markers at 20-cM intervals, $M(x)$ peaks at 47 cM in an incorrect interval (bold curve). Note the similarity in shape between the LODs in Figure 1 and the limiting forms depicted here.

Wright Kong (1997 *Genetics*)

robust locus estimate for non-normal phenotype

large sample size &
dense marker map:
no need for normality

but what happens for
modest sample sizes?

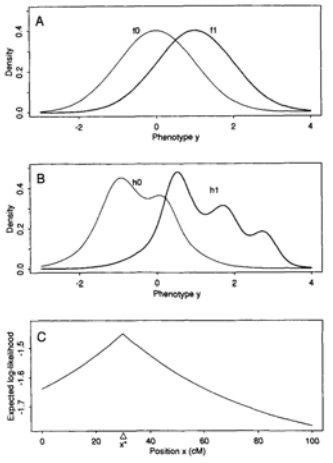


FIGURE 2.—Misspecification of the phenotype model. (A) The assumed distributions f_0 and f_1. (B) The true distributions h_0, h_1. (C) The expected log-likelihood across the chromosome when the markers are dense. Despite the misspecification, the function is maximized at exactly the true location $x^* = 30$ cM (indicated by △).

Wright Kong (1997 *Genetics*)

What shape is your histogram?

- histogram conditional on known QT genotype
  - pr(Y|qq, θ) model shape with genotype qq
  - pr(Y|Qq, θ) model shape with genotype Qq
  - pr(Y|QQ, θ) model shape with genotype QQ
- is the QTL at a given locus λ?
  - no QTL pr(Y|qq, θ) = pr(Y|Qq, θ) = pr(Y|QQ, θ)
  - QTL present mixture if genotype unknown
- mixture across possible genotypes
  - sum over Q = qq, Qq, QQ
  - pr($Y|X,\lambda,\theta$) = sum_Q pr($Q|X,\lambda$) pr($Y|Q,\theta$)

interval mapping likelihood

- likelihood: basis for scanning the genome
  - product over $i = 1,\ldots,n$ individuals

$$L(\theta,\lambda|Y) = \text{product}_i \text{ pr}(Y_i|X_i,\lambda)$$
$$= \text{product}_i \text{ sum}_Q \text{ pr}(Q|X_i,\lambda) \text{ pr}(Y_i|Q,\theta)$$

- problem: unknown phenotype model
  - parametric pr($Y|Q,\theta$) = $f(Y \mid \mu, G_Q, \sigma^2)$
  - semi-parametric pr($Y|Q,\theta$) = $f(Y)$ exp($Y\beta_Q$)
  - non-parametric pr($Y|Q,\theta$) = $F_Q(Y)$

useful models & transformations

- binary trait (yes/no, hi/lo, …)
  - map directly as another marker
  - categorical: break into binary traits?
  - mixed binary/continuous: condition on $Y > 0$?
- known model for biological mechanism
  - counts Poisson
  - fractions binomial
  - clustered negative binomial
- transform to stabilize variance
  - counts $\sqrt{Y}$ = sqrt(Y)
  - concentration log(Y) or log($Y+c$)
  - fractions arcsin($\sqrt{Y}$)
- transform to symmetry (approx. normal)
  - fraction log(Y/(1-Y)) or log(($Y+c$)/(1+c-Y))
- empirical transform based on histogram
  - watch out: hard to do well even without mixture
  - probably better to map untransformed, then examine residuals

semi-parametric QTL

- phenotype model $\text{pr}(Y|Q,\theta) = f(Y)\exp(Y\beta_Q)$
  - unknown parameters $\theta = (f, \beta)$
    - $f(Y)$ is a (unknown) density if there is no QTL
    - $\beta = (\beta_{qq}, \beta_{Qq}, \beta_{QQ})$
    - $\exp(Y\beta_Q)$ `tilts' f based on genotype Q and phenotype Y
- test for QTL at locus λ
  - $\beta_Q = 0$ for all Q, or $\text{pr}(Y|Q,\theta) = f(Y)$
- includes many standard phenotype models

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

  Poisson $\text{pr}(Y|Q,\theta) = \text{Poisson}(G_Q)$

  exponential, binomial, …, but not negative binomial

QTL for binomial data

- approximate methods: marker regression
  - Zeng (1993,1994); Visscher et al. (1996); McIntyre et al. (2001)
- interval mapping, CIM
  - Xu Atchley (1996); Yi Xu (2000)
  - $Y \sim$ binomial$(1,\pi)$, π depends on genotype Q
  - pr$(Y|Q) = (\pi_Q)^Y (1 - \pi_Q)^{(1-Y)}$
  - substitute this phenotype model in EM iteration
- or just map it as another marker!
  - but may have complex

EM algorithm for binomial QTL

- E-step: posterior probability of genotype Q

$$\text{pr}(Q \mid Y_i, X_i, \lambda, \pi_Q) = \frac{\text{pr}(Q \mid X_i, \lambda)(\pi_Q)^{Y_i}(1-\pi_Q)^{(1-Y_i)}}{\text{sum}_Q \text{ of numerator}}$$

- M-step: MLE of binomial probability π_Q

$$\pi_Q = \frac{\text{sum}_i \; Y_i \text{pr}(Q \mid Y_i, X_i, \lambda, \pi_Q)}{\text{sum}_i \; \text{pr}(Q \mid Y_i, X_i, \lambda, \pi_Q)}$$

threshold or latent variable idea

- "real", unobserved phenotype Z is continuous
- observed phenotype Y is ordinal value
  - no/yes; poor/fair/good/excellent
  - $\text{pr}(Y = j) = \text{pr}(\tau_{j-1} < Z \leq \tau_j)$
  - $\text{pr}(Y \leq j) = \text{pr}(Z \leq \tau_j)$
- use logistic regression idea (Hackett Weller 1995)
  - substitute new phenotype model in to EM algorithm
  - or use Bayesian posterior approach
  - extended to multiple QTL (papers in press)

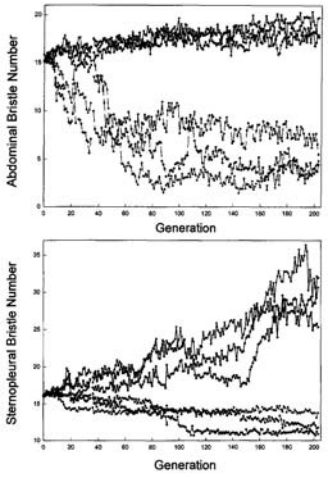
$$\text{pr}(Y \leq j \mid Q) = \text{pr}(Z \leq \tau_j \mid Q) = [1 + \exp(\mu + G_Q - \tau_j)]^{-1}$$

quantitative & qualitative traits

- Broman (2003): spike in phenotype
  - large fraction of phenotype has one value
  - map binary trait (is/is not that value)
  - map continuous trait given not that value
- multiple traits
  - Williams et al. (1999)
    - multiple binary & normal traits
    - variance component analysis
  - Corander Sillanpaa (2002)
    - multiple discrete & continuous traits
    - latent (unobserved) variables

other parametric approaches

- Poisson counts
  - Mackay Fry (1996)
    - trait = bristle number
  - Shepel et al (1998)
    - trait = tumor count
- negative binomial
  - Lan *et al.* (2001)
    - number of tumors
- exponential
  - Jansen (1992)



Mackay Fry (1996 *Genetics*)

semi-parametric empirical likelihood

- phenotype model $\text{pr}(Y|Q,\theta) = f(Y)\exp(Y\beta_Q)$
  - "point mass" at each measured phenotype Y_i
  - subject to distribution constraints for each Q:
    $1 = \text{sum}_i f(Y_i)\exp(Y_i\beta_Q)$
- non-parametric empirical likelihood (Owen 1988)

$$\begin{aligned} L(\theta,\lambda|Y,X) &= \text{product}_i\,[\text{sum}_Q\,\text{pr}(Q|X_i,\lambda)f(Y_i)\exp(Y_i\beta_Q)] \\ &= \text{product}_i\,f(Y_i)\,[\text{sum}_Q\,\text{pr}(Q|X_i,\lambda)\exp(Y_i\beta_Q)] \\ &= \text{product}_i\,f(Y_i)\,w_i \end{aligned}$$

  - weights $w_i = w(Y_i|X_i,\beta,\lambda)$ rely only on flanking markers
    - 4 possible values for BC, 9 for F2, etc.
- profile likelihood: $L(\lambda|Y,X) = \max_\theta L(\theta,\lambda|Y,X)$

semi-parametric formal tests

- clever trick: use partial empirical LOD
  - Zou, Fine, Yandell (2002 *Biometrika*)
  - Lange, Whittaker (2001 *Genetics*) GEE
- has same formal behavior as parametric LOD
  - single locus test: approximately χ^2 with 1 d.f.
  - genome-wide scan: can use same critical values
  - permutation test: possible with some work
- can estimate cumulative distributions
  - nice properties (converge to Gaussian processes)

log empirical likelihood details

$\log(L(\theta,\lambda|Y,X)) = \text{sum}_i \log(f(Y_i)) + \log(w_i)$

now profile with respect to β,λ

$\log(L(\beta,\lambda|Y,X)) = \text{sum}_i \log(f_i) + \log(w_i)$

$$+ \text{sum}_Q \alpha_Q(1-\text{sum}_i f_i \exp(Y_i\beta_Q))$$

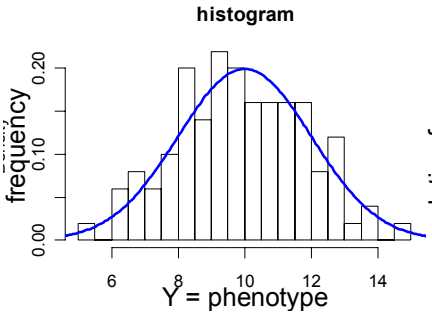
partial likelihood: set Lagrange multipliers α_Q to 0

