Seattle Summer Institute 2006
Advanced QTL
Brian S. Yandell
University of Wisconsin-Madison

- Bayesian QTL mapping & model selection
- data examples in detail
- multiple phenotypes & microarrays
- software demo & automated strategy

contact information & resources

- email: byandell@wisc.edu
- web: www.stat.wisc.edu/~yandell/statgen
  - QTL & microarray resources
  - references, software, people
- thanks:
  - students: Jaya Satagopan, Pat Gaffney, Fei Zou, Amy Jin, W. Whipple Neely
  - faculty/staff: Alan Attie, Michael Newton, Nengjun Yi, Gary Churchill, Hong Lan, Christina Kendziorski, Tom Osborn, Jason Fine, Tapan Mehta, Hao Wu, Samprit Banerjee, Daniel Shriner

Bayesian Interval Mapping

1. 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

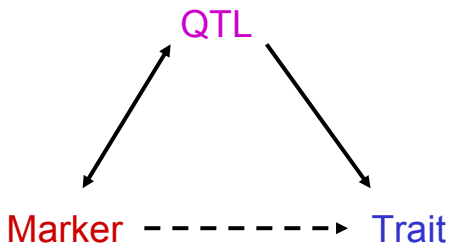
cross two inbred lines

→ linkage disequilibrium

→ associations

→ linked segregating QTL

(after Gary Churchill)



pragmatics of multiple QTL

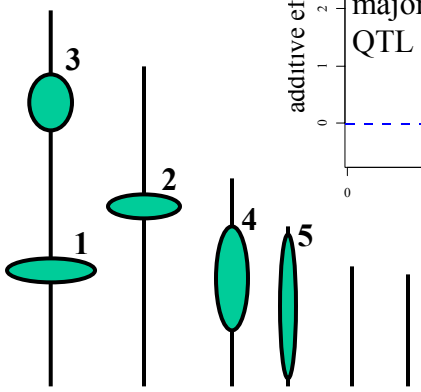
- evaluate some objective for model given data
  - classical likelihood
  - Bayesian posterior
- search over possible genetic architectures (models)
  - number and positions of loci
  - gene action: additive, dominance, epistasis
- estimate "features" of model
  - means, variances & covariances, confidence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

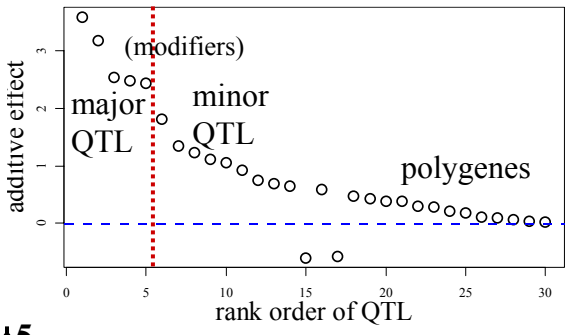
advantages of multiple QTL approach

- 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

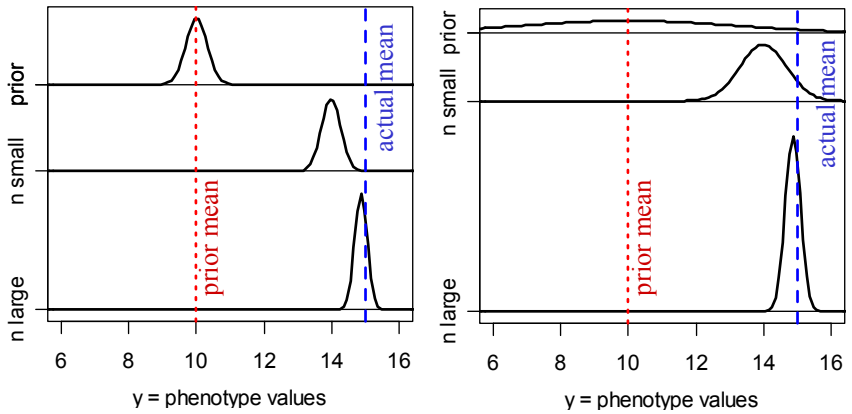
QTL below detection level?

