NCSU Summer Institute 2004
QTL II
Brian S. Yandell
University of Wisconsin-Madison

- Model: selection for multiple QTL
- Pheno: extensions beyond normal data
- Bayes: interval mapping with prior info
- Traits: multiple phenotypes & microarrays

contact information & resources

- email: byandell@wisc.edu
- web: www.stat.wisc.edu/~yandell/statgen
  - QTL & microarray resources
  - references, software, people

- thanks:
  - students: Chunfang “Amy” Jin, Fei Zou, Pat Gaffney, Jaya Satagopan
  - faculty/staff: Alan Attie, Hong Lan, Michael Newton, Christina Kendziorski, Tom Osborn, Jason Fine

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

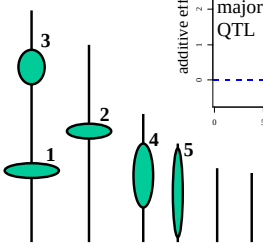
- evaluate some objective for model given data
  - classical likelihood
  - Bayesian posterior
- search over possible genetic architectures (models)
  - number and positions of loci
  - gene action: additive, dominance, epistasis
- estimate "features" of model
  - means, variances & covariances, confidence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

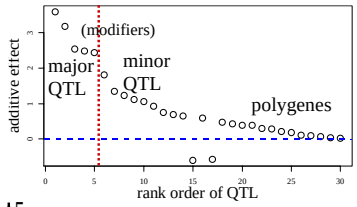
advantages of multiple QTL approach

- 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

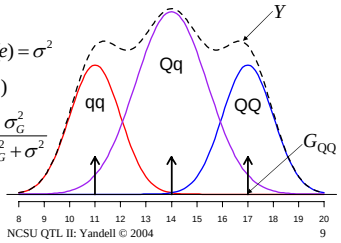
- trait = mean + genetic + environment
- genetic effect uncorrelated with environment
- pr(trait Y | genotype Q, effects θ)

$$Y = G_Q + e$$

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

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

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



two QTL with epistasis

- same phenotype model overview

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

- partition of genotypic value with epistasis

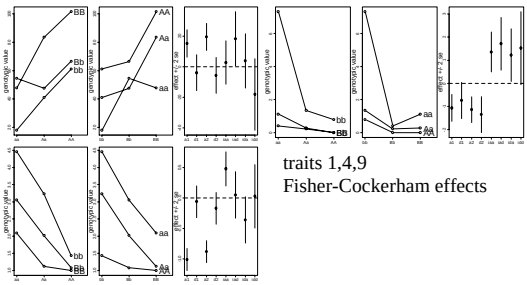
$$G_Q = \mu + \beta_1(Q) + \beta_2(Q) + \beta_{12}(Q)$$

- 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,…}

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

- 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

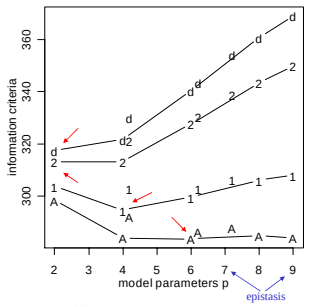
- LOD or likelihood ratio compares model
  - $L(p)$ = log likelihood for a particular model with p parameters
  - $\log(LR) = L(p_2) - L(p_1)$
  - $LOD = \log_{10}(LR) = \log(LR)/\log(10)$
- p = number of model degrees of freedom
  - consider models with m QTL and all 2-QTL epistasis terms
  - BC: $p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

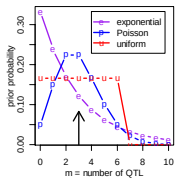
$$B_{12} = \frac{\text{pr}(\text{model}_1 \mid Y)/\text{pr}(\text{model}_2 \mid Y)}{\text{pr}(\text{model}_1)/\text{pr}(\text{model}_2)} = \frac{\text{pr}(Y \mid \text{model}_1)}{\text{pr}(Y \mid \text{model}_2)}$$

- 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
  pr($Y|Q,\theta$) ~ $N(G_Q, \sigma^2)$, 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 ~ $N(\mu_0, \kappa_0\sigma^2)$ grand mean
  $\beta(Q)$ ~ $N(0, \kappa_1\sigma^2)$ model-independent genotypic effect
  $\beta_j(Q)$ ~ $N(0, \kappa_1\sigma^2/|M|)$ effects down-weighted by size of M
- determine hyper-parameters via Empirical Bayes

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

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

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

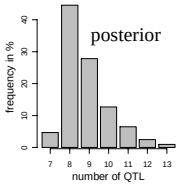
LS estimate $\hat{G}_Q = \text{sum}_i[\text{sum}_{j\in M}\hat{\beta}_j(Q_i)] = \text{sum}_i w_{iQ} Y$

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

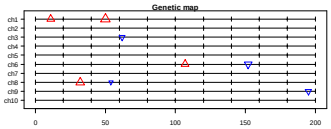
shrinkage $B_Q = \kappa/(\kappa + C_Q) \to 1$

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

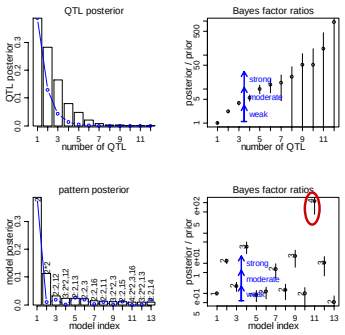
- 108 plants from double haploid
  - similar genetics to backcross: follow 1 gamete
  - parents are Major (biennial) and Stellar (annual)
- 300 markers across genome
  - 19 chromosomes
  - average 6cM between markers
    - median 3.8cM, max 34cM
  - 83% markers genotyped
- phenotype is days to flowering
  - 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

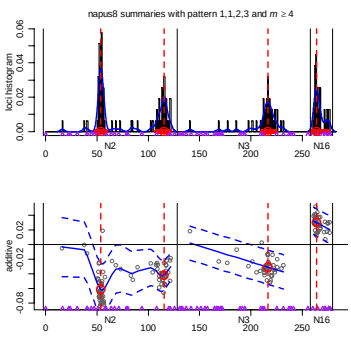
evidence suggests
4-5 QTL
N2(2-3),N3,N16



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

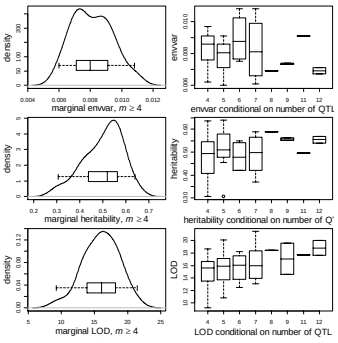


Bayesian model diagnostics

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)

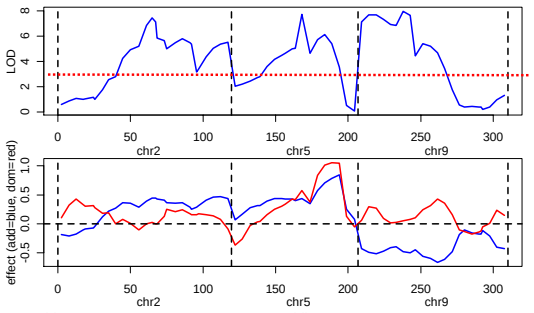
but note change with *m*



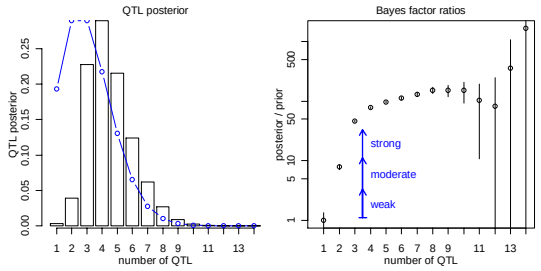
studying diabetes in an F2

- segregating cross of inbred lines
  - B6.ob x BTBR.ob → 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,…

