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

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?

2. **selecting a class of QTL models**
3. **comparing QTL models**
   - QTL model selection criteria
   - issues of detecting epistasis
4. **simulations and data studies**
   - simulation with 8 QTL
   - plant BC, animal F2 studies
   - searching through QTL models

### 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: minimize prediction error

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

### 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 $\Rightarrow$ 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^p 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
  - Avoids pitfalls of one "best" model
  - Examines "better" models with more probable QTL
- Build $m = \text{number of QTL detected into QTL model}$
  - 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
- $\theta = \text{gene action}$
  - Additive (A) or general (A+D) effects
  - Epistatic interactions (AA, AD, …, or other types?)
- $\lambda = \text{location of QTL}$
  - Known locations?
  - Widely spaced (no 2 in marker interval) or arbitrarily close?
- $m = \text{number of QTL}$
  - Single QTL?
  - Multiple QTL: known or unknown number?

Normal phenotype

- Trait = mean + genetic + environment
- Genetic effect uncorrelated with environment
- $\text{Pr}(\text{trait} | \text{genotype}, \text{effects})$

Two QTL with epistasis

- Same phenotype model overview
  - $Y = G_0 + \epsilon$, var(\epsilon) = $\sigma^2$
- Partition of genotypic value with epistasis
  - $G_0 = \mu + \beta_1(Q) + \beta_2(Q) + \beta_3(Q)$
- Partition of genetic variance
  - var($G_0$) = $\sigma^2 + \sigma_1^2 + \sigma_2^2$

Epistasis examples

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

- Multiple QTL with epistasis
  - Same overview model
    - $Y = G_0 + \epsilon$, var(\epsilon) = $\sigma^2$
  - Sum over multiple QTL in model $M = \{1,2,12,…\}$
    - $G_0 = \mu + \sum_{\beta_j(M)} \beta_j(Q)$
  - Partition genetic variance in same manner
    - var($G_0$) = $\sum_{\beta_j(M)} \sigma_j^2$
  - 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
  - 2 QTL in F2: 4 df
  - 3 QTL in F2: 20 df
- 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?

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

information criteria: likelihoods

- \( L(p) \) = likelihood for model with \( p \) parameters
- common information criteria:
  - Akaike AIC = \(-2 \log(L(p)) + 2p \)
  - Bayes/Schwartz BIC = \(-2 \log(L(p)) + p \log(n) \)
  - BIC-delta BIC = \(-2 \log(L(p)) + \delta(p) \log(n) \)
  - general form: IC = \(-2 \log(L(p)) + p \Delta(n) \)

Bayes factors & BIC

- 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

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

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

- partition genotype-specific mean into QTL effects

\[
GQ = \mu + \beta(Q) = \mu + \sum_{j=1}^{M} \beta_j(Q)
\]

- priors on mean and effects
  - $\mu \sim N(\mu_0, \kappa_0 \sigma^2)$ grand mean
  - $\beta(Q) \sim N(0, \kappa_1 \sigma^2)$ model-independent genotypic effect
  - $\beta_j(Q) \sim N(0, \kappa_1 \sigma^2/M)$ effects down-weighted by size of $M$

- determine hyper-parameters via Empirical Bayes

\[
\mu_0 = \bar{y} \text{ and } \kappa_1 = \frac{b^2}{\bar{y} - \mu_0} = \frac{\sigma_0^2}{\sigma^2}
\]

multiple QTL posteriors

- phenotype influenced by genotype & environment
  $pr(Y|Q, \theta) \sim N(GQ, \sigma^2)$, or $Y = GQ + \text{environment}$

- relation of posterior mean to LS estimate

\[
\hat{G}_Q = \frac{\sum w_i \hat{y}_i}{\sum w_i} = \frac{\sum w_i (G_Q + \text{environment})}{\sum w_i}
\]

\[
\text{variance } V(\hat{G}_Q) = \sum w_i \sigma_0^2 - \frac{\sigma^2}{\sum w_i}
\]

\[
\text{shrinkage } B_0 = \kappa / (\kappa + \sigma_0^2) \rightarrow 1
\]

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 $= 97\%$

- notice which chromosomes have persistent loci
  - best pattern found $42\%$ of the time

loci pattern across genome
**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

**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.Cb x BTHR.Cb → 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 tissue: 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