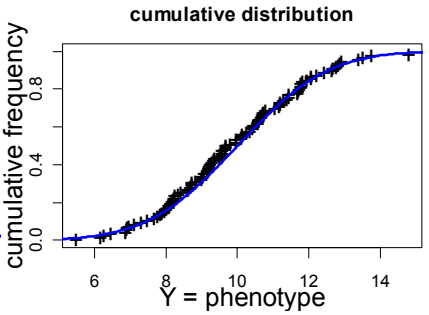
point mass density estimates

$$f_i = \left[\text{sum}_Q \exp(Y_i\beta_Q)\, p(Q \mid X, \lambda)\right]^{-1}$$

$$\text{with } p(Q \mid X, \lambda) = \text{sum}_i \text{pr}(Q \mid X_i, \lambda)$$

histograms and CDFs



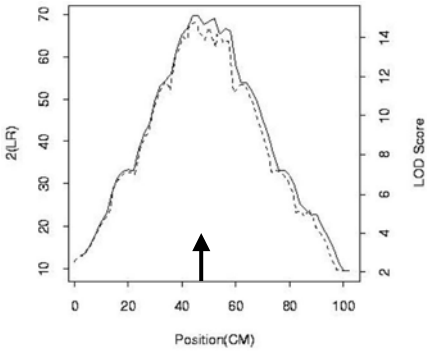


histograms capture shape but are not very accurate

CDFs are more accurate but not always intuitive

rat study of breast cancer
Lan *et al.* (2001 *Genetics*)

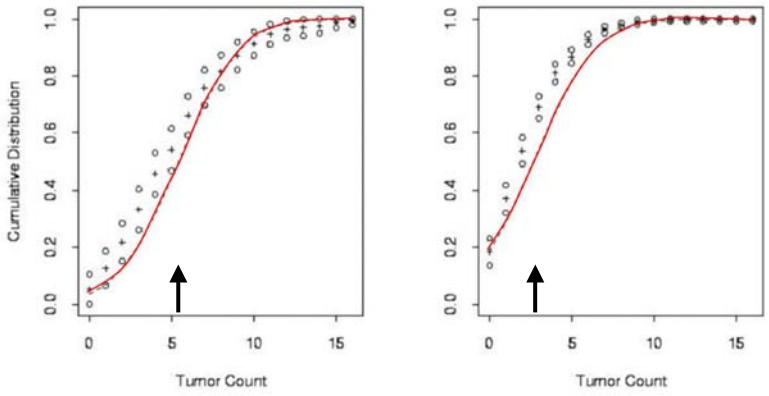
- rat backcross
  - two inbred strains
    - Wistar-Furth susceptible
    - Wistar-Kyoto resistant
  - backcross to WF
  - 383 females
  - chromosome 5, 58 markers
- search for resistance genes
- Y = # mammary carcinomas
- where is the QTL?



dash = normal
solid = semi-parametric

what shape histograms by genotype?

WF/WF WKy/WF



line = normal, + = semi-parametric, o = confidence interval

non-parametric methods

- phenotype model pr($Y|Q,\theta$) = $F_Q(Y)$
  - $\theta = F = (F_{qq},F_{Qq},F_{QQ})$ arbitrary distribution functions
- interval mapping Wilcoxon rank-sum test
  - replaced Y by rank(Y)
    - (Kruglyak Lander 1995; Poole Drinkwater 1996; Broman 2003)
  - claimed no estimator of QTL effects
- non-parametric shift estimator
  - semi-parametric shift (Hodges-Lehmann)
    - Zou (2001) thesis, Zou, Yandell, Fine (2002 in review)
  - non-parametric cumulative distribution
    - Fine, Zou, Yandell (2001 in review)
- stochastic ordering (Hoff et al. 2002)

rank-sum QTL methods

- phenotype model $\text{pr}(Y|Q,\theta) = F_Q(Y)$
- replace Y by rank(Y) and perform IM
  - extension of Wilcoxon rank-sum test
  - fully non-parametric (Kruglyak Lander 1995; Poole Drinkwater 1996)
- Hodges-Lehmann estimator of shift β
  - most efficient if $\text{pr}(Y|Q,\theta) = F(Y+Q\beta)$
  - find β that matches medians
    - problem: genotypes Q unknown
    - resolution: Haley-Knott (1992) regression scan
  - works well in practice, but theory is elusive
    - Zou, Yandell Fine (*Genetics*, in review)

non-parametric QTL CDFs

- estimate non-parametric phenotype model
  - cumulative distributions $F_Q(y) = \text{pr}(Y \le y\,|Q)$
  - can use to check parametric model validity
- basic idea:

  $\text{pr}(Y \le y\,|X,\lambda) = \text{sum}_Q\, \text{pr}(Q|X,\lambda)F_Q(y)$
  - depends on X only through flanking markers
  - few possible flanking marker genotypes
    - 4 for BC, 9 for F2, etc.

finding non-parametric QTL CDFs

- cumulative distribution $F_Q(y) = \text{pr}(Y \leq y \,|Q)$
- $F = \{F_Q$, all possible QT genotypes $Q\}$
  - BC with 1 QTL: $F = \{F_{QQ}, F_{Qq}\}$
- find F to minimize over all phenotypes y

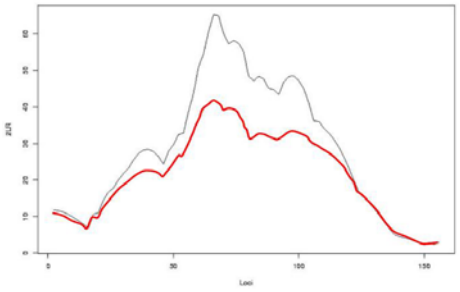
  $\text{sum}_i\, [I(Y_i \leq y) - \text{sum}_Q\, \text{pr}(Q|X,\lambda)F_Q(y)]^2$
- looks complicated, but simple to implement

non-parametric CDF properties

- readily extended to censored data
  - time to flowering for non-vernalized plants
- nice large sample properties
  - estimates of $F(y) = \{F_Q(y)\}$ jointly normal
  - point-wise, experiment-wise confidence bands
- more robust to heavy tails and outliers
- can use to assess parametric assumptions

what QTL influence flowering time? no vernalization: censored survival

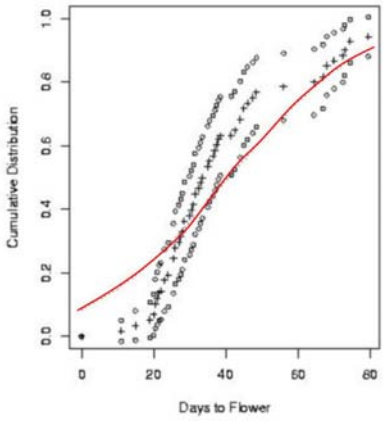
- *Brassica napus*
  - Major female
    - needs vernalization
  - Stellar male
    - insensitive
  - 99 double haploids
- *Y* = log(days to flower)
  - over 50% Major at QTL never flowered
  - log not fully effective



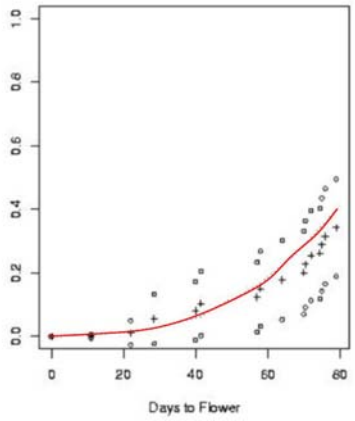
grey = normal, red = non-parametric

what shape is flowering distribution?

B. napus Stellar



B. napus Major



line = normal, + = non-parametric, o = confidence interval

Bayesian Interval Mapping

- multiple QTL likelihood
  - compare CIM, MIM, imputation, BIM
  - *Drosophila* shape example
- Bayesian idea
  - Who was Bayes? What is Bayes theorem?
  - Bayesian
  - 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…

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

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
- study properties of posterior $\text{pr}(\theta,\lambda,Q \mid Y, X)$
  - sample from posterior in some clever way
    - multiple imputation or MCMC

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

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

shape phenotype in BC study indexed by PC1

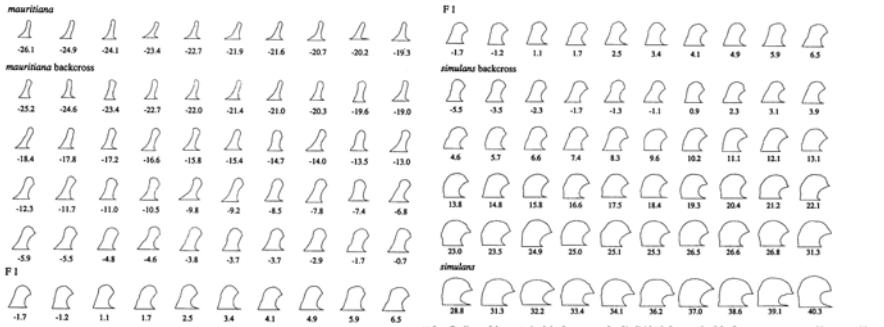


FIGURE 6.—Outlines of the posterior lobe from a sample of individuals from each of the five groups: pure *mauritiana*, *mauritiana* backcross, F_1, *simulans* backcross, and pure *simulans*. Within each group, the outlines are presented in order of their PC1 score (sampled at even intervals from the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

shape phenotype via PC

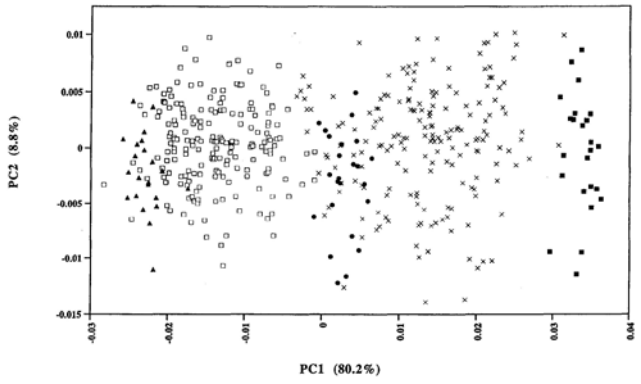


FIGURE 5.—A plot of the first two principal components of the Fourier coefficients from posterior lobe outlines. Many individuals from each of five genotypic classes are represented. Each point represents an average of scores from the left and right sides of an individual (with a few exceptions for which 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.

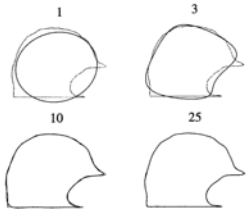


FIGURE 2.—The effect of harmonic number on the accuracy of reconstruction of a posterior lobe outline by elliptical Fourier analysis.