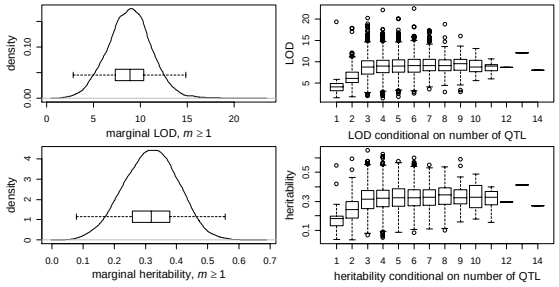
Multiple Interval Mapping
SCD1: multiple QTL plus epistasis!



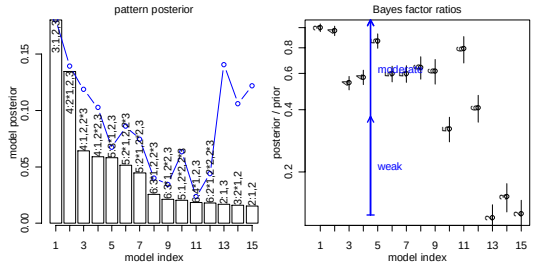
Bayesian model assessment: number of QTL for SCD1



Bayesian LOD and h^2 for SCD1

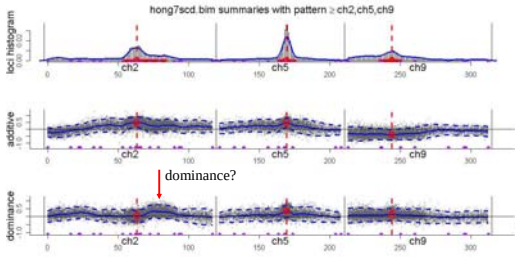


Bayesian model assessment: chromosome QTL pattern for SCD1

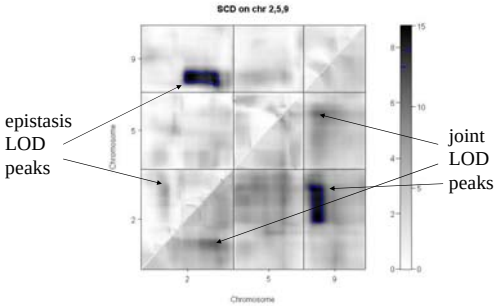


trans-acting QTL for SCD1

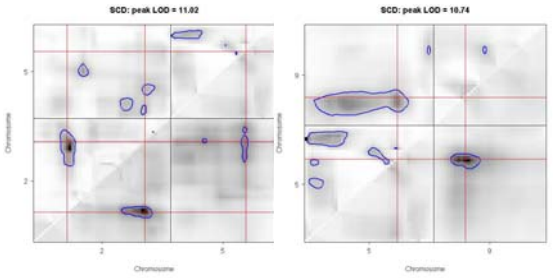
(no epistasis yet: see Yi, Xu, Allison 2003)



2-D scan: assumes only 2 QTL!

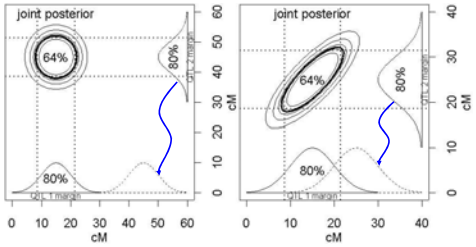


sub-peaks can be easily overlooked!



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

unlinked loci linked loci



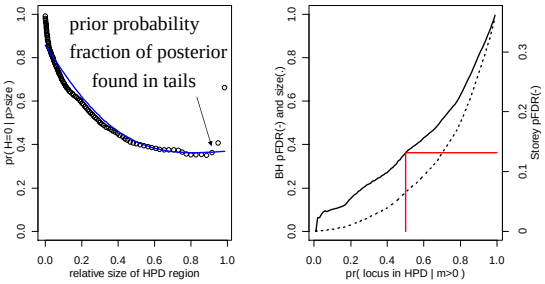
false detection rates and thresholds

- 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



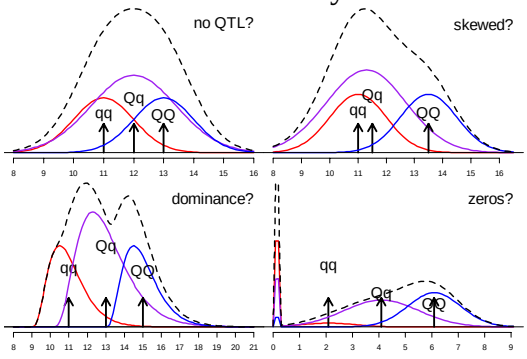
Extending the Phenotype Model

- bottom line for normal phenotype model
  - may work well to pick up loci
  - may be poor at estimating effects if data not normal

1. 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

large sample size
dense vs. sparse markers

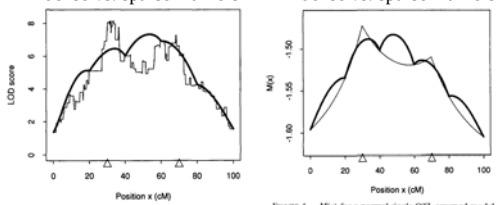


FIGURE 1.—The two-QTL true model with a QTL at 30 cM and a second QTL of somewhat smaller effect at 70 cM (true locations indicated by △). A normal single-QTL model is assumed and the LOD score for 100 simulated individuals is given for dense markers (thin curve) and markers at 20-cM intervals (bold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

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 50 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?

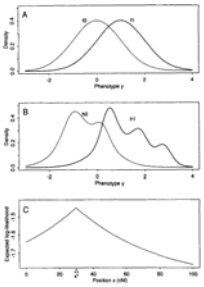


FIGURE 2.—Misspecification of the phenotype model. (A) The assumed distributions f_0 and f_1. (B) The true distributions h_0, h_1. (C) The expected log-likelihood across the chromosome when the markers are dense. Despite the misspecification, the function is maximized at exactly the true location $a^* = 50$ 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,…,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

2. semi-parametric QTL

- phenotype model 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 pr($Y|Q,\theta$) = $f(Y)$
- includes many standard phenotype models

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

  Poisson pr($Y|Q,\theta$) = 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 ~ binomial(1,π), 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
  - $\text{pr}(Y = j) = \text{pr}(\tau_{j-1} < Z \le \tau_j)$
  - $\text{pr}(Y \le j) = \text{pr}(Z \le \tau_j)$
- 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)

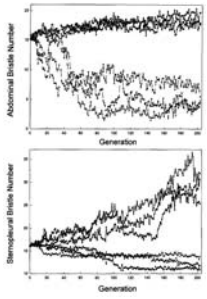
$$\text{pr}(Y \le j \mid Q) = \text{pr}(Z \le \tau_j \mid Q) = [1 + \exp(\mu + G_Q - \tau_j)]^{-1}$$