- 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. Bayesian QTL mapping

- 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?
    - prior: anywhere on the table
    - posterior: more likely toward right end of table

Bayes posterior for normal data



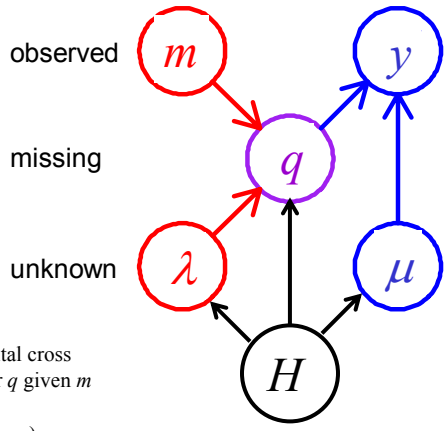
small prior variance large prior variance

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	$\mu \sim N\left(b_n\bar{y}_\bullet + (1-b_n)\mu_0, b_n\sigma^2/n\right)$ with $\bar{y}_\bullet = \underset{\{i=1,\ldots,n\}}{\text{sum}} y_i / n$
fudge factor (shrinks to 1)	$b_n = \frac{\kappa n}{\kappa n + 1} \to 1$

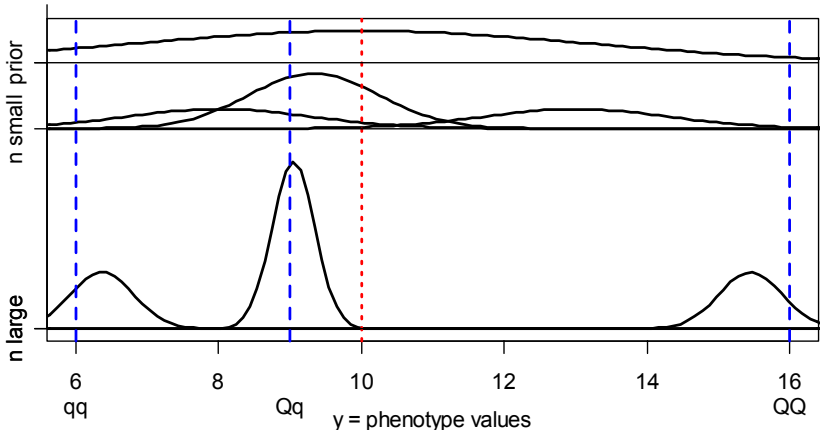
Bayesian QTL: key players

- observed measurements
  - y = phenotypic trait
  - m = markers & linkage map
  - i = individual index (1,…,n)
- missing data
  - missing marker data
  - q = QT genotypes
    - alleles QQ, Qq, or qq at locus
- unknown quantities
  - λ = QT locus (or loci)
  - μ = phenotype model parameters
  - H = QTL model/genetic architecture
- pr($q|m,\lambda,H$) genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for q given m
- pr($y|q,\mu,H$) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters μ (could be non-parametric)



after
Sen Churchill (2001)

pr($y|q,\mu$) phenotype model



Bayes posterior QTL means

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

prior: $\mu_q \sim N\left(\bar{y}_\bullet, \kappa\sigma^2\right)$

posterior: $\mu_q \sim N\left(b_q\bar{y}_q + (1-b_q)\bar{y}_\bullet, b_q\sigma^2/n_q\right)$

$$n_q = \text{count}\{q_i = q\}, \bar{y}_q = \underset{\{q_i=q\}}{\text{sum}} y_i / n_q$$

fudge factor: $b_q = \dfrac{\kappa n_q}{\kappa n_q + 1} \to 1$

partition of multiple QTL effects

- partition genotype-specific mean into QTL effects
  μ_q = mean + main effects + epistatic interactions
  $\mu_q = \mu + \beta_q = \mu + \text{sum}_{j \text{ in } H}\, \beta_{qj}$
- priors on mean and effects