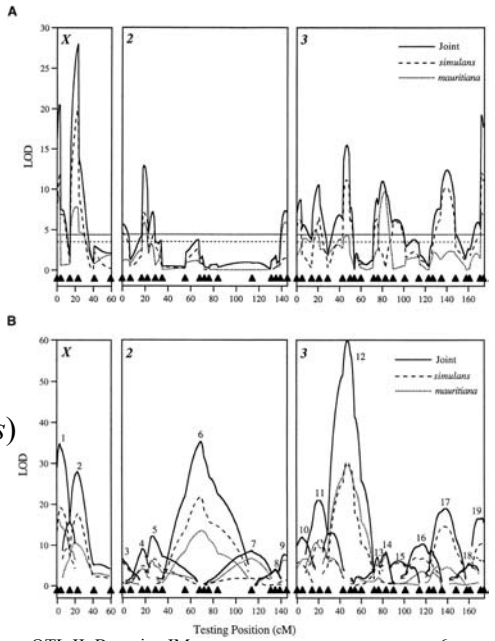
Liu et al. (1996) *Genetics*

CIM vs. MIM

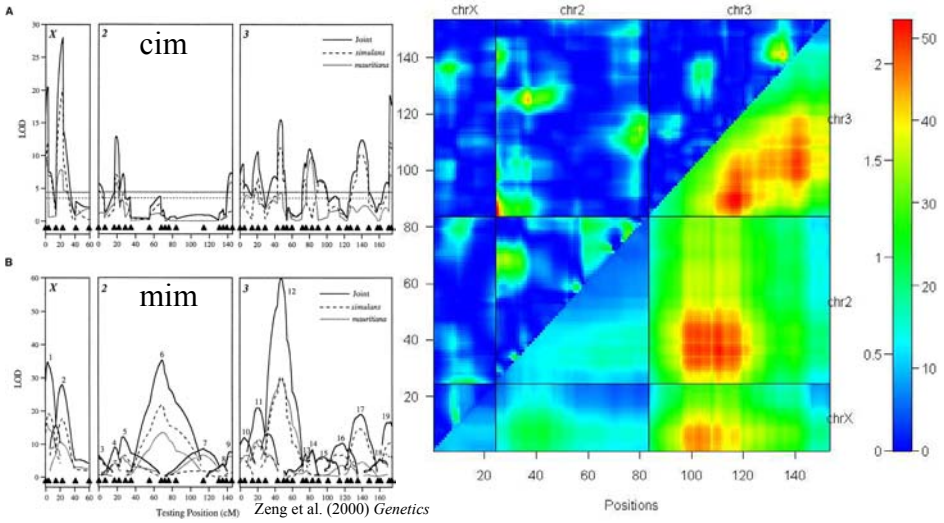
composite interval mapping
(Liu et al. 1996 *Genetics*)
narrow peaks
miss some QTL

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

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



CIM, MIM and IM pairscan

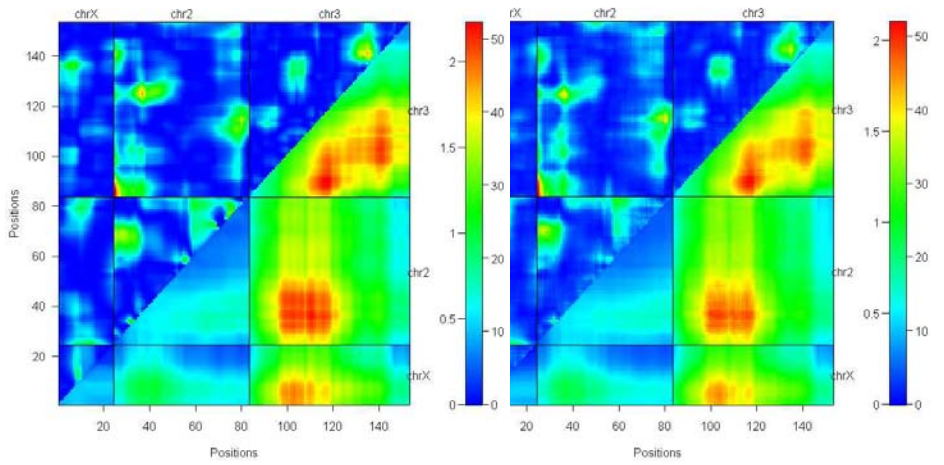


Zeng et al. (2000) *Genetics*

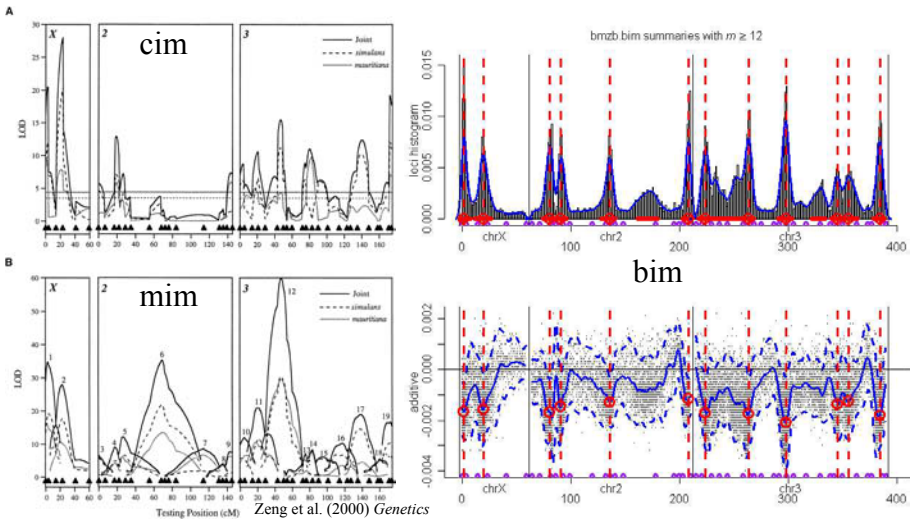
2 QTL + epistasis: IM versus multiple imputation

IM pairscan

multiple imputation



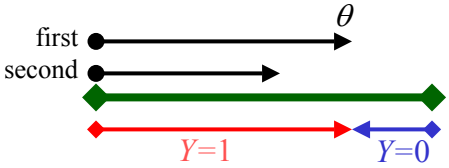
multiple QTL: CIM, MIM and BIM



Zeng et al. (2000) *Genetics*

who was Bayes?

- Reverend Thomas Bayes (1702-1761)
  - part-time mathematician
  - buried in Bunhill Cemetary, Moongate, London
  - famous paper in 1763 *Phil Trans Roy Soc London*
  - was Bayes the first with this idea? (Laplace)
- billiard balls on rectangular table
  - two balls tossed at random (uniform) on table
  - where is first ball if the second is to its left (right)?

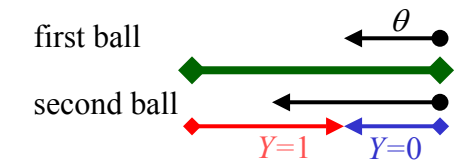


prior $\quad \Pr(\theta) \quad = 1$

likelihood $\quad \Pr(Y \mid \theta) = \theta^{Y}(1-\theta)^{1-Y}$

posterior $\quad \Pr(\theta \mid Y) = ?$

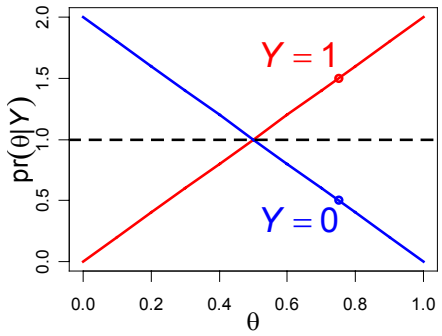
where is the first ball?



prior $\quad \Pr(\theta) \quad = 1$

likelihood $\quad \Pr(Y \mid \theta) = \theta^{Y}(1-\theta)^{1-Y}$

posterior $\quad \Pr(\theta \mid Y) = ?$



$$\Pr(\theta \mid Y) = \frac{\Pr(Y \mid \theta)\Pr(\theta)}{\Pr(Y)}$$

$$\Pr(Y) = \int_0^1 \theta^{Y}(1-\theta)^{1-Y}\, d\theta = \frac{1}{2}$$

$$\Pr(\theta \mid Y) = \begin{cases} 2\theta & Y = 1 \\ 2(1-\theta) & Y = 0 \end{cases}$$

(now throw second ball *n* times)

what is Bayes theorem?

- before and after observing data
  - prior: $\Pr(\theta)$ = pr(parameters)
  - posterior: $\Pr(\theta|Y)$ = pr(parameters|data)
- posterior = likelihood * prior / constant
  - usual likelihood of parameters given data
  - normalizing constant pr(*Y*) depends only on data
    - constant often drops out of calculation

$$\Pr(\theta \mid Y) = \frac{\Pr(\theta, Y)}{\Pr(Y)} = \frac{\Pr(Y \mid \theta) \times \Pr(\theta)}{\Pr(Y)}$$

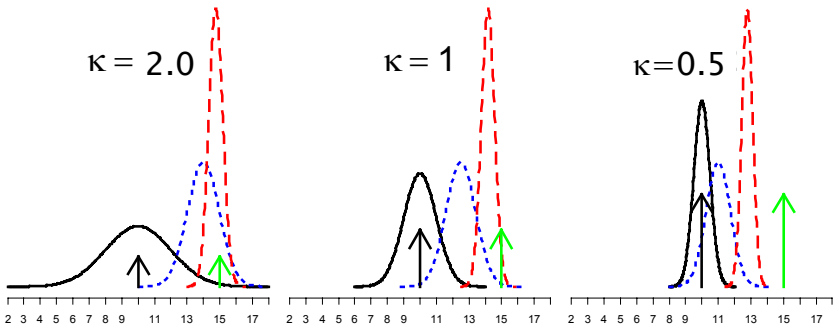
Bayes for normal data

$Y = \mu + E$	posterior for single individual
environ	$E \sim N(0, \sigma^2)$, σ^2 known
likelihood	$pr(Y \mid \mu, \sigma^2) = N(Y \mid \mu, \sigma^2)$
prior	$pr(\mu \mid \mu_0, \sigma^2, \kappa) = N(\mu \mid \mu_0, \kappa\sigma^2)$
posterior	$N(\mu \mid \mu_0 + B_1(Y - \mu_0),\ B_1\sigma^2)$
$Y_i = \mu + E_i$	posterior for sample of n individuals
	shrinkage weights B_n go to 1

$$pr(\mu \mid Y, \mu_0, \sigma^2, \kappa) = N\left(G \middle| \mu_0 + B_n(\overline{Y}_\bullet - \mu_0), B_n \frac{\sigma^2}{n}\right)$$

$$\text{with } \overline{Y}_\bullet = \text{sum}\frac{Y_i}{n}, B_n = \frac{\kappa n}{\kappa n + 1} \to 1$$

effect of prior variance on posterior



normal prior, posterior for $n = 1$, posterior for $n = 5$, true mean
(solid black) (dotted blue) (dashed red) (green arrow)