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

- partial empirical LOD
  - Zou, Fine, Yandell (2002 *Biometrika*)
- conditional empirical LOD
  - Zou, Fine (2003 *Biometrika*); Jin, Fine, Yandell (2004)
- 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)

partial empirical likelihood

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

now profile with respect to β,λ

$$\log(L(\beta,\lambda|Y,X)) = \text{sum}_i \log(f_i) + \log(w_i) + \text{sum}_Q\,\alpha_Q(1 - \text{sum}_i\, f_i \exp(Y_i\beta_Q))$$

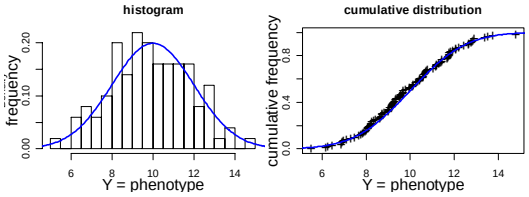
partial likelihood: set Lagrange multipliers α_Q to 0

force f to be a distribution that sums to 1

point mass density estimates

$$f_i = (\text{sum}_i w_i)^{-1} \text{ with } w_i = \text{sum}_Q \exp(Y_i\beta_Q)\text{pr}(Q \mid X_i, \lambda)$$

histograms and CDFs

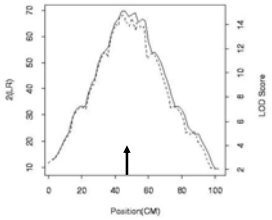


histograms capture shape
but are not very accurate

CDFs are more accurate
but not always intuitive

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

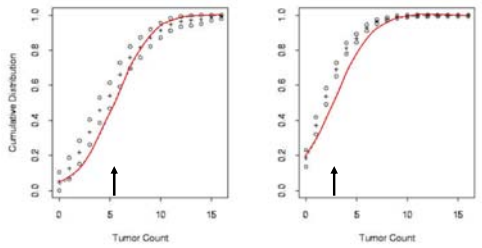
- rat backcross
  - two inbred strains
    - Wistar-Furth susceptible
    - Wistar-Kyoto resistant
  - backcross to WF
  - 383 females
  - chromosome 5, 58 markers
- search for resistance genes
- *Y* = # mammary carcinomas
- where is the QTL?



dash = normal
solid = semi-parametric

what shape histograms by genotype?

WF/WF WKy/WF



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

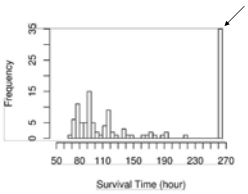
conditional empirical LOD

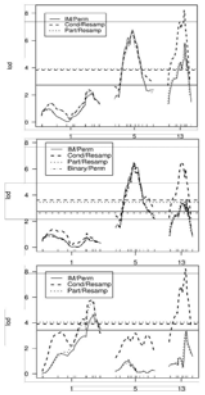
- partial empirical LOD has problems
  - tests for F2 depends on unknown weights
  - difficult to generalize to multiple QTL
- conditional empirical likelihood unbiased
  - examine genotypes given phenotypes
  - does not depend on *f*(*Y*)
  - pr(X_i) depends only on mating design
  - unbiased for selective genotyping (Jin et al. 2004)

$$\text{pr}(X_i|Y_i,\theta,\lambda,Q) = \exp(Y_i\beta_Q)\ \text{pr}(Q|X_i|\lambda)\ \text{pr}(X_i) \,/\, \text{constant}$$

spike data example

Boyartchuk et al. (2001); Broman (2003)
133 markers, 20 chromosomes
116 female mice
Listeria monocytogenes infection





new resampling threshold method

- EM locally approximates LOD by quadratic form
- use local covariance of β estimates to further approximate
  - relies on *n* independent standard normal variates $Z = (Z_1,\ldots,Z_n)$
  - one set of variates *Z* for the entire genome!
- repeatedly resample independent standard normal variates *Z*
  - no need to recompute maximum likelihood on new samples
  - intermediate EM calculations used directly
- evaluate threshold as with usual permutation test
  - extends naturally to multiple QTL
- results shown in previous figure

$$LOD(\lambda) \approx n\hat{\beta}^{\mathrm{T}}(\lambda)S(\lambda)\hat{\beta}(\lambda) \approx Z^{\mathrm{T}}C^{\mathrm{T}}(\lambda)S(\lambda)C(\lambda)Z$$
$$\text{cov}\left(\sqrt{n}\hat{\beta}(\lambda)\right) = -C^{\mathrm{T}}(\lambda)C(\lambda)$$
$$\sqrt{n}\hat{\beta}(\lambda) \approx C(\lambda)Z, \text{ with } Z \sim N(0,I)$$

3. non-parametric methods

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

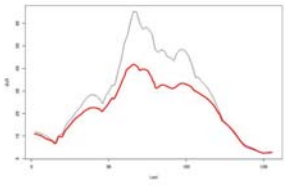
  $\text{sum}_i\ [I(Y_i \le y) - \text{sum}_Q\ \text{pr}(Q|X,\lambda)F_Q(y)]^2$
- looks complicated, but simple to implement

non-parametric CDF properties

- readily extended to censored data
  - time to flowering for non-vernalized plants
  - Fine, Zou, Yandell (2004 *Biometrics J)*
- 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

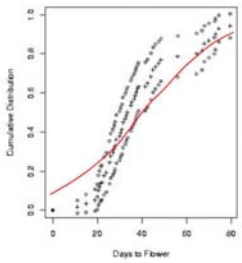
- *Brassica napus*
  - Major female
    - needs vernalization
  - Stellar male
    - insensitive
  - 99 double haploids
- Y = log(days to flower)
  - over 50% Major at QTL never flowered
  - log not fully effective



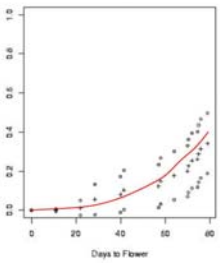
grey = normal, red = non-parametric

what shape is flowering distribution?

B. napus Stellar



B. napus Major

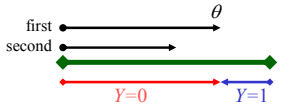


line = normal, + = non-parametric, o = confidence interval

Bayesian Interval Mapping

1. who was Bayes? what is Bayes theorem?

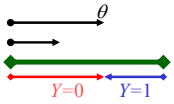
- Reverend Thomas Bayes (1702-1761)
  - part-time mathematician
  - buried in Bunhill Cemetary, Moongate, London
  - famous paper in 1763 *Phil Trans Roy Soc London*
  - was Bayes the first with this idea? (Laplace?)
- basic idea (from Bayes' original example)
  - two billiard balls tossed at random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



prior $\quad \text{pr}(\theta) = 1$
likelihood $\quad \text{pr}(Y|\theta) = \theta^{1-Y}(1-\theta)^{Y}$
posterior $\quad \text{pr}(\theta|Y) = ?$

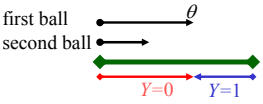
what is Bayes theorem?