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!
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
    - $A = \{ \lambda | \text{LOD}(\lambda) > \text{threshold} \}$
    - extends naturally to multiple QTL

pFDR and QTL posterior

- positive false detection rate
  - pFDR = pr( no QTL at λ | Y, X, \lambda in A )
  - pFDR = \frac{pr(H=0) \times \text{size}}{pr(m=0) \times \text{size} + pr(m>0) \times \text{power}}
  - power = posterior = pr(QTL in A | Y, X, m>0)
  - size = (length of $A$) / (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

prior probability
fraction of posterior found in tails

relative size of HPD region
pr locus in HPD (m=0)

relative size of HPD region
pr locus in HPD (m>0)
Extending the Phenotype Model

1. limitations of parametric models
   - diagnostic tools for QTL analysis
   - QTL mapping with other parametric "families"
   - quick fixes via data transformations

2. semi-parametric approaches

3. non-parametric approaches
   - 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)
     - 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

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 (1–2)\(r\)
  - plot effect vs. location

marker density & sample size: 2 QTL

modest sample size
   - dense vs. sparse markers

large sample size
   - dense vs. sparse markers

robust locus estimate for non-normal phenotype

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

but what happens for modest sample sizes?
What shape is your histogram?

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

interval mapping likelihood

- likelihood: basis for scanning the genome
  - product over $i = 1, \ldots, n$ individuals
    $L(\theta, \lambda|Y) = \prod_i p(Y_i|X_i, \lambda)$
  - $p(Y|X, \lambda, \theta) = \sum_Q p(Q|X, \lambda) p(Y|Q, \theta)$

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
  - Poisson
  - binomial
  - clustered negative binomial
- transform to stabilize variance
  - counts $\sqrt{Y}$ or log($Y+c$)
  - fractions $\arcsin(\sqrt{Y})$
- transform to symmetry (approx. normal)
  - Poisson $p(Y|Q, \theta) = \text{Poisson}(G_Q)$
- empirical transform based on histogram
  - watch out: hard to do well even without mixture
  - probably better to map untransformed, then examine residuals

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)
  - $Y \sim \text{binomial}(1, \pi)$, $\pi$ depends on genotype $Q$
  - $p(Y|Q) = (\pi_Q)^Y (1-\pi_Q)^{1-Y}$
  - 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 $Q$
  $p(Q|X, Y, \lambda, \pi_Q) \propto p(Y|X, \lambda, \pi_Q)^Y (1-\pi_Q)^{1-Y}$
- M-step: MLE of binomial probability $\pi_Q$
  $\pi_Q = \frac{\sum_Y Y p(Q|Y, X, \lambda, \pi_Q)}{\sum_Y p(Q|Y, X, \lambda, \pi_Q)}$
threshold or latent variable idea

- "real", unobserved phenotype $Z$ is continuous
- observed phenotype $Y$ is ordinal value
  - no/yes; poor/fair/good/excellent
  - $p_r(Y=j) = p_r(\tau_{j-1} < Z \leq \tau_j)$
- $p_r(Y \leq j) = p_r(Z \leq \tau_j)$

- 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)

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

other parametric approaches

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

semi-parametric empirical likelihood

- phenotype model $p_r(Y|Q, \theta) = f(Y \exp(Y\beta Q))$
  - “point mass” at each measured phenotype $Y_i$
  - subject to distribution constraints for each $Q$:
    - $1 = \sum_i f(Y_i \exp(Y_i\beta Q))$

- non-parametric empirical likelihood (Owen 1988)
  - $L(\theta, \lambda|Y, X) = \prod_i f(Y_i \exp(Y_i\beta Q))$
  - 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: $L(\lambda|Y, X) = \max_{\lambda} L(\theta, \lambda|Y, X)$

semi-parametric formal tests

- partial empirical LOD
  - Zou, Fine, Yandell (2002 Biometrika)
- conditional empirical LOD

- has same formal behavior as parametric LOD
  - single locus test: approximately $\chi^2$ 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

$$\log(L(\theta|Y, X)) = \sum_i \log(f(Y_i)) + \log(w_i)$$

now profile with respect to $\beta, \lambda$

$$\log(L(\beta, \lambda|Y, X)) = \sum_i \log(f(Y_i)) + \log(w_i) + \sum_{Q} \alpha_Q (1-\sum_i f_i \exp(Y_i\beta_Q))$$

partial likelihood: set Lagrange multipliers $\alpha_Q$ to 0

force $f_i$ to be a distribution that sums to 1

point mass density estimates

$$f_i = (\sum_i w_i)^{-1}$$

with $w_i = \sum_{Q} \exp(Y_i\beta_Q) p_r(Q|X_i, \lambda)$
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?