Bayesian priors for QTL

- locus λ may be uniform over genome
  - pr($\lambda \mid X$) = 1 / length of genome
- missing genotypes Q
  - pr($Q \mid X, \lambda$)
  - recombination model is formally a prior
- effects $\theta = (\mu, G, \sigma^2)$, $G = (G_{QQ}, G_{Qq}, G_{qq})$
  - conjugate priors for normal phenotype
  - $\mu \sim N(0, \kappa_0\sigma^2)$
  - $G_Q \sim N(0, \kappa\sigma^2)$
  - $\sigma^2 \sim$ inverse-$\chi^2(\nu,\tau^2)$, or $\nu\tau^2 / \sigma^2 \sim \chi^2$

details of phenotype priors

- priors depend on "hyper-parameters"
- $\mu \sim N(\mu_0, \kappa_0\sigma^2)$ grand mean
- $G_Q \sim N(0, \kappa\sigma^2)$
  - $\kappa\sigma^2 \approx \sigma_G^2$ = genetic variance
  - $\kappa \approx \sigma_G^2 / \sigma^2 = h^2 / (1-h^2)$
  - $h^2 = \sigma_G^2 / (\sigma_G^2 + \sigma^2)$ = heritability
- $\sigma^2 \sim$ inverse-$\chi^2(\nu,\tau^2)$, or $\nu\tau^2 / \sigma^2 \sim \chi^2$
  - $\tau^2 \approx s^2$ = total sample variance
  - ν = prior degrees of freedom = small integer

posterior by QT genetic value

$Y = \mu + G_Q + E$ genetics Q = qq,Qq,QQ

environment $E \sim N(0,\sigma^2)$, σ^2 known

parameters $\theta = (\mu, G, \sigma^2)$

likelihood $\text{pr}(Y \mid Q, G, \sigma^2) = N(Y \mid G_Q, \sigma^2)$

prior $\text{pr}(G_Q \mid \sigma^2, \kappa) = N(G_Q \mid 0, \kappa\sigma^2)$

posterior:

$$\text{pr}(G_Q \mid Y, Q, \sigma^2, \mu, \kappa) = N\left(G_Q \middle| B_Q(\bar{Y}_Q - \mu), B_Q \frac{\sigma^2}{n_Q}\right)$$

$$n_Q = \text{count}\{Q_i = Q\}, \bar{Y}_Q = \underset{\{i:Q_i=Q\}}{\text{sum}} \frac{Y_i}{n_Q}, B_Q = \frac{\kappa n_Q}{\kappa n_Q + 1} \to 1$$

Empirical Bayes: choosing hyper-parameters

How do we choose hyper-parameters μ, κ?

Empirical Bayes: marginalize over prior

estimate μ, κ from marginal posterior

likelihood $\text{pr}(Y_i \mid Q_i, G, \sigma^2) = N(Y_i \mid G(Q_i), \sigma^2)$

prior $\text{pr}(G_Q \mid \sigma^2, \kappa) = N(G_Q \mid 0, \kappa\sigma^2)$

marginal $\text{pr}(Y_i \mid \sigma^2, \mu, \kappa_0, \kappa) = N(Y_i \mid \mu, (\kappa_0 + \kappa + 1)\sigma^2)$

estimates $\hat{\mu}_0 = \bar{Y}_\bullet,\ s^2 = \text{sum}_i (Y_i - \bar{Y}_\bullet)^2 / n$

$\hat{\sigma}^2 = s^2/(\kappa + 1) \approx s^2/(1 - h^2)$

EB posterior $$\text{pr}(G_Q \mid Y) = N\left(G_Q \middle| B_Q(\bar{Y}_Q - \bar{Y}_\bullet), B_Q \frac{\hat{\sigma}^2}{n_Q}\right)$$

What if variance σ^2 is unknown?

- recall that sample variance is proportional to chi-square
  - $\text{pr}(s^2 \mid \sigma^2) = \chi^2\,(ns^2/\sigma^2 \mid n)$
  - or equivalently, $ns^2/\sigma^2 \mid \sigma^2 \sim \chi_n^2$
- conjugate prior is inverse chi-square
  - $\text{pr}(\sigma^2 \mid \nu,\tau^2) = \text{inv-}\chi^2\,(\sigma^2 \mid \nu,\tau^2)$
  - or equivalently, $\nu\tau^2/\sigma^2 \mid \nu,\tau^2 \sim \chi_\nu^2$
  - empirical choice: $\tau^2 = s^2/3$, $\nu=6$
    - $E(\sigma^2 \mid \nu,\tau^2) = s^2/2$, $\text{Var}(\sigma^2 \mid \nu,\tau^2) = s^4/4$
- posterior given data
  - $\text{pr}(\sigma^2 \mid Y,\nu,\tau^2) = \text{inv-}\chi^2\,(\sigma^2 \mid \nu+n,\ (\nu\tau^2+ns^2)/(\nu+n))$
  - weighted average of prior and data

joint effects posterior details

$Y_i = \mu + G(Q_i) + E_i$ genetic Q_i = qq,Qq,QQ
environ $E \sim \text{N}(0,\sigma^2)$
parameters $\theta = (\mu, G, \sigma^2)$

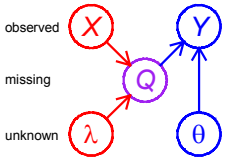
likelihood $\text{pr}(Y_i \mid Q_i, G, \sigma^2) = \text{N}(Y_i \mid G(Q_i), \sigma^2)$

prior $\text{pr}(G_Q \mid \sigma^2, \kappa) = \text{N}(G_Q \mid 0,\ \sigma^2/\kappa)$
$\text{pr}(\sigma^2 \mid \nu,\tau^2) = \text{inv-}\chi^2\,(\sigma^2 \mid \nu,\tau^2)$

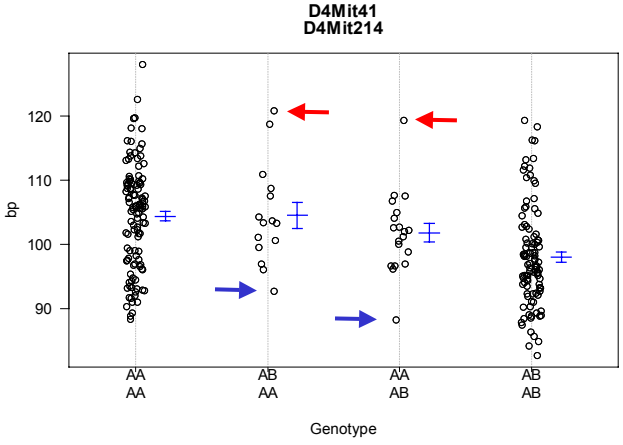
posterior:
$$\text{pr}(G_Q \mid Y, Q, \sigma^2, \kappa) = \text{N}\left(G_Q \middle| B_Q(\bar{Y}_Q - \bar{Y}_\cdot), B_Q\frac{\sigma^2}{n_Q}\right)$$
$$\text{pr}(\sigma^2 \mid Y, Q, G_Q, \nu, \tau^2) = \text{inv-}\chi^2\left(\sigma^2 \middle| \nu+n, \frac{\nu\tau^2 + ns_Q^2}{\nu+n}\right)$$
$$\text{with } B_Q = \frac{n_Q}{\kappa+n_Q}, s_Q^2 = \text{sum}_i\left(Y_i - G(Q_i)\right)^2/n$$

uncertainty in QTL genotype *Q*

- how to improve guess on *Q* with data, parameters?
  - prior recombination: pr($Q \mid X_i, \lambda$)
  - posterior recombination: pr($Q \mid Y_i, X_i, \theta, \lambda$)
- main philosophies for assessing likelihood
  - maximum likelihood: study peak(s)
  - Bayesian analysis: study whole shape
- implementation methodologies
  - Expectation-Maximization (EM)
  - Markov chain Monte Carlo (MCMC)
  - multiple imputation
  - genetic algorithms, GEE, …



how does phenotype *Y* affect *Q*?



what are probabilities for genotype *Q* between markers?

recombinants AA:AB

all 1:1 if ignore *Y* and if we use *Y*?

posterior on QTL genotypes

- full conditional of Q given data, parameters
  - proportional to prior $\text{pr}(Q \mid X_i, \lambda)$
    - weight toward Q that agrees with flanking markers
  - proportional to likelihood $\text{pr}(Y_i \mid Q, \theta)$
    - weight toward Q so that group mean $G_Q \approx Y_i$
- phenotype and flanking markers may conflict
  - posterior recombination balances these two weights

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

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
- hard to sample (λ, Q, θ, m) from joint posterior
  - update (λ, Q, θ) from full conditionals for m-QTL model
  - update m using reversible jump technology

$$(\lambda, Q, \theta, m) \sim \text{pr}(\lambda, Q, \theta, m \mid Y, X)$$

$$(\lambda, Q, \theta, m)_1 \rightarrow (\lambda, Q, \theta, m)_2 \rightarrow \cdots \rightarrow (\lambda, Q, \theta, m)_N$$

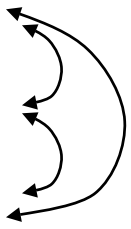
MCMC sampling of (λ, Q, θ)

- sample missing genotypes Q
  - decouples effects θ from QTL λ
  - but Q depends on (θ,λ) and vice versa
- cycle updates using full conditionals:

$$\lambda \sim \frac{\text{pr}(Q \mid X, \lambda)\text{pr}(\lambda \mid X)}{\text{pr}(Q \mid X)}$$

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

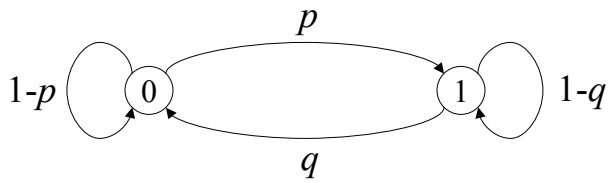
$$\theta \sim \frac{\text{pr}(Y \mid Q, \theta)\text{pr}(\theta)}{\text{pr}(Y \mid Q)}$$



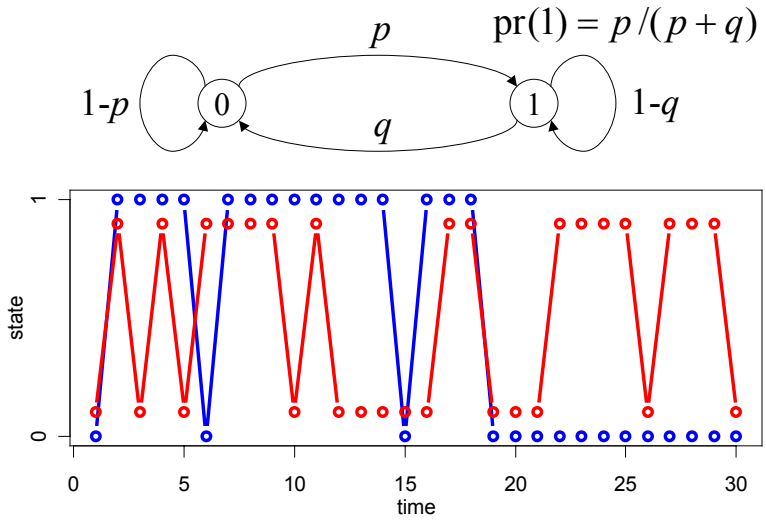
What is a Markov chain?