- where is first ball if the second is to its left (right)?
- prior: probability of parameter before observing data
  - pr(θ) = pr(parameter)
  - equal chance of being anywhere on the table
- posterior: probability of parameter after observing data
  - pr(θ | Y) = pr(parameter | data)
  - more likely to left if first ball is toward the right end of table
- likelihood: probability of data given parameters
  - pr(Y | θ) = pr(data | parameter)
  - basis for classical statistical inference
- Bayes theorem
  - posterior = likelihood * prior / pr(data)
  - normalizing constant pr(Y) often drops out of calculation

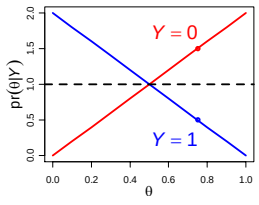


$$\text{pr}(\theta|Y) = \frac{\text{pr}(\theta, Y)}{\text{pr}(Y)} = \frac{\text{pr}(Y|\theta) \times \text{pr}(\theta)}{\text{pr}(Y)}$$

where is the second ball given the first?



prior $\quad \text{pr}(\theta) = 1$
likelihood $\quad \text{pr}(Y|\theta) = \theta^{1-Y}(1-\theta)^{Y}$
posterior $\quad \text{pr}(\theta|Y) = ?$



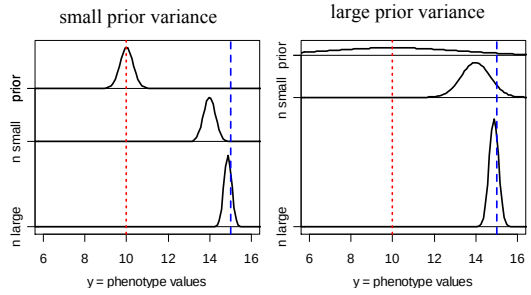
$\text{prior} : \text{pr}(\theta) = 1$

$\text{likelihood} : \text{pr}(Y|\theta) = \begin{cases} \theta & \text{if } Y = 0 \\ 1-\theta & \text{if } Y = 1 \end{cases}$

$\text{marginal} : \text{pr}(Y) = \frac{1}{2}$

$\text{posterior} : \text{pr}(\theta|Y) = \frac{\text{pr}(Y|\theta)\text{pr}(\theta)}{\text{pr}(Y)} = \begin{cases} 2\theta & \text{if } Y = 0 \\ 2(1-\theta) & \text{if } Y = 1 \end{cases}$

Bayes posterior for normal data



Bayes posterior for normal data

model	$Y_i = \mu + E_i$
environment	$E \sim N(0, \sigma^2)$, σ^2 known
likelihood	$Y \sim N(\mu, \sigma^2)$
prior	$\mu \sim N(\mu_0, \kappa\sigma^2)$, κ known
posterior:	mean tends to sample mean
single individual	$\mu \sim N(\mu_0 + B_1(Y_1 - \mu_0), B_1\sigma^2)$
sample of n individuals	$\mu \sim N\left(B_n\bar{Y}_\bullet + (1-B_n)\mu_0, B_n\frac{\sigma^2}{n}\right)$
	with $\bar{Y}_\bullet = \text{sum}\frac{Y_i}{n}$
fudge factor (shrinks to 1)	$B_n = \frac{\kappa n}{\kappa n + 1} \to 1$

2. Bayesian inference for QTL

- develop priors on unknowns
  - unknowns:
    - missing genotypes Q
    - effects $\theta = (G_Q, \sigma^2)$
    - loci λ (see next section)
  - use empirical Bayes to set useful priors
- study posterior for unknowns given data
  - data:
    - phenotypes Y
    - markers & linkage map X
  - marginal posteriors for effects θ, loci λ

Bayesian priors for QTL

- missing genotypes Q
  - $\text{pr}(Q \mid X, \lambda)$
  - recombination model is formally a prior
- effects $\theta = (G_Q, \sigma^2)$
  - $\text{pr}(\theta) = \text{pr}(G_Q \mid \sigma^2)\,\text{pr}(\sigma^2)$
  - use conjugate priors for normal phenotype
    - $\text{pr}(G_Q \mid \sigma^2)$ = normal
    - $\text{pr}(\sigma^2)$ = inverse chi-square
- each locus λ may be uniform over genome
  - $\text{pr}(\lambda \mid X)$ = 1 / length of genome
- combined prior
  - $\text{pr}(Q, \theta, \lambda \mid X) = \text{pr}(Q \mid X, \lambda)\,\text{pr}(\theta)\,\text{pr}(\lambda \mid X)$

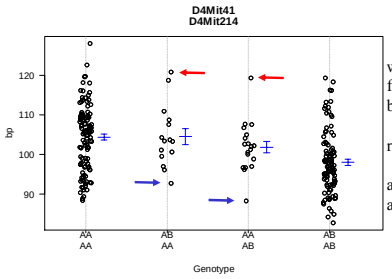
Bayesian model posterior

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

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

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

how does phenotype Y improve posterior for genotype Q?



what are probabilities for genotype Q between markers?

recombinants AA:AB

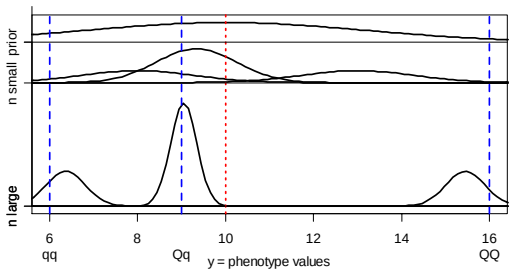
all 1:1 if ignore Y and if we use Y?

posterior on QTL genotypes

- full conditional of Q given data, parameters
  - proportional to prior $\text{pr}(Q \mid X, \lambda)$
    - 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

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

posterior genotypic means G_Q



genetic effect posterior given Q

posterior centered on sample genotypic mean
but shrunken slightly toward overall mean

prior: $G_Q \sim N\left(\overline{Y}_\bullet, \kappa\sigma^2\right)$

posterior: $G_Q \sim N\left(B_Q\overline{Y}_Q + (1-B_Q)\overline{Y}_\bullet, B_Q\frac{\sigma^2}{n_Q}\right)$

$$n_Q = \text{count}\{Q_i = Q\}, \overline{Y}_Q = \underset{\{i:Q_i=Q\}}{\text{sum}} \frac{Y_i}{n_Q}$$

fudge factor: $B_Q = \frac{\kappa n_Q}{\kappa n_Q + 1} \to 1$

What if variance σ^2 is unknown?

- sample variance is proportional to chi-square
  - $ns^2/\sigma^2 \sim \chi^2(n)$
  - likelihood of sample variance s^2 given n, σ^2
- conjugate prior is inverse chi-square
  - $\nu\tau^2/\sigma^2 \sim \chi^2(\nu)$
  - prior of population variance σ^2 given ν, τ^2
- posterior is weighted average of likelihood and prior
  - $(\nu\tau^2+ns^2)/\sigma^2 \sim \chi^2(\nu+n)$
  - posterior of population variance σ^2 given n, s^2, ν, τ^2
- empirical choice of hyper-parameters
  - $\tau^2 = s^2/3$, $\nu=6$
  - $E(\sigma^2 \mid \nu,\tau^2) = s^2/2$, $\text{Var}(\sigma^2 \mid \nu,\tau^2) = s^4/4$

3. Markov chain sampling of architectures

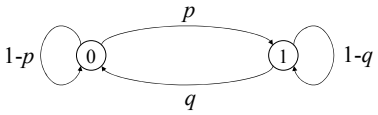
- construct Markov chain around posterior
  - want posterior as stable distribution of Markov chain
  - in practice, the chain tends toward stable distribution
    - initial values may have low posterior probability
    - burn-in period to get chain mixing well
- hard to sample (λ,Q,θ,m) from joint posterior
  - update (λ,Q,θ) from full conditionals for m-QTL model
  - update m using reversible jump technology

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

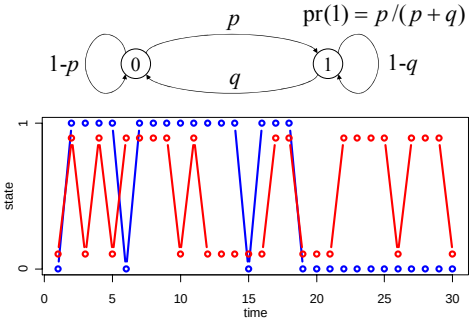
What is a Markov chain?