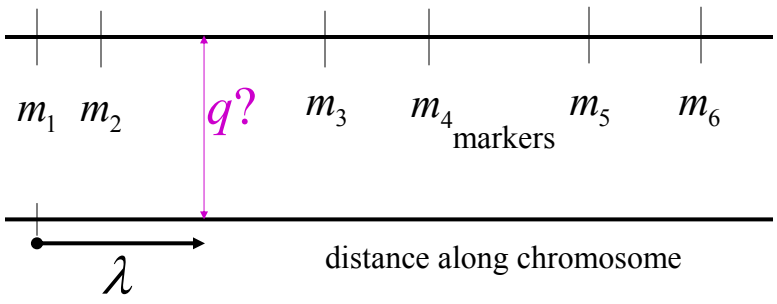
μ	~	$N(\mu_0, \kappa_0\sigma^2)$	grand mean
β_q	~	$N(0, \kappa_1\sigma^2)$	model-independent genotypic effect
β_{qj}	~	$N(0, \kappa_1\sigma^2/\lvert H\rvert)$	effects down-weighted by size of H

- determine hyper-parameters via empirical Bayes

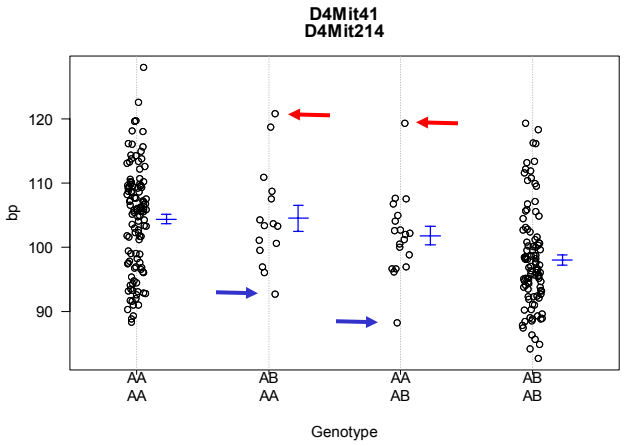
$$\mu_0 \approx \bar{Y}_\bullet \text{ and } \kappa_1 \approx \frac{h^2}{1-h^2} = \frac{\sigma_G^2}{\sigma^2}$$

pr($q|m,\lambda$) recombination model

pr($q|m,\lambda$) = pr(geno | map, locus) ≈
pr(geno | flanking markers, locus)



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 for q depends data for individual i
  - proportional to prior $\text{pr}(q \mid m_i, \lambda)$
    - weight toward q that agrees with flanking markers
  - proportional to likelihood $\text{pr}(y_i|q,\mu)$
    - weight toward q so that group mean $\mu_q \approx y_i$
- phenotype and prior recombination may conflict
  - posterior recombination balances these two weights
  - this is "E step" in EM for classical QTL analysis

$$\text{pr}(q \mid y_i, m_i, \mu, \lambda) = \frac{\text{pr}(q \mid m_i, \lambda)\text{pr}(y_i \mid q, \mu)}{\text{pr}(y_i \mid m_i, \mu, \lambda)}$$

Bayesian model posterior

- augment data (y,m) with unknowns q
- study unknowns (μ,λ,q) given data (y,m)
  - properties of posterior pr($\mu,\lambda,q \mid y,m$)
- sample from posterior in some clever way
  - multiple imputation or MCMC

$$\text{pr}(q,\mu,\lambda \mid y,m) = \frac{\text{pr}(y \mid q,\mu)\text{pr}(q \mid m,\lambda)\text{pr}(\mu)\text{pr}(\lambda \mid m)}{\text{pr}(y \mid m)}$$

$$\text{pr}(\mu,\lambda \mid y,m) = \text{sum}_q \ \text{pr}(q,\mu,\lambda \mid y,n)$$