- future given present is independent of past
- update chain based on current value
  - can make chain arbitrarily complicated
  - chain converges to stable pattern $\pi()$ we wish to study

$$\text{pr}(1) = p/(p+q)$$



Markov chain idea



Gibbs sampler idea

- want to study two correlated normals
- could sample directly from bivariate normal
- Gibbs sampler:
  - sample each from its full conditional
  - pick order of sampling at random
  - repeat *N* times

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}\Big|\,\mu,\rho \sim N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix},\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$

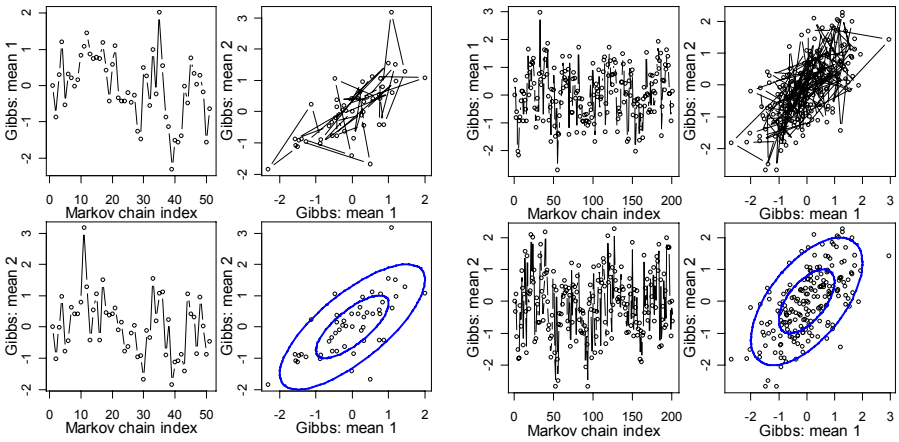
$$\theta_1 \mid \theta_2,\mu,\rho \sim N\left(\mu_1+\rho(\theta_2-\mu_2),1-\rho^2\right)$$

$$\theta_2 \mid \theta_1,\mu,\rho \sim N\left(\mu_2+\rho(\theta_1-\mu_1),1-\rho^2\right)$$

Gibbs sampler samples: $\rho = 0.6$

$N = 50$ samples

$N = 200$ samples



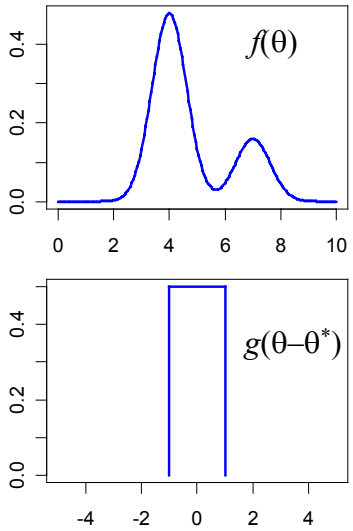
Gibbs Sampler: effects & genotypes

- for given locus λ, can sample effects θ and genotypes Q
  - effects parameter vector $\theta = (G, \sigma^2)$ with $G = (G_{qq}, G_{Qq}, G_{QQ})$
  - missing genotype vector $Q = (Q_1, Q_2, \ldots, Q_n)$
- Gibbs sampler: update one at a time via full conditionals
  - randomly select order of unknowns
  - update each given current values of all others, locus λ and data (Y,X)
    - sample variance σ^2 given Y,Q and genetic values G
    - sample genotype Q_i given markers X_i and locus λ
  - can do block updates if more efficient
    - sample all genetic values G given Y,Q and variance σ^2

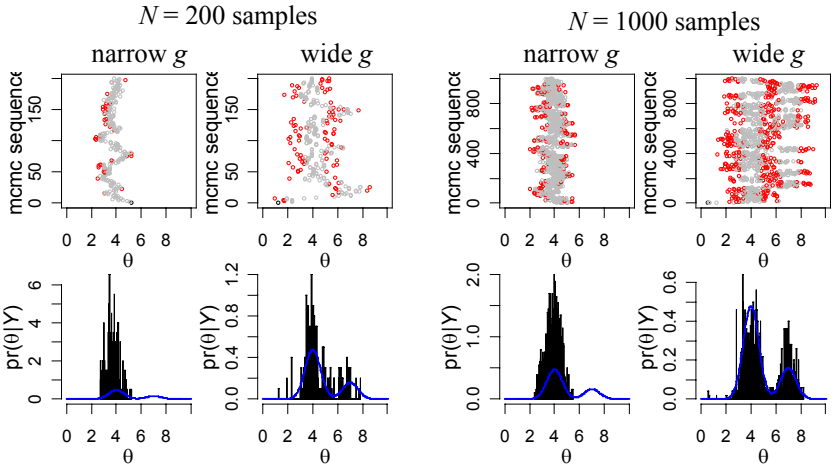
Metropolis-Hastings idea

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

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



Metropolis-Hastings samples



full conditional for locus

- cannot easily sample from locus full conditional

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

  $= \text{pr}(\lambda)\ \text{pr}(Q \mid X, \lambda)$ /constant
- cannot explicitly determine full conditional
  - difficult to normalize
  - need to average over all possible genotypes over entire map
- Gibbs sampler will not work
  - but can use method based on ratios of probabilities…

Metropolis-Hastings Step

- pick new locus based upon current locus
  - propose new locus from distribution $q(\ )$
    - pick value near current one?
    - pick uniformly across genome?
  - accept new locus with probability $a()$
- Gibbs sampler is special case of M-H
  - always accept new proposal
- acceptance insures right stable distribution
  - accept new proposal with probability A
  - otherwise stick with current value

$$A(\lambda_{old}, \lambda_{new}) = \min\left(1, \frac{\pi(\lambda_{new} \mid \mathrm{x}^*) q(\lambda_{new}, \lambda_{old})}{\pi(\lambda_{old} \mid \mathrm{x}^*) q(\lambda_{old}, \lambda_{new})}\right)$$

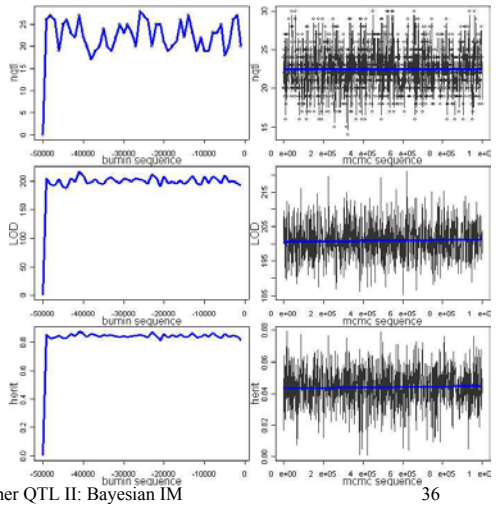
Markov chain Monte Carlo

- can study arbitrarily complex models
  - need only specify how parameters affect each other
  - can reduce to specifying full conditionals
- construct Markov chain with "right" model
  - joint posterior of unknowns as limiting "stable" distribution
  - update unknowns given data and all other unknowns
    - sample from full conditionals
    - cycle at random through all parameters
  - next step depends only on current values
- nice Markov chains have nice properties
  - sample summaries make sense
  - consider almost as random sample from distribution
  - ergodic theorem and all that stuff

MCMC diagnostics for *Dm* shape

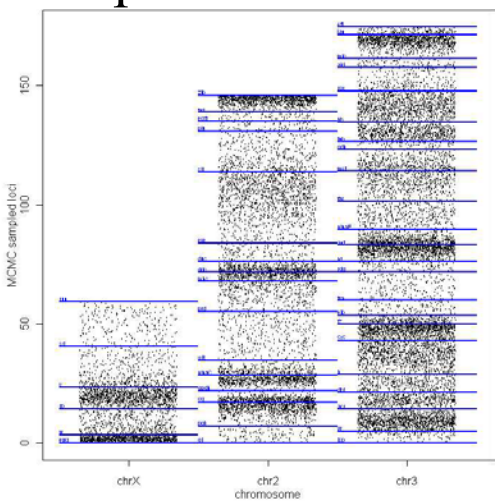
- *m* ~ Poisson(15) prior on number of QTL
- Bayesian LOD (log posterior density)
- Heritability