- 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
- toy problem
  - two states (0,1)
  - move chances depend on current state
  - what is the chance of being in state 1?

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



Markov chain idea

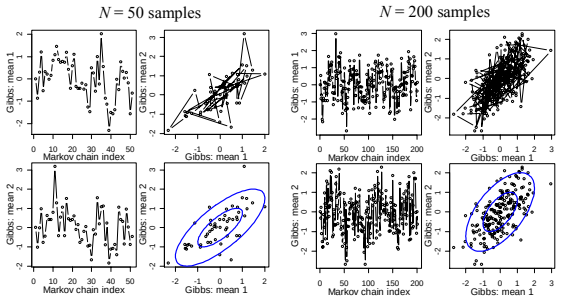


Gibbs sampler idea

- toy problem
  - want to study two correlated effects
  - could sample directly from their bivariate distribution
- instead use Gibbs sampler:
  - sample each effect from its full conditional given the other
  - pick order of sampling at random
  - repeat many times

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$
$$\theta_1 \sim N\left(\mu_1 + \rho(\theta_2 - \mu_2), 1-\rho^2\right)$$
$$\theta_2 \sim N\left(\mu_2 + \rho(\theta_1 - \mu_1), 1-\rho^2\right)$$

Gibbs sampler samples: $\rho = 0.6$



MCMC sampling of (λ, Q, θ)

- Gibbs sampler
  - effects $\theta = (G_Q, \sigma^2)$
  - genotypes Q
  - *not* loci λ

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

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

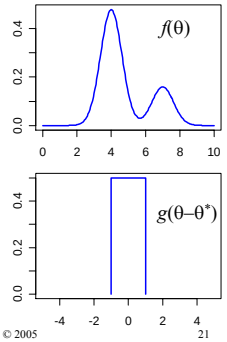
$$\theta \sim \frac{\text{pr}(Y \mid Q, \theta)\text{pr}(\theta)}{\text{pr}(Y \mid Q)}$$

- extension of Gibbs sampler
  - Metropolis-Hastings sampler
  - does not require normalization
  - loci λ: pr($Q \mid X$) difficult to compute

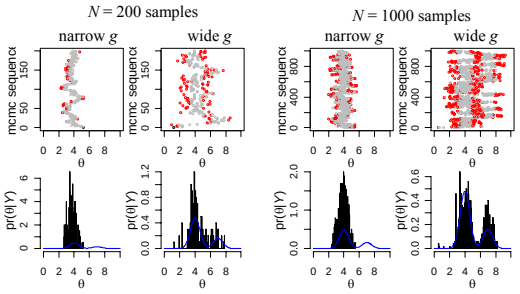
Metropolis-Hastings idea

- want to study distribution $f(\theta)$
  - take Monte Carlo samples
    - unless too complicated
  - take samples using ratios of f
- 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
  $\text{pr}(\lambda \mid Y, X, \theta, Q) = \text{pr}(\lambda \mid X, Q)$
  $= \text{pr}(\lambda)\ \text{pr}(Q \mid X, \lambda) / \text{constant}$
- to explicitly determine constant, must average
  - over all possible genotypes
  - over entire map
- Gibbs sampler will not work in general
  - but can use method based on ratios of probabilities
  - Metropolis-Hastings is extension of Gibbs sampler

Metropolis-Hastings Step

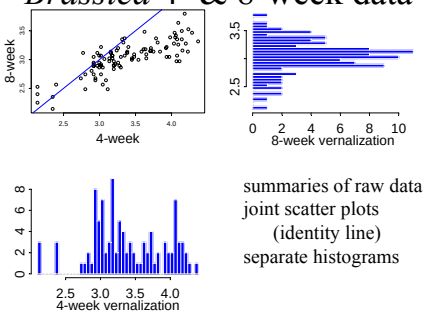
- pick new locus based upon current locus
  - propose new locus from some distribution $g(\ )$
    - pick value near current one? (usually)
    - pick uniformly across genome? (sometimes)
  - accept new locus with probability A
    - otherwise stick with current value

$$A(\lambda_{old}, \lambda_{new}) = \min\left(1, \frac{\text{pr}(\lambda_{new})\text{pr}(Q \mid X, \lambda_{new})g(\lambda_{new}, \lambda_{old})}{\text{pr}(\lambda_{old})\text{pr}(Q \mid X, \lambda_{old})g(\lambda_{old}, \lambda_{new})}\right)$$

Brassica napus data

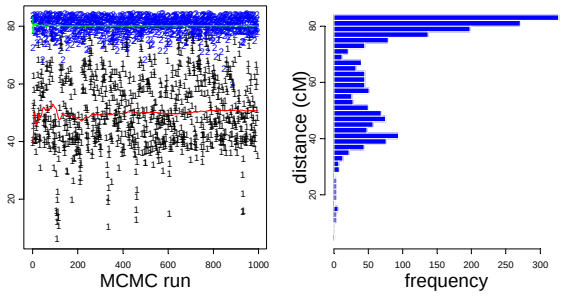
- 4-week & 8-week vernalization effect
  - log(days to flower)
- genetic cross of
  - Stellar (annual canola)
  - Major (biennial rapeseed)
- 105 F1-derived double haploid (DH) lines
  - homozygous at every locus (*QQ* or *qq*)
- 10 molecular markers (RFLPs) on LG9
  - two QTLs inferred on LG9 (now chromosome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

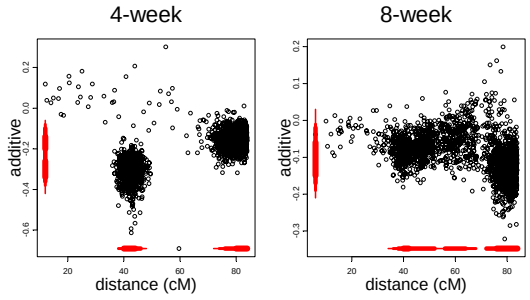
4-week vernalization
- longer time to flower
- larger LOD at 40cM
- modest LOD at 80cM
- loci well determined