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 \( f(Y) \)
- \( p(X|Y, \theta, \lambda, Q) \) depends only on mating design
- unbiased for selective genotyping (Jin et al. 2004)

new resampling threshold method
- EM locally approximates LOD by quadratic form
- use local covariance of \( \hat{\beta} \) estimates to further approximate
  - relies on \( n \) independent standard normal variates \( Z = (Z_1, \ldots, Z_n) \)
  - one set of variates 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

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 \( \hat{\beta} \) estimates to further approximate
  - relies on \( n \) independent standard normal variates \( Z = (Z_1, \ldots, Z_n) \)
  - one set of variates 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

\[
\text{LOD}(\lambda) \approx n \hat{\beta}(\lambda) S(\lambda) \hat{\beta}(\lambda) + Z^T C(\lambda) S(\lambda) C(\lambda) Z
\]
\[
\text{cov}(\hat{\beta}(\lambda)) \approx -C(\lambda) S(\lambda) C(\lambda)
\]
\[
\sqrt{n} \hat{\beta}(\lambda) \approx C(\lambda) Z, \text{ with } Z \sim N(0, I)
\]
3. non-parametric methods

- phenotype model $p(Y|Q, \theta) = F_Q(Y)$
- $\theta = F(\theta_Q, F_Q(Q))$ 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)
  - non-parametric cumulative distribution
    - Fine, Zou, Yandell (2001)
  - stochastic ordering (Hoff et al. 2002)

rank-sum QTL methods

- phenotype model $p(Y|Q, \theta) = F_Q(Y)$
- 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 $\beta$
  - most efficient if $p(Y|Q, \theta) = F(Y + \beta)$
  - find $\beta$ that matches medians
    - problem: genotypes $Q$ unknown
    - resolution: Haley-Knott (1992) regression scan
  - works well in practice, but theory is elusive
- Fine, Zou, Yandell (2001)

non-parametric QTL CDFs

- estimate non-parametric phenotype model
  - cumulative distributions $F_Q(y) = p(Y \leq y | Q)$
  - can use to check parametric model validity
- basic idea:
  $p(Y \leq y | X, \lambda) = \sum_Q p(Q|X, \lambda) F_Q(y)$
  - 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 $F_Q(y) = p(Y \leq y | Q)$
- $F = \{F_Q, \text{all possible QT genotypes } Q\}$
- BC with 1 QTL: $F = \{F_{QQ}, F_{Qq}\}$
- find $F$ to minimize over all phenotypes $y$
  $\sum_i [I(Y_i \leq y) - \sum_Q p(Q|X, \lambda) F_Q(y)]^2$
- 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 $F_Q(y) = \{F_Q(y)\}$ 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
  - Stellar male
  - insensitive
  - 99 double haploids
  - $Y = \log(\text{days to flower})$ 
  - over 50% Major at QTL never flowered
  - log not fully effective

grey = normal, red = non-parametric
what shape is flowering distribution?

\[ B. napus \text{ Stellar} \quad B. napus \text{ Major} \]

- line = normal
- * = non-parametric
- o = confidence interval
Bayesian Interval Mapping

1. Who was Bayes? 2-6
   - What is Bayes theorem?

2. Bayesian inference for QTL 7-14

3. Markov chain sampling 15-29
   - for fixed number of QTL \( m \)

4. Sampling across architectures 30-40
   - handling epistasis

what is Bayes theorem?

- where is first ball if the second is to its left (right)?
- prior: probability of parameter before observing data
  \[ \text{pr}(\theta) = \text{pr}(\text{parameter}) \]
  - equal chance of being anywhere on the table
- posterior: probability of parameter after observing data
  \[ \text{pr}(\theta|Y) = \text{pr}(\theta \mid \text{data}) \]
  - more likely to left if first ball is toward the right end of table
- likelihood: probability of data given parameters
  \[ \text{pr}(Y|\theta) = \text{pr}(\text{data} | \theta) \]
  - basis for classical statistical inference
- Bayes theorem:
  \[ \text{pr}(\theta|Y) = \frac{\text{pr}(\theta,Y)}{\text{pr}(Y)} = \frac{\text{pr}(Y|\theta) \times \text{pr}(\theta)}{\text{pr}(Y)} \]

1. who was Bayes? what is Bayes theorem?