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



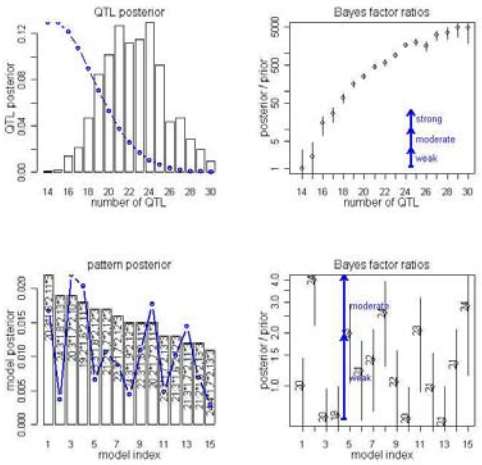
MCMC sampled loci

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



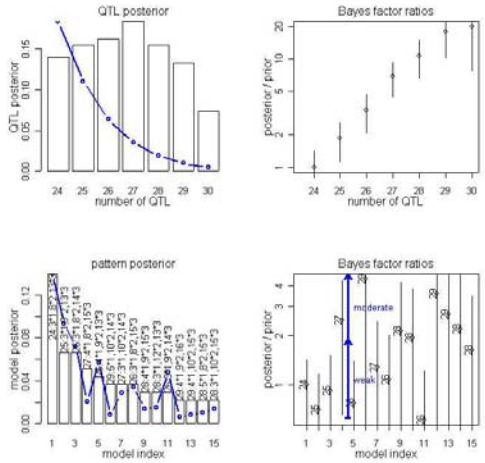
MCMC model selection

- 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



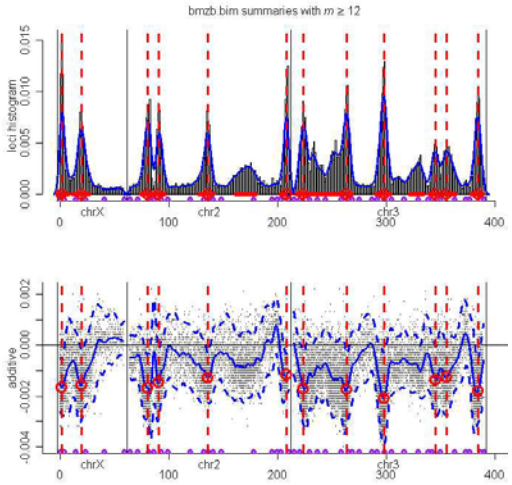
MCMC model selection restricted to "better models"

- 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

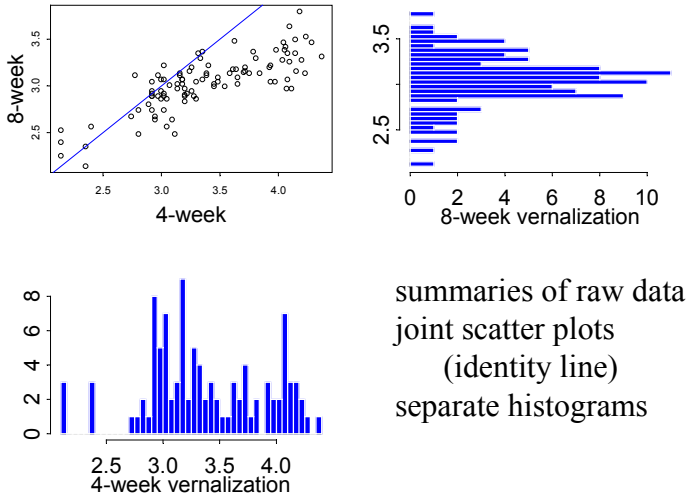
- model averaging
  - over all models
  - 1000 samples
- histogram of loci
  - marginal posteriors
  - superimposed on genome
  - 12 peaks identified
- scatterplot: loci & effects
  - smoothed mean ± 2 SE



Brassica napus data

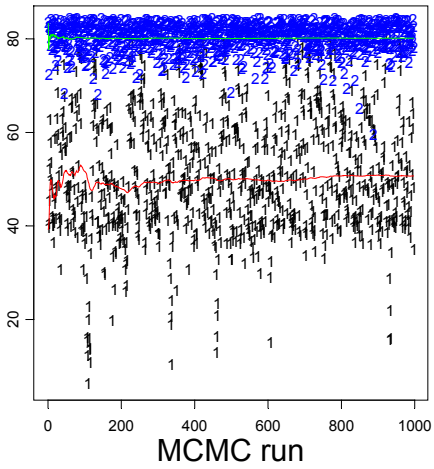
- 4-week & 8-week vernalization effect
  - log(days to flower)
- genetic cross of
  - Stellar (annual canola)
  - Major (biennial rapeseed)
- 105 F1-derived double haploid (DH) lines
  - homozygous at every locus (*QQ* or *qq*)
- 10 molecular markers (RFLPs) on LG9
  - two QTLs inferred on LG9 (now chromosome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

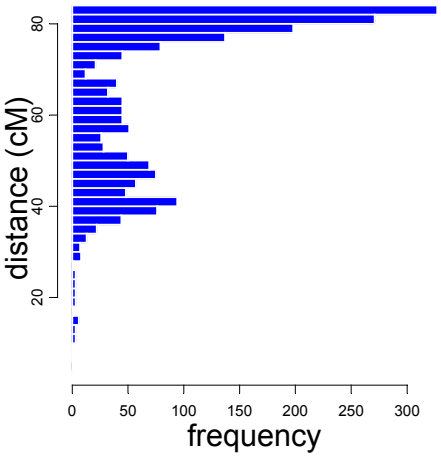
Brassica 4- & 8-week data



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

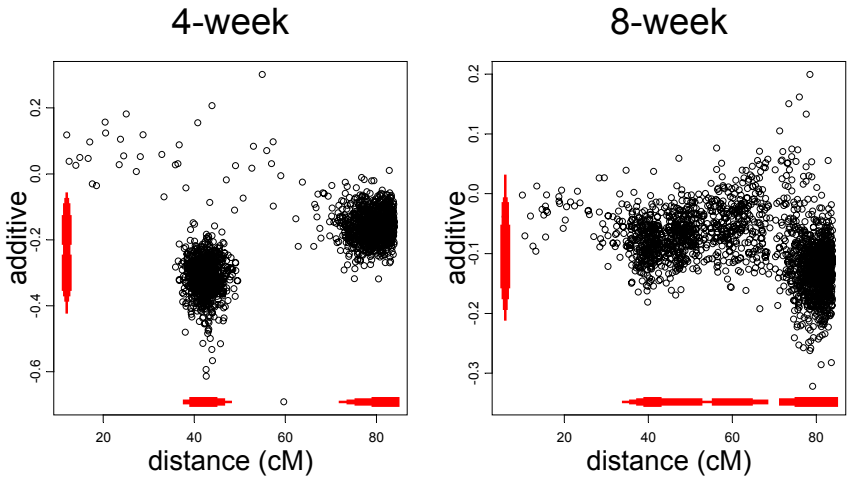
cM	add
40	.30
80	.16

8-week vernalization

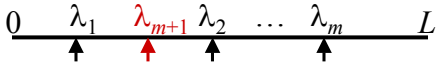
- shorter time to flower
- larger LOD at 80cM
- modest LOD at 40cM
- loci poorly determined

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

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

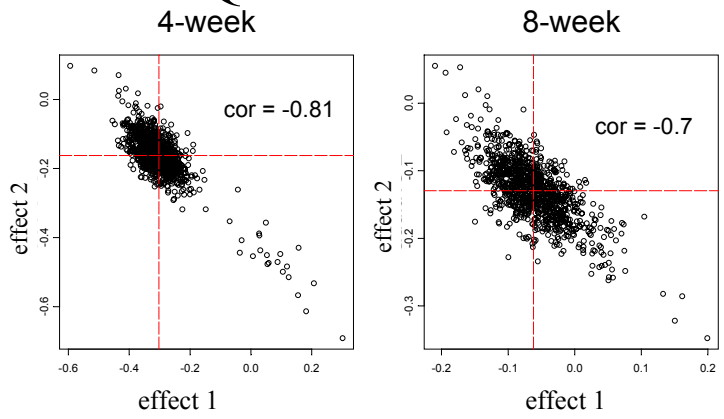
model selection in regression

- 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

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

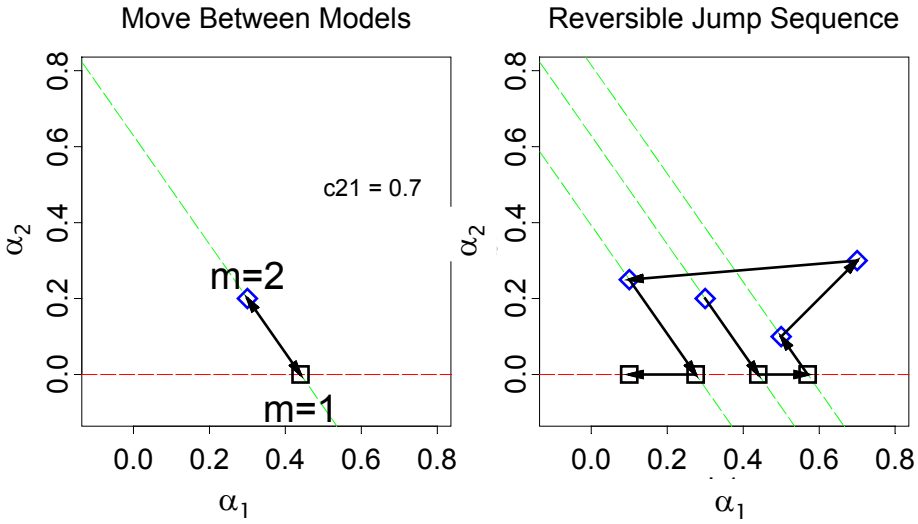
$$m = 2 : Y_i = \mu + \alpha_1(Q_{i1} - \overline{Q}_1) + \alpha_2(Q_{i1} - \overline{Q}_1) + e_i$$

collinear QTL = correlated effects

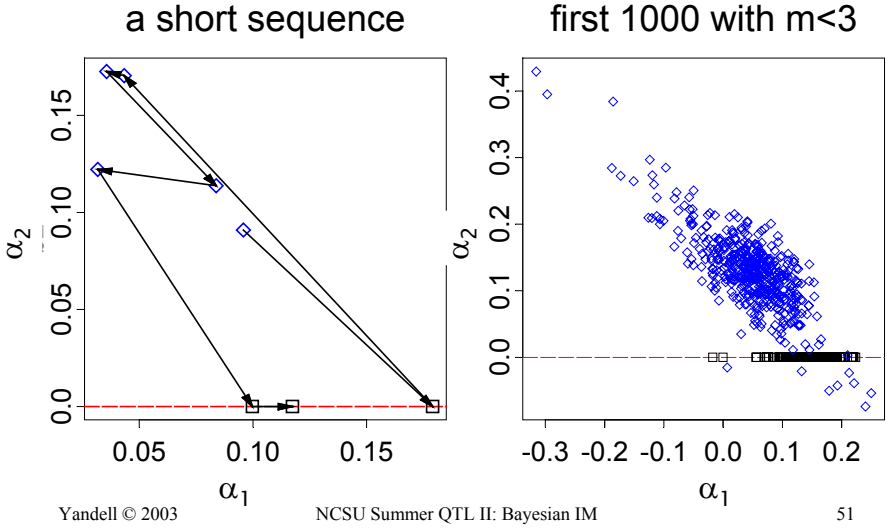


- linked QTL: collinear genotypes & correlated effect estimates
  - –sum of linked effects usually well determined
- which QTL to go after in breeding, genome walking?