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



4. sampling across architectures

- search across genetic architectures *M* of various sizes
  - allow change in *m* = number of QTL
  - allow change in types of epistatic interactions
- compare architectures
  - Bayes factors: previous talk
- methods for search
  - reversible jump MCMC
  - Gibbs sampler with loci indicators
- complexity of epistasis
  - Fisher-Cockerham effects model
  - general multi-QTL interaction & limits of inference

reversible jump issues

- use reversible jump MCMC to change m
  - adjust to change of variables between models
    - bookkeeping helps in comparing models
  - Green (1995); Richardson Green (1997)
- think model selection in multiple regression
  - but regressors (QTL genotypes) are unknown
  - linked loci = collinear regressors = correlated effects
  - consider only additive genetic effects here
    - genotype coding Q= –1, 0, 1 centered on average genotype

$$G(Q) = \mu + \beta(Q) \text{ with } \beta(Q) = \alpha \times (Q - \overline{Q})$$

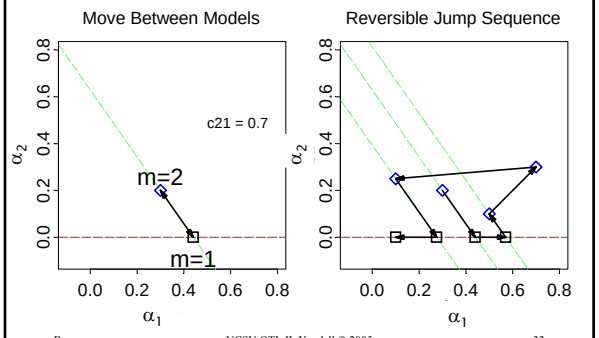
model selection in regression

- consider known genotypes Q at 2 known loci λ
  - models with 1 or 2 QTL
- jump between 1-QTL and 2-QTL models
  - adjust parameters when model changes
  - α and α_1 differ due to collinearity of QTL genotypes

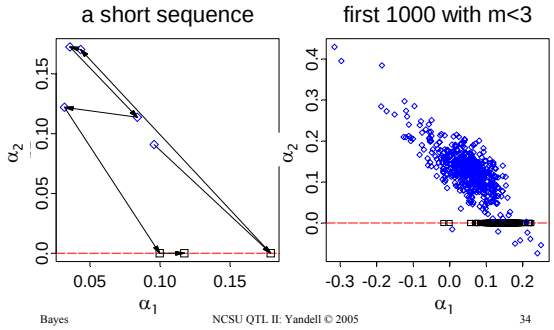
$$m = 1 : Y = \mu + \alpha(Q_1 - \overline{Q}_1) + e$$

$$m = 2 : Y = \mu + \alpha_1(Q_1 - \overline{Q}_1) + \alpha_2(Q_1 - \overline{Q}_1) + e$$

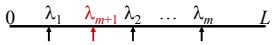
geometry of reversible jump



geometry allowing Q and λ to change



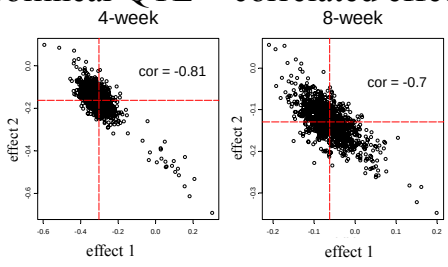
reversible jump MCMC



Metropolis-Hastings updates: 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

collinear QTL = correlated effects



- linked QTL = collinear genotypes
  - ➢ correlated estimates of effects (negative if in coupling phase)
  - ➢ sum of linked effects usually fairly constant

R/bim: our RJ-MCMC software

- R: www.r-project.org
  - freely available statistical computing application R
  - library(bim) builds on Broman's library(qtl)
- QTLCart: statgen.ncsu.edu/qtlcart
- www.stat.wisc.edu/~yandell/qtl/software/Bmapqtl
- genesis
  - 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
  - incorporated into QTLCart (S Wang 2003)
  - built as official R library (H Wu, Yandell, Gaffney, CF Jin 2003)

Gibbs sampler with loci indicators

- partition genome into intervals
  - at most one QTL per interval
  - interval = marker interval or large chromosome region
- use loci indicators in each interval
  - $\delta = 1$ if QTL in interval
  - $\delta = 0$ if no QTL
- Gibbs sampler on loci indicators
  - still need to adjust genetic effects for collinearity of Q
  - see work of Nengjun Yi (and earlier work of Ina Hoeschele)

$$Y = \mu + \delta_1 \alpha_1 (Q_1 - \overline{Q}_1) + \delta_2 \alpha_2 (Q_1 - \overline{Q}_1) + e$$

epistatic interactions

- model space issues
  - 2-QTL interactions only?
  - Fisher-Cockerham partition vs. tree-structured?
  - general interactions among multiple QTL
- model search issues
  - 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)

limits of epistatic inference

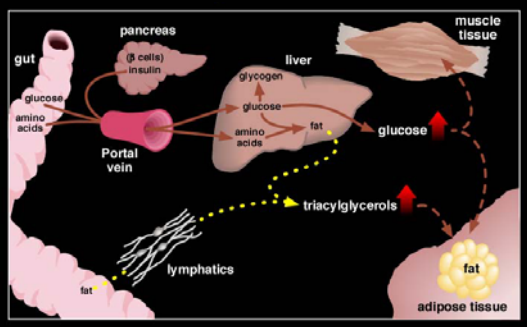
- power to detect effects
  - epistatic model size grows exponentially
    - $|M| = 3^m$ for general interactions
  - power depends on ratio of n to model size
    - want $n / |M|$ to be fairly large (say > 5)
    - $n = 100$, $m = 3$, $n / |M| \approx 4$
- empty cells mess up adjusted (Type 3) tests
  - missing q_1Q_2 / q_1Q_2 or $q_1Q_2q_3 / q_1Q_2q_3$ genotype
  - null hypotheses not what you would expect
  - can confound main effects and interactions
  - can bias AA, AD, DA, DD partition

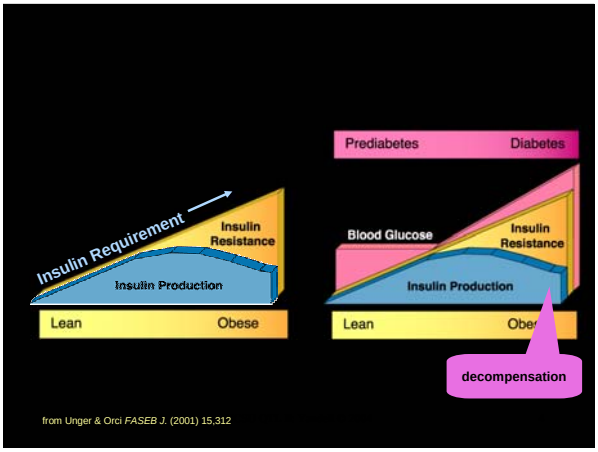
Multiple Traits & Microarrays

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

Type 2 Diabetes Mellitus



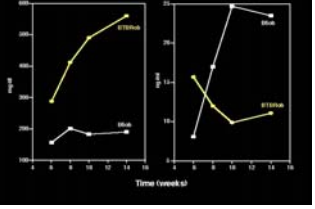


Insulin Resistant Mice


Bill Dove


BTBR strain


glucose insulin



(courtesy AD Attie)

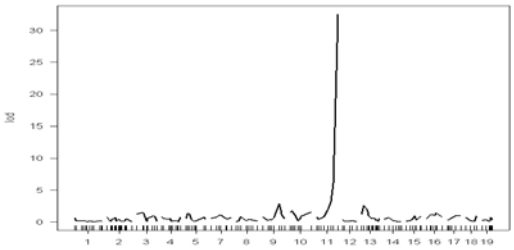
studying diabetes in an F2

- 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
    - (Nadler et al. 2000 *PNAS*; Ntambi et al. 2002 *PNAS*)
  - RT-PCR for a few mRNA on 108 F2 mice liver tissues
    - (Lan et al. 2003 *Diabetes;* Lan et al. 2003 *Genetics*)
  - Affymetrix microarrays on 60 F2 mice liver tissues
    - design (Jin et al. 2004 *Genetics* tent. accept)
    - analysis (work in prep.)

why map gene expression as a quantitative trait?