Bayesian priors for QTL

- missing genotypes q
  - pr($q \mid m, \lambda$)
  - recombination model is formally a prior
- effects (μ, σ^2)
  - prior = pr($\mu_q \mid \sigma^2$) pr(σ^2)
  - use conjugate priors for normal phenotype
    - pr($\mu_q \mid \sigma^2$) = normal
    - pr(σ^2) = inverse chi-square
- each locus λ may be uniform over genome
  - pr($\lambda \mid m$) = 1 / length of genome
- combined prior
  - pr($q, \mu, \lambda \mid m$) = pr($q \mid m, \lambda$) pr(μ) pr($\lambda \mid m$)

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 (q, μ , λ, H) from joint posterior
  - update (q,μ,λ) from full conditionals for model H
  - update genetic architecture H

$$(q,\mu,\lambda,H) \sim \text{pr}(q,\mu,\lambda,H \mid y,m)$$

$$(q,\mu,\lambda,H)_1 \rightarrow (q,\mu,\lambda,H)_2 \rightarrow \cdots \rightarrow (q,\mu,\lambda,H)_N$$

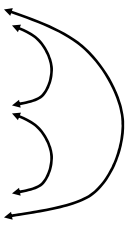
MCMC sampling of (λ,q,μ)

- Gibbs sampler
  - genotypes q
  - effects μ
  - *not* loci λ

$$q \sim \text{pr}(q \mid y_i, m_i, \mu, \lambda)$$

$$\mu \sim \frac{\text{pr}(y \mid q,\mu)\text{pr}(\mu)}{\text{pr}(y \mid q)}$$

$$\lambda \sim \frac{\text{pr}(q \mid m,\lambda)\text{pr}(\lambda \mid m)}{\text{pr}(q \mid m)}$$



- Metropolis-Hastings sampler
  - extension of Gibbs sampler
  - does not require normalization
    - pr(q | m) = sum$_\lambda$ pr(q | m, λ) pr(λ)

full conditional for locus

- cannot easily sample from locus full conditional

$$\begin{aligned} \text{pr}(\lambda \mid y,m,\mu,q) &= \text{pr}(\lambda \mid m,q) \\ &= \text{pr}(q \mid m,\lambda)\,\text{pr}(\lambda) / \text{constant} \end{aligned}$$

- constant is very difficult to compute explicitly
  - must average over all possible loci λ over genome
  - must do this for every possible genotype q
- Gibbs sampler will not work in general
  - but can use method based on ratios of probabilities
  - Metropolis-Hastings is extension of Gibbs sampler

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

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

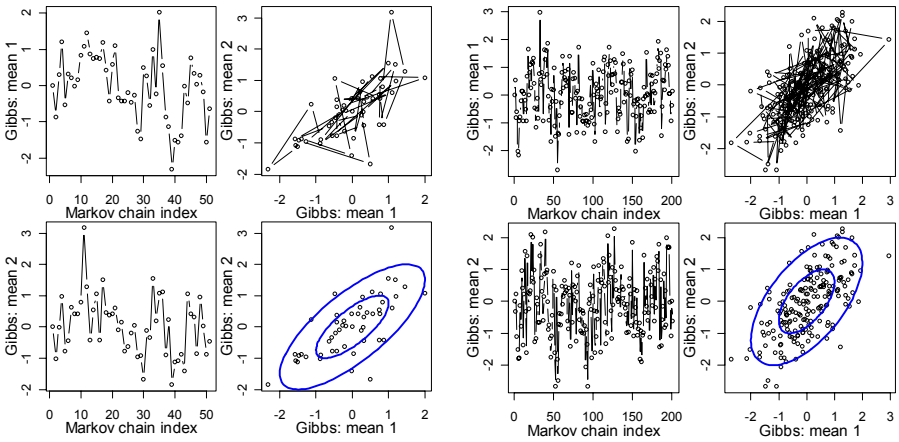
$$\mu_1 \sim N\left(\rho\mu_2, 1-\rho^2\right)$$