Geometry of Reversible Jump



QT `additive` Reversible Jump



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

Multiple Traits & Microarrays

- why study multiple traits together?
  - diabetes case study
  - central dogma via microarrays
- why are traits correlated?
  - close linkage or pleiotropy?
- how to handle high throughput?
  - dimension reduction: multivariate stats
  - principal components on phenotypes

1 why study multiple traits together?

- avoid reductionist approach to biology
  - address physiological/biochemical mechanisms
  - Schmalhausen (1942); Falconer (1952)
- separate close linkage from pleiotropy
  - 1 locus or 2 linked loci?
- identify epistatic interaction or canalization
  - influence of genetic background
- establish QTL x environment interactions
- decompose genetic correlation among traits
- increase power to detect QTL

how to map multiple traits?

- WinQTL/QTL Cartographer: IM & CIM
  - Jiang Zeng (1995); Vieira et al. (2000)
  - statgen.ncsu.edu/qtlcart
- MultiQTL: 1-2 QTL with PC on residuals
  - Korol et al. (2001)
  - www.multiqtl.com
- QTL Express: Haley-Knott regression
  - Knott Haley (2000)
  - qtl.cap.ed.ac.uk
- SOLAR: outbred pedigrees
  - Almasy Blangero (1997); Williams et al. (1999)

Type 2 Diabetes mellitus

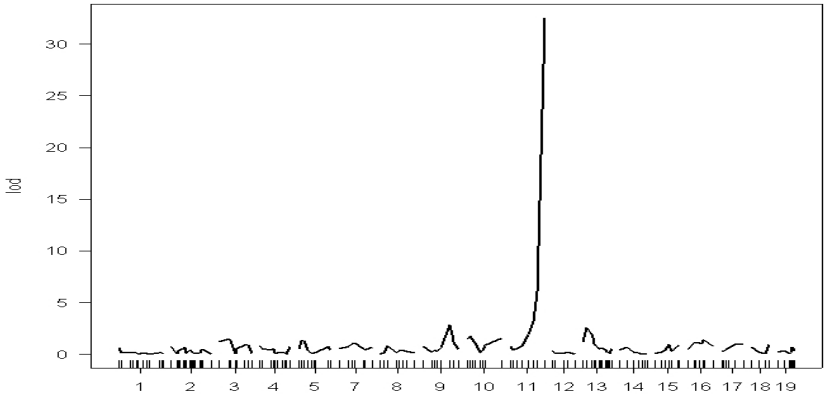
insulin requirement

insulin resistant mice

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,…

LOD map for PDI: *cis*-regulation



why map gene expression as a quantitative trait?

- *cis*- or *trans*-action?
  - does gene control its own expression?
  - evidence for both modes (Brem et al. 2002 *Science*)
- mechanics of gene expression mapping
  - measure gene expression in intercross (F2) population
  - map expression as quantitative trait (QTL technology)
  - adjust for multiple testing via false discovery rate
- research groups working on expression QTLs
  - review by Cheung and Spielman (2002 *Nat Gen Suppl*)
  - Kruglyak (Brem et al. 2002 *Science*)
  - Doerge et al. (Purdue); Jansen et al. (Waginingen)
  - Williams et al. (U KY); Lusis et al. (UCLA) (Schadt et al. 2003 *Nature*)
  - Dumas et al. (2000 *J Hypertension*)

mapping microarray data

- overview, wish lists
  - Jansen, Nap (2001 *Trends Gen*); Cheung, Spielman (2002 *Nat Gen Suppl*); Doerge (2002 *Nat Rev Gen*); Bochner (2003 *Nat Rev Gen*)
- single gene expression as trait (single QTL)
  - Dumas et al. (2000 *J Hypertens*)
- microarray scan via 1 QTL interval mapping
  - Brem et al. (2002 *Science*); Schadt et al. (2003 *Nature*)
  - found *cis* and *trans* acting genes
- multivariate and multiple QTL approach
  - Lan et al. (2003 *Genetics*)

central dogma via microarrays
Bochner (2003 *Nat Rev Gen*)

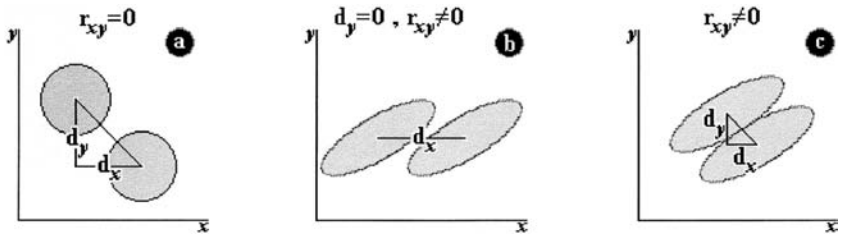
idea of mapping microarrays
(Jansen, Nap 2001 *Trends Gen*)

goal: unravel biochemical pathways (Jansen, Nap 2001 *Trends Gen*)

2 why are traits correlated?

- environmental correlation
  - non-genetic, controllable by design
  - historical correlation (learned behavior)
  - physiological correlation (same body)
- genetic correlation
  - pleiotropy
    - one gene, many functions
    - common biochemical pathway, splicing variants
  - close linkage
    - two tightly linked genes
    - genotypes *Q* are collinear

interplay of pleiotropy & correlation



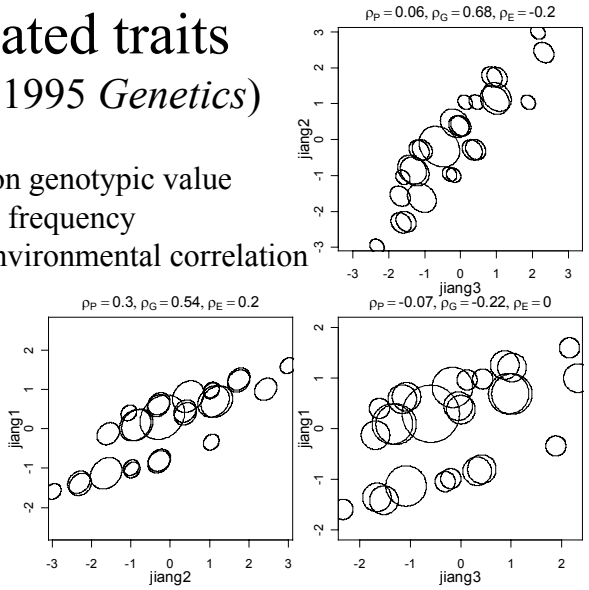
pleiotropy only correlation only both

Korol et al. (2001 *Genetics*)

3 correlated traits
(Jiang Zeng 1995 *Genetics*)

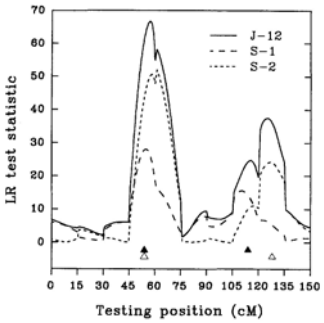
ellipses centered on genotypic value
width for nominal frequency
main axis angle environmental correlation
3 QTL, F2
27 genotypes

note signs of genetic and environmental correlation



pleiotropy or close linkage?

2 traits, 2 qtl/trait
pleiotropy @ 54cM
linkage @ 114,128cM
Jiang, Zeng (1995 *Genetics*)



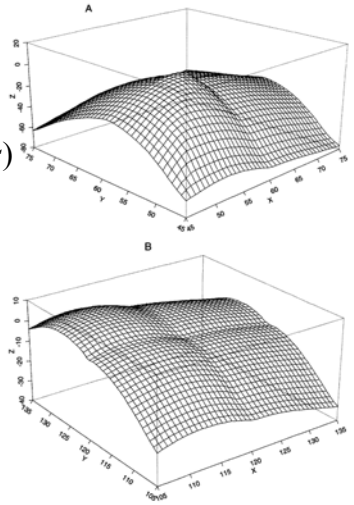
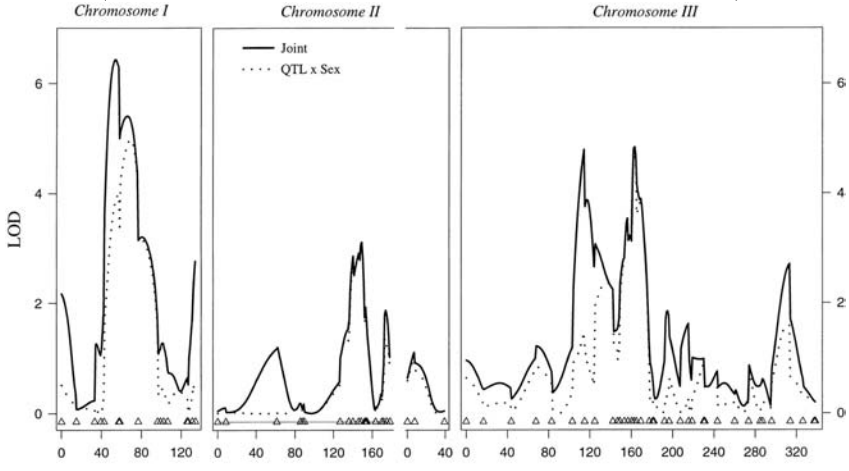


FIGURE 2.—Two-dimensional log-likelihood surfaces (expressed as deviations from the maximum of the log-likelihoods on the diagonal) for the test of pleiotropy *vs.* close linkage are presented for two regions: the region between 45 and 75 cM of Figure 1 (A) and the region between 105 and 135 cM (B). *X* is the testing position for a QTL affecting trait 1 and *Y* is the testing position for a QTL affecting trait 2. On the diagonal of *X*-*Y* plane, two QTL are located in the same position and statistically are treated as one pleiotropic QTL. *Z* is the likelihood ratio test statistic scaled to zero at the maximum point of the diagonal.