- *cis*- or *trans*-action?
  - does gene control its own expression?
  - or is it influenced by one or more other genomic regions?
  - evidence for both modes (Brem et al. 2002 Science)
- simultaneously measure all mRNA in a tissue
  - ~5,000 mRNA active per cell on average
  - ~30,000 genes in genome
  - use genetic recombination as natural experiment
- mechanics of gene expression mapping
  - measure gene expression in intercross (F2) population
  - map expression as quantitative trait (QTL)
  - adjust for multiple testing

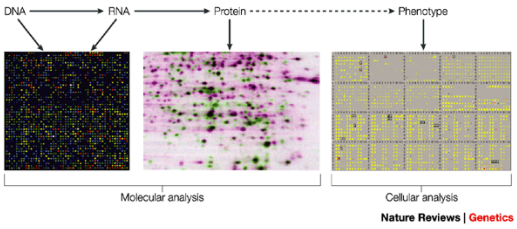
LOD map for PDI: *cis*-regulation (Lan et al. 2003)



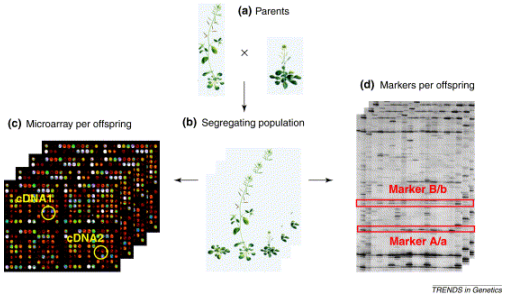
mapping microarray data

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

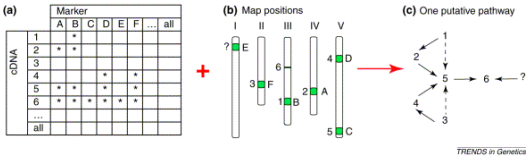
central dogma via microarrays (Bochner 2003)



idea of mapping microarrays (Jansen Nap 2001)



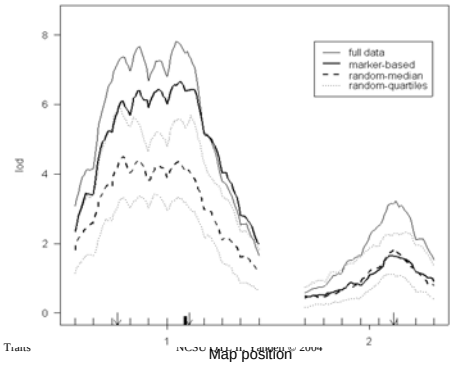
goal: unravel biochemical pathways (Jansen Nap 2001)



2. design issues for expensive phenotypes (thanks to CF “Amy” Jin)

- microarray analysis ~ $1000 per mouse
  - can only afford to assay 60 of 108 in panel
  - wish to not lose much power to detect QTL
- selective phenotyping
  - genotype all individuals in panel
  - select subset for phenotyping
  - previous studies can provide guide

simulated LOD profiles with 3 QTL

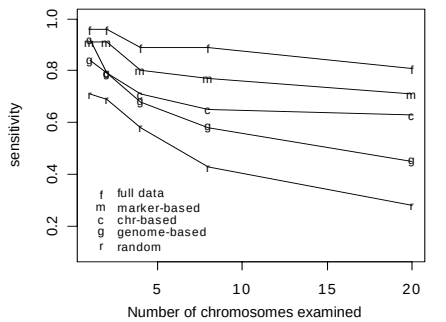


selective phenotyping

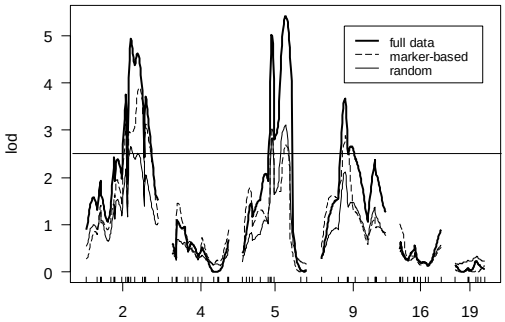
- genotype all individuals in panel
  - whole genome or selected genomic regions?
  - maintain high power in selected regions
  - sensitivity similar to random sample in other regions
- select subset for phenotyping
  - select individuals with large genetic distance
  - use experimental design concepts (Jin et al. 2004)
- previous studies: key regions of chr 2,4,5,9,16,19
  - QTL for important physiological traits

comparison of different selection methods

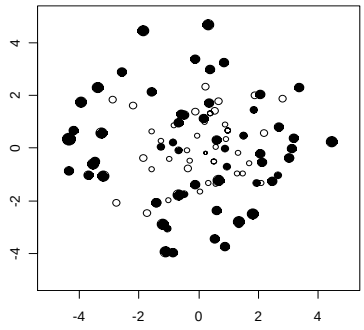
sensitivity = pr(detect QTL | QTL is real)



LOD profile of SCD trait



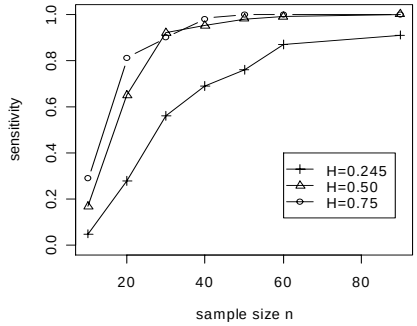
multidimensional scaling of mice selection
(close points have similar genotypes)



is this relevant to large QTL studies?

- why not phenotype entire mapping panel?
  - selectively phenotype subset of 50-67%
  - may capture most effects
  - with little loss of power
- two-stage selective phenotyping?
  - genotype & phenotype subset of 100-300
    - could selectively phenotype using whole genome
  - QTL map to identify key genomic regions
  - selectively phenotype subset using key regions

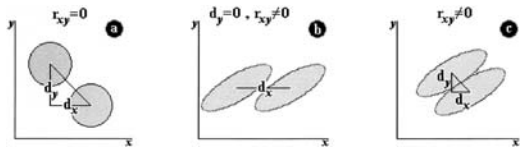
sensitivity = pr(detect QTL | QTL is real)
depends on heritability and proportion sampled (of *N*=100)



3. 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



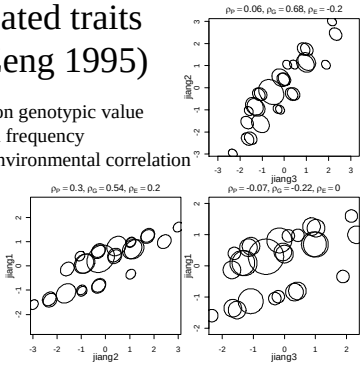
pleiotropy only correlation only both

Korol et al. (2001)