$$\mu_2 \sim N\left(\rho\mu_1, 1-\rho^2\right)$$

Gibbs sampler samples: $\rho = 0.6$

$N = 50$ samples

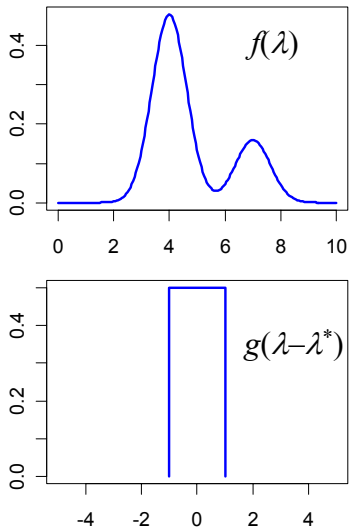
$N = 200$ samples



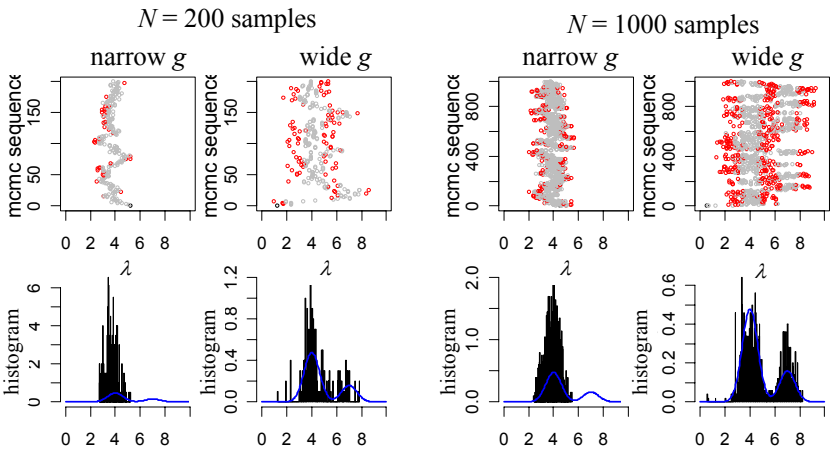
Metropolis-Hastings idea

- want to study distribution $f(\lambda)$
  - take Monte Carlo samples
    - unless too complicated
  - take samples using ratios of f
- Metropolis-Hastings samples:
  - propose new value λ^*
    - near (?) current value λ
    - from some distribution g
  - accept new value with prob a
    - Gibbs sampler: $a = 1$ always

$$a = \min\left(1, \frac{f(\lambda^*)g(\lambda^* - \lambda)}{f(\lambda)g(\lambda - \lambda^*)}\right)$$



Metropolis-Hastings samples



4. sampling across architectures

- search across genetic architectures *M* of various sizes
  - allow change in number of QTL
  - allow change in types of epistatic interactions
- 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

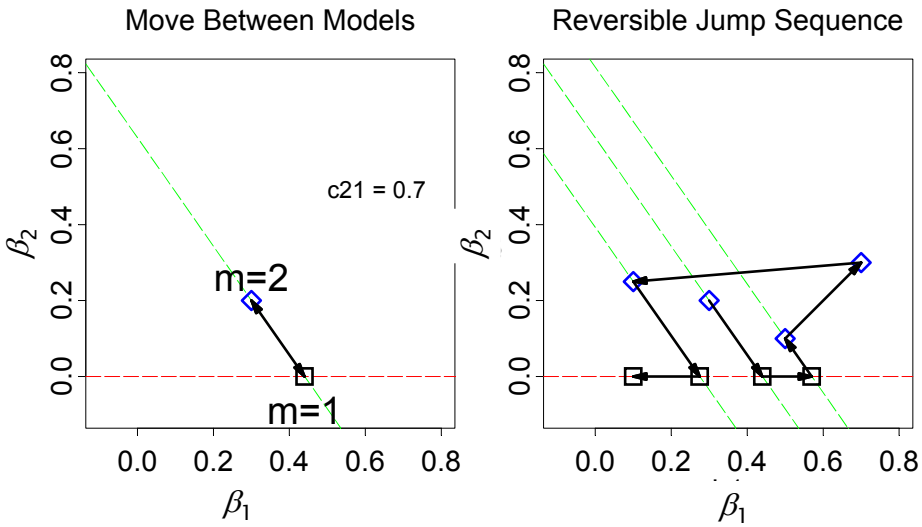
model selection in regression

- 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
  - β_{q1} estimate changes between models 1 and 2
  - due to collinearity of QTL genotypes