\[
\begin{align*}
\text{prior } \text{pr}(\theta) & = 1 \\
\text{likelihood } \text{pr}(Y|\theta) & = \theta_i (1-\theta)_{1-\theta} \\
\text{posterior } \text{pr}(\theta|Y) & = ?
\end{align*}
\]

Bayes posterior for normal data

model
\[ Y_i = \mu + E_i \]
environment \[ E \sim N(0, \sigma^2) \]
likelihood \[ Y \sim N(\mu, \sigma^2) \]
prior \[ \mu \sim N(\mu_0, \kappa \sigma^2) \]

\[
\begin{align*}
\text{posterior: mean tends to sample mean} \\
\text{single individual} & = \mu \sim N(\mu_0 + B_i (Y_i - \mu_0), B_i^2 \sigma^2) \\
sample of n individuals & = \mu \sim N \left( \frac{B_i \sum (1-B_i) \mu_0 + B_i \mu_i}{n} \sigma^2 \right) \\
\text{with } & \sum = \sum_i \frac{1}{n} \\
\text{fudge factor} & (\text{shrinks to 1}) \rightarrow \frac{B_i}{m+n} \\
\end{align*}
\]
2. Bayesian inference for QTL

- develop priors on unknowns
  - unknowns:
    - missing genotypes $Q$
    - effects $\theta = (GQ, \sigma^2)$
    - loci $\lambda$ (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 $\theta$ loci $\lambda$

Bayesian priors for QTL

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

Bayesian model posterior

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

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

posterior on QTL genotypes

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

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

Bayesian model posterior

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

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

Bayesian priors for QTL

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

how does phenotype Y improve posterior for genotype Q?

posterior genotypic means $G_Q$
posterior centered on sample genotypic mean but shrunken slightly toward overall mean

prior: \( G_0 \sim N\left(\bar{y}, \sigma^2\right) \)

posterior: \( G_i \sim N\left(B_y \bar{y} + (1 - B_y) \bar{y}, B_y \sigma^2 + n_0 \right) \)

fudge factor: \( B_y = \frac{kn_0}{kn_0 + 1} \rightarrow 1 \)

What if variance \( \sigma^2 \) is unknown?

• sample variance is proportional to chi-square
  \( \sigma^2 / \sigma^2 \sim \chi^2(n) \)
  likelihood of sample variance \( s^2 \) given \( n, \sigma^2 \)

• conjugate prior is inverse chi-square
  \( \nu \tau^2 / \sigma^2 \sim \chi^2(\nu) \)
  prior of population variance \( \sigma^2 \) given \( \nu, \tau^2 \)

• posterior is weighted average of likelihood and prior
  \( (\nu \tau^2 + ns^2) / \sigma^2 \sim \chi^2(\nu + n) \)
  posterior of population variance \( \sigma^2 \) given \( n, \sigma^2, \nu, \tau^2 \)

• empirical choice of hyper-parameters
  \( \tau^2 = s^2 / 3, \nu = 6 \)
  \( E(\sigma^2 | \nu, \tau^2) = s^2 / 2, \text{Var}(\sigma^2 | \nu, \tau^2) = s^4 / 4 \)

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 \((\lambda, Q, \theta, m)\) from joint posterior
  - update \((\lambda, Q, \theta)\) from full conditionals for \(m\)-QTL model
  - update \(m\) using reversible jump technology

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

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 at we wish to study
• toy problem
  - two states \((0,1)\)
  - move chances depend on current state
  - what is the chance of being in state \(1\)?

What is variance \( \sigma^2 \)?

\[ \text{What is variance } \sigma^2 \text{ of unknown?} \]

\[ \begin{align*}
  \text{• sample variance is proportional to chi-square} \\
  \sigma^2 / \sigma^2 \sim \chi^2(n) \\
  \text{likelihood of sample variance } s^2 \text{ given } n, \sigma^2 \\
  \text{• conjugate prior is inverse chi-square} \\
  \nu \tau^2 / \sigma^2 \sim \chi^2(\nu) \\
  \text{prior of population variance } \sigma^2 \text{ given } \nu, \tau^2 \\
  \text{• posterior is weighted average of likelihood and prior} \\
  (\nu \tau^2 + ns^2) / \sigma^2 \sim \chi^2(\nu + n) \\
  \text{posterior of population variance } \sigma^2 \text{ given } n, s^2, \nu, \tau^2 \\
  \text{• empirical choice of hyper-parameters} \\
  \tau^2 = s^2 / 3, \nu = 6 \\
  E(\sigma^2 | \nu, \tau^2) = s^2 / 2, \text{Var}(\sigma^2 | \nu, \tau^2) = s^4 / 4 \\
\end{align*} \]
Gibbs sampler samples: $\rho = 0.6$