3 correlated traits (Jiang Zeng 1995)

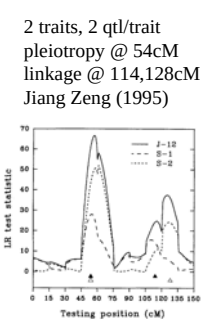
ellipses centered on genotypic value
width for nominal frequency
main axis angle environmental correlation
3 QTL, F2
27 genotypes

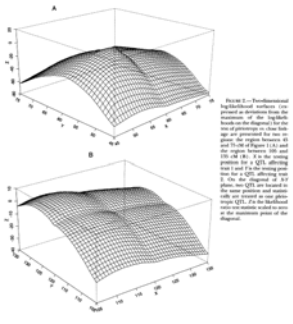
note signs of genetic and environmental correlation



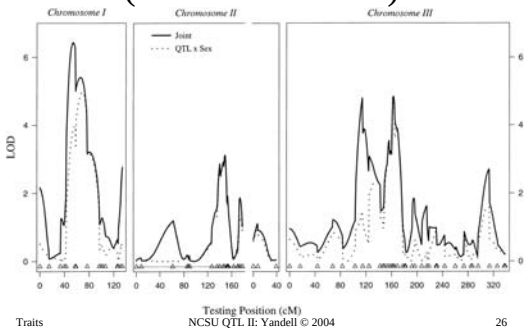
pleiotropy or close linkage?

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





QTL x sex interaction (Vieira et al. 2000)

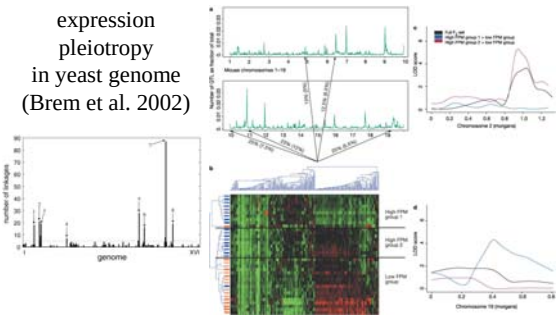


4. high throughput dilemma

- 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
- 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)

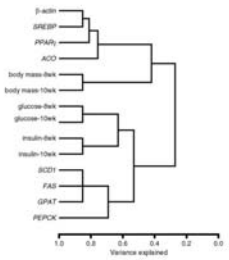
coordinated expression in mouse genome (Schadt et al. 2003)

expression pleiotropy in yeast genome (Brem et al. 2002)



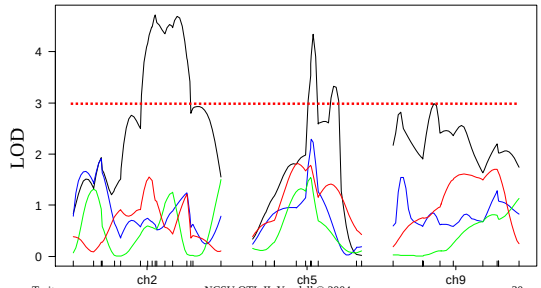
high throughput: which genes are the key players?

Lan et al., mapping mRNA, Figure 2



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

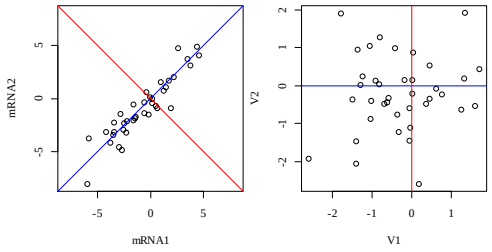
SCD1, FAS,GPAT, PEPCK: *trans*-regulation by multiple QTL?



from gene expression to super-genes

- PC or SVD decomposition of multiple traits
  - Y = t traits × 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

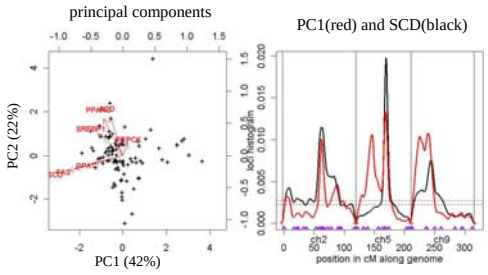
PC simply rotates & rescales to find major axes of variation



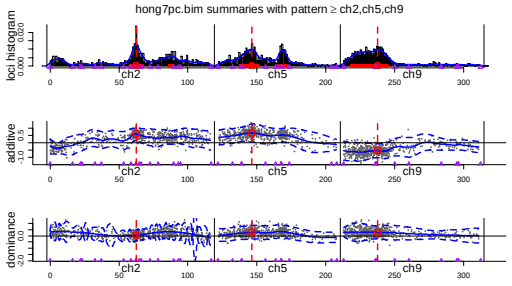
QTL via Principal Components

- *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

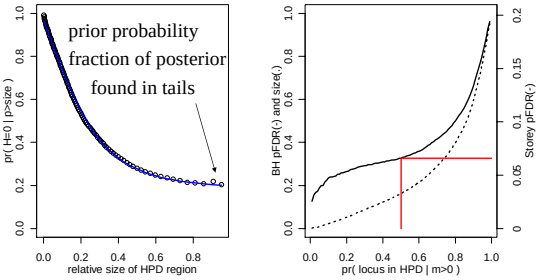
multivariate screen for gene expressing mapping



mapping first diabetes PC as a trait



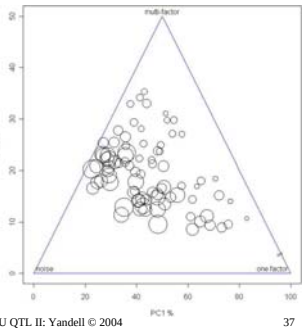
pFDR for PC1 analysis



PC across microarray functional groups

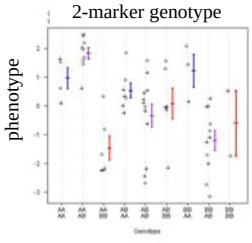
1500+ mRNA of 30,000
85 functional groups
60 mice
2-35 mRNA / group
which are interesting?

examine PC1, PC2
size = # unique mRNA

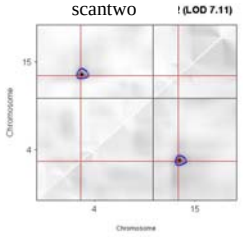


PC-guided search of mRNA
(red lines at main QTL for PC1)

2-marker genotype

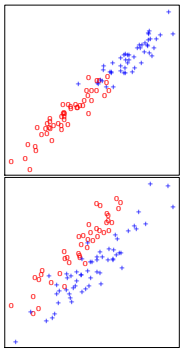


scantwo



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
  - redo at each putative QTL
  - Gilbert and le Roy (2003,2004)



how to map multiple traits?

- WinQTL/QTL Cartographer: IM & CIM
  - Jiang Zeng (1995); statgen.ncsu.edu/qtlcart
- MultiQTL: 1-2 QTL with PC on residuals
  - Korol et al. (2001); www.multiqtl.com
- 1-2 QTL with DA across traits
  - Gilbert and le Roy (2003, 2004)
- QTL Express: Haley-Knott regression
  - Knott Haley (2000); qtl.cap.ed.ac.uk
- SOLAR: outbred pedigrees
  - Almasy Blangero (1997); Williams et al. (1999)