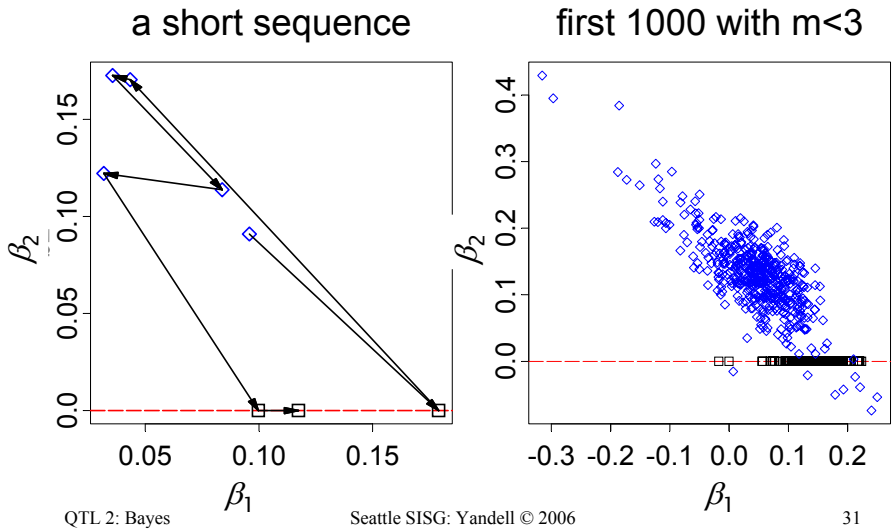
$$m = 1 : \mu_q = \mu + \beta_{q1}$$

$$m = 2 : \mu_q = \mu + \beta_{q1} + \beta_{q2}$$

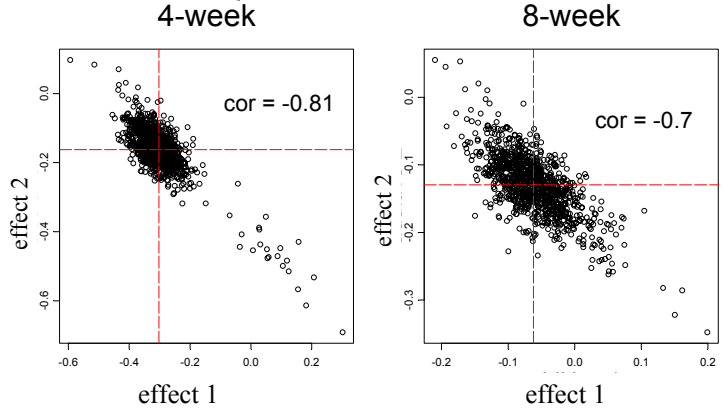
geometry of reversible jump



geometry allowing q and λ 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

reversible jump MCMC idea

- 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 and innovate new genotypes & genotypic 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
- Satagopan Yandell (1996, 1998); Sillanpaa Arjas (1998); Stevens Fisch (1998)
  - these build on RJ-MCMC idea of Green (1995); Richardson Green (1997)