MCMC sampling of $(\lambda, Q, \theta)$

- Gibbs sampler
  - effects $\theta = (G, \sigma^2)$
  - genotypes $Q$
  - not loci $\lambda$

- extension of Gibbs sampler
  - Metropolis-Hastings sampler
  - does not require normalization
  - loci $\lambda$: $\text{pr}(Q | X, \lambda)$ 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
  - propose new value $\theta^\ast$
    - from some distribution $g(\theta, \theta^\ast)$
    - accept new value with prob $A$
    - otherwise stick with current value

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

Metropolis-Hastings samples

full conditional for locus

- cannot easily sample from locus full conditional
  $\text{pr}(\lambda | Y, X, Q) = \frac{\text{pr}(\lambda | X, \lambda)}{\text{pr}(Q | X, \lambda)}$
- 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

$A(\lambda^\ast, \lambda^\ast) = \min\left(1, \frac{\text{pr}(\lambda^\ast)\text{pr}(Q | X, \lambda^\ast)g(\lambda^\ast, \lambda^\ast)}{\text{pr}(\lambda)\text{pr}(Q | X, \lambda)g(\lambda, \lambda^\ast)}\right)$
**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**

summarizes 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**

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

4-week vernalization
8-week vernalization

<table>
<thead>
<tr>
<th>cM</th>
<th>add</th>
<th>cM</th>
<th>add</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>0.30</td>
<td>40</td>
<td>0.06</td>
</tr>
<tr>
<td>80</td>
<td>0.16</td>
<td>80</td>
<td>0.13</td>
</tr>
</tbody>
</table>

**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
      \[ G(Q) = \mu + \beta(Q) \text{ with } \beta(Q) = \alpha \times (Q - \bar{Q}) \]

model selection in regression

• consider known genotypes \( Q \) at 2 known loci \( \lambda \)
  – models with 1 or 2 QTL
• jump between 1-QTL and 2-QTL models
  – adjust parameters when model changes
  – \( \alpha_1 \) and \( \alpha_2 \) differ due to collinearity of QTL genotypes
\[ m = 1 : Y = \mu + \alpha_1 (Q_1 - \bar{Q}_1) + e \]
\[ m = 2 : Y = \mu + \alpha_1 (Q_1 - \bar{Q}_1) + \alpha_2 (Q_2 - \bar{Q}_2) + e \]

geometry of reversible jump

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

geometry allowing \( Q \) and \( \lambda \) to change

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](http://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](http://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)
  \[
  Y = \mu + \delta_1 \alpha_1 (Q_1 - \bar{Q}_1) + \delta_2 \alpha_2 (Q_2 - \bar{Q}_2) + e
  \]

**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?

**Limits of epistatic inference**

- **Power to detect effects**
  - Epistatic model size grows exponentially
    - \(|M| = 3^n\) for general interactions
  - Power depends on ratio of \( n \) to model size
    - Want \( n / |M| \) to be fairly large (say > 5)
    - \( n = 100, m = 3, n / |M| = 4 \)
- **Empty cells mess up adjusted (Type 3) tests**
  - Missing \( q_1Q_2 / q_1Q_2 \) or \( q_1Q_2 / q_2Q_1 \) 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?  2-13
   - diabetes case study
   - central dogma via microarrays
2. design issues for expensive phenotypes  14-21
   - selective phenotyping
3. why are traits correlated?  22-26
   - close linkage or pleiotropy?
4. how to handle high throughput?  27-40
   - 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

Type 2 Diabetes Mellitus

Insulin Requirement
from Unger & Orci
FASEB J. (2001) 15,312

decompensation

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; Niambi 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

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

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 = \( \text{pr( detect QTL | QTL is real )} \)

LOD profile of SCD trait

<table>
<thead>
<tr>
<th>LOD profile of SCD trait</th>
<th>full data</th>
<th>marker-based</th>
<th>random</th>
<th>chr-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of chromosomes examined

LOD profile of SCD trait

<table>
<thead>
<tr>
<th>LOD profile of SCD trait</th>
<th>full data</th>
<th>marker-based</th>
<th>random</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

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

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

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

high throughput:
which genes are the key players?

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

QTL x sex interaction
(Vieira et al. 2000)

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

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

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 $Y = UD\bar{W}$
    - $U, W$ = 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 correalted 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)

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)