QTL x sex interaction
(Vieira et al. 2000 *Genetics*)



high throughput dilemma

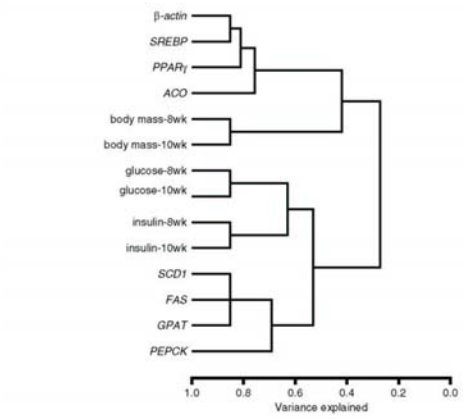
- want to focus on gene expression network
  - ideally capture pathway in a few dimensions
  - allow for complicated genetic architecture
- may have multiple controlling loci
  - could affect many genes in coordinated fashion
  - could show evidence of epistasis
  - quick assessment via interval mapping may be misleading
    - Brem et al. (2002 *Science*); Schadt et al. (2003 *Nature*)
- try mapping principle components as super-traits
  - capture key multivariate features of multiple traits
  - elicit biochemical pathways
    - Henderson et al. Hoeschele (2001); Ong Page (2002)

coordinated gene expression

- Brem et al. (2002 *Science*)
  - pleiotropy in yeast genome
- Schadt et al. (2003 *Nature*)
  - coordinated expression in mouse genome

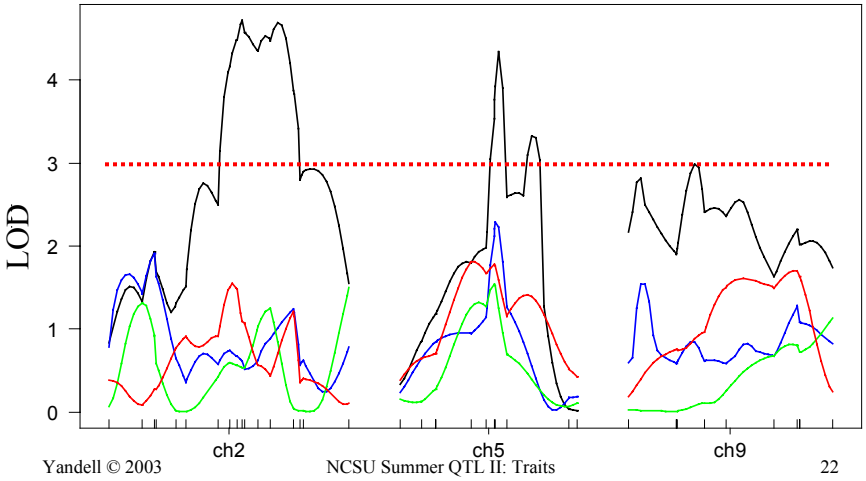
high throughput: which genes are the key players?

Lan et al., mapping mRNA, Figure 2 (2003 *Genetics*)



- clustering of expression seed by insulin, glucose
- advantage: subset relevant to trait
- disadvantage: still many genes to study

SCD1, FAS,GPAT, PEPCK: *trans*-regulation by multiple QTL?



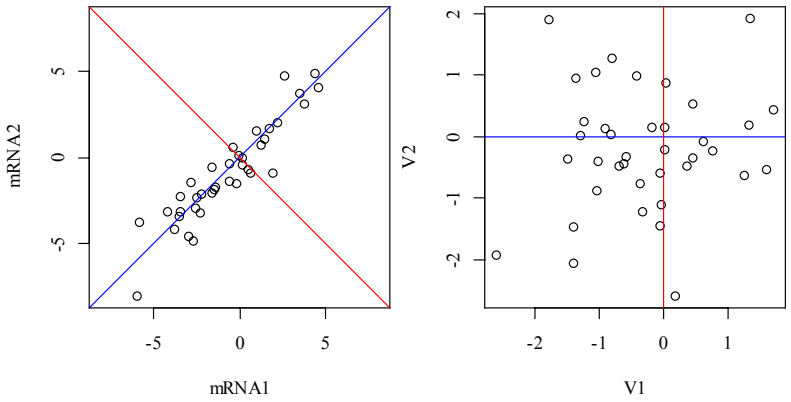
from gene expression to super-genes

- PC or SVD decomposition of multiple traits
  - $Y = t$ traits $\times$ n individuals
  - decompose as $Y = UDW^{T}$
    - U,W = ortho-normal transforms (eigen-vectors)
    - D = diagonal matrix with singular values
- transform problem to principal components
  - W_1 and W_2 uncorrelated "super-traits"
- interval map each PC separately
  - $W_1 = \mu^*_1 + G^*_{1Q} + e^*_1$
- may only need to map a few PCs

Alter et al. (2000 *PNAS*)

- yeast cell cycle
- singular value decomposition
  - graphical display of decomposition
  - see supplement to *PNAS* article

PC simply rotates & rescales to find major axes of variation



PC summary of shape phenotype

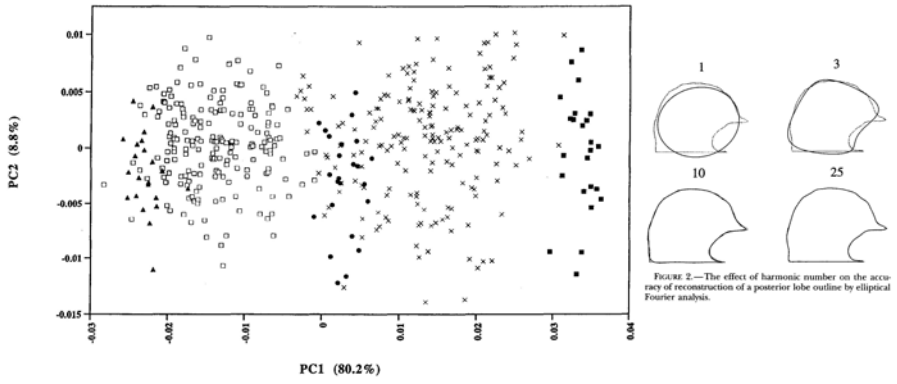


FIGURE 2.—The effect of harmonic number on the accuracy of reconstruction of a posterior lobe outline by elliptical Fourier analysis.

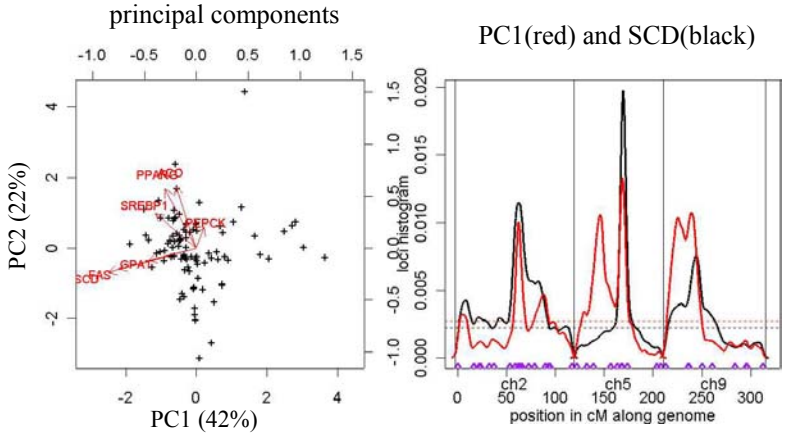
FIGURE 5.—A plot of the first two principal components of the Fourier coefficients from posterior lobe outlines. Many individuals from each of five genotypic classes are represented. Each point represents an average of scores from the left and right sides of an individual (with a few exceptions for which 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*

QTL via Principal Components

- *Drosophila* gonad shape
  - Liu et al. (1996); Zeng et al. (2000)
- other refs of interest
  - Weller et al. (1996); Mangin et al. (1998); Olson et al. (1999); Mahler et al. (2002)
- problems
  - PC may have no relation to genetics!
  - residuals from QTL correlated across PCs
  - PC is descriptive summary, not interpretive

multivariate screen for gene expressing mapping

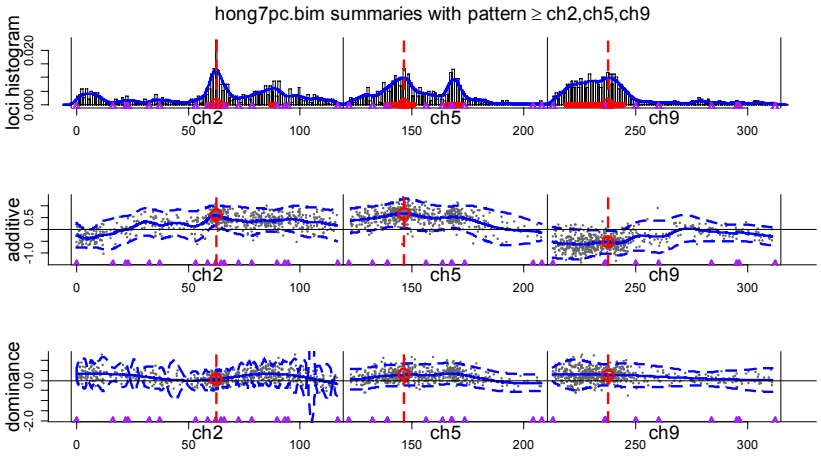
(Lan et al. 2003 *Genetics*)



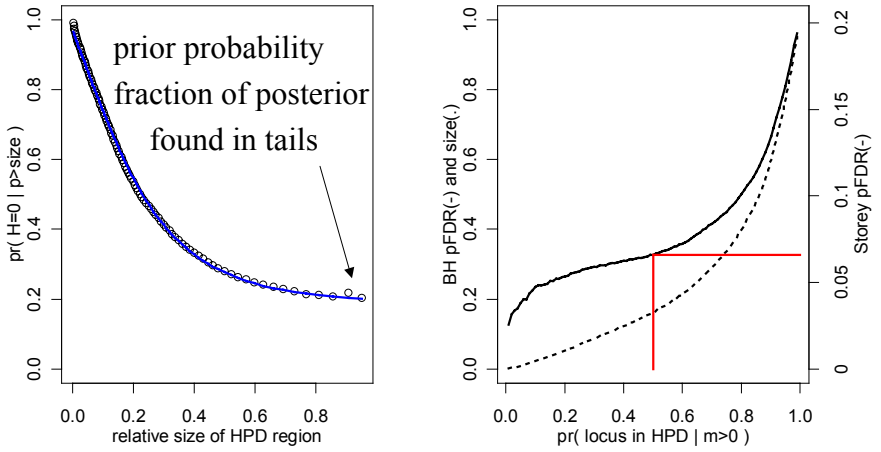
Relation of Composite Phenotypes to Individual mRNA Expressions

(West et al. 2001 *PNAS*)

mapping first diabetes PC as a trait



pFDR for PC1 analysis



improvements on PC?

- what is our goal?
  - reduce dimensionality
  - focus on QTL
- PC reduces dimensionality
  - but may not relate to genetics
- canonical discriminant analysis
  - rotate to improve discrimination
  - but need to know QTL first!
- open research area!