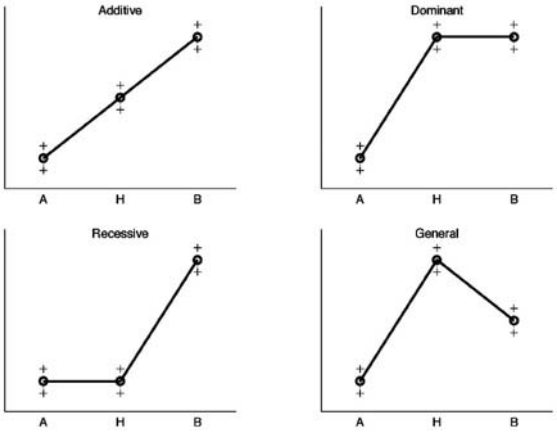
Gibbs sampler with loci indicators

- consider only QTL at pseudomarkers
  - every 1-2 cM
  - modest approximation with little bias
- use loci indicators in each pseudomarker
  - $\delta = 1$ if QTL present
  - $\delta = 0$ if no QTL present
- Gibbs sampler on loci indicators δ
  - relatively easy to incorporate epistasis
  - Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005 *Genetics*)
    - (see earlier work of Nengjun Yi and Ina Hoeschele)

$$\mu_q = \mu + \delta_1 \beta_{q1} + \delta_2 \beta_{q2}$$

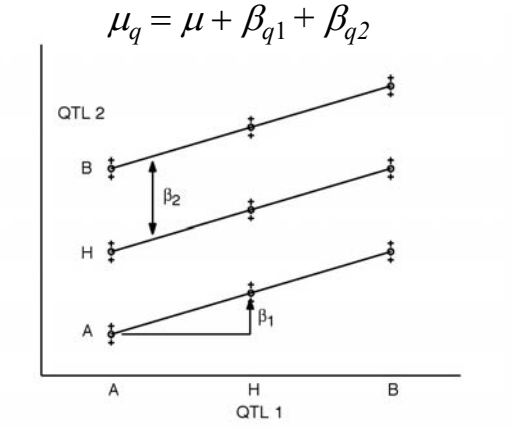
5. Gene Action and Epistasis

additive, dominant, recessive, general effects of a single QTL (Gary Churchill)



additive effects of two QTL (Gary Churchill)

$$\mu_q = \mu + \beta_{q1} + \beta_{q2}$$



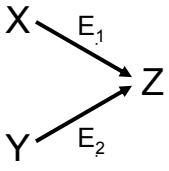
Epistasis (Gary Churchill)

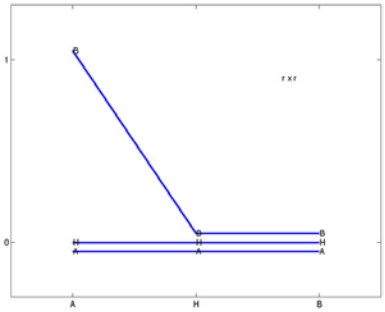
The allelic state at one locus can mask or uncover the effects of allelic variation at another.

- W. Bateson, 1907.

epistasis in parallel pathways (GAC)

- Z keeps trait value low
- neither E_1 nor E_2 is rate limiting
- loss of function alleles are segregating from parent A at E_1 and from parent B at E_2



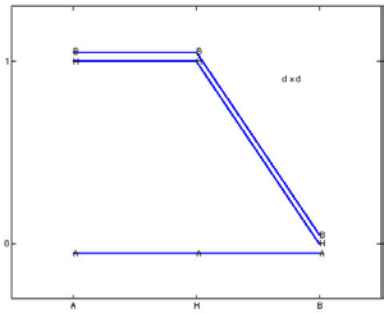


epistasis in a serial pathway (GAC)

- Z keeps trait value high

$$X \xrightarrow{E_1} Y \xrightarrow{E_2} Z$$

- neither E_1 nor E_2 is rate limiting
- loss of function alleles are segregating from parent B at E_1 and from parent A at E_2



QTL with epistasis

- same phenotype model overview

$$y = \mu_q + e, \text{var}(e) = \sigma^2$$

- partition of genotypic value with epistasis

$$\mu_q = \mu + \beta_{q1} + \beta_{q2} + \beta_{q12}$$

- partition of genetic variance & heritability

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

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma^2} = h_1^2 + h_2^2 + h_{12}^2$$

epistatic interactions

- model space issues
  - 2-QTL interactions only?
    - or general interactions among multiple QTL?
  - partition of effects
    - Fisher-Cockerham or tree-structured or ?
- 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 *et al.* (2005) *Genetics*

limits of epistatic inference

- power to detect effects
  - epistatic model size grows exponentially
    - $|H| = 3^{nqtl}$ for general interactions
  - power depends on ratio of n to model size
    - want $n / |H|$ to be fairly large (say > 5)
    - $n = 100$, $nqtl = 3$, $n / |H| \approx 4$
- 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

6. 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

Bayes factors & BIC

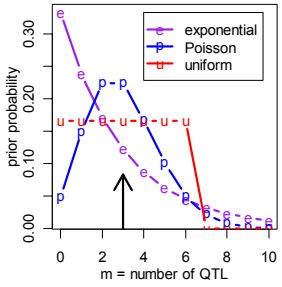
$$B_{12} = \frac{\text{pr}(\text{ model}_1 \mid Y) / \text{pr}(\text{ model}_2 \mid Y)}{\text{pr}(\text{ model}_1) / \text{pr}(\text{ model}_2)} = \frac{\text{pr}(Y \mid \text{model}_1)}{\text{pr}(Y \mid \text{model}_2)}$$

- 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)

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

Bayes factors and genetic model H

- H = number of QTL
  - prior pr(H) chosen by user
  - posterior pr($H|y,m$)
    - sampled marginal histogram
    - shape affected by prior pr(H)



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

- pattern of QTL across genome
- gene action and epistasis

issues in computing Bayes factors

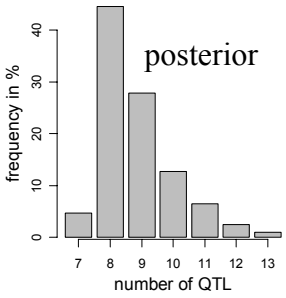
- *BF* insensitive to shape of prior on *nqtl*
  - geometric, Poisson, uniform
  - precision improves when prior mimics posterior
- *BF* sensitivity to prior variance on effects θ
  - prior variance should reflect data variability
  - resolved by using hyper-priors
    - automatic algorithm; no need for user tuning
- easy to compute Bayes factors from samples
  - sample posterior using MCMC
  - posterior pr(*nqtl*|*y*,*m*) is marginal histogram

examples in detail

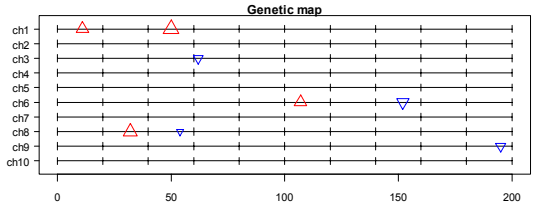
- simulation study (after Stephens & Fisch (1998)
- days to flower for *Brassica napus* (plant) (n = 108)
  - single chromosome with 2 linked loci
  - whole genome
- gonad shape in *Drosophila* spp. (insect) (n = 1000)
  - multiple traits reduced by PC
  - many QTL and epistasis
- expression phenotype (SCD1) in mice (n = 108)
  - multiple QTL and epistasis
- obesity in mice (n = 421)
  - epistatic QTLs with no main effects

simulation with 8 QTL

- 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

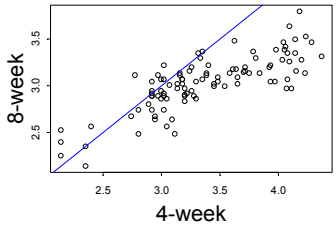
- notice which chromosomes have persistent loci
- best pattern found 42% of the time

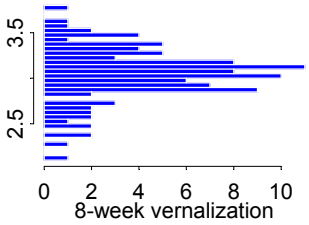
	Chromosome										
m	**1**	2	**3**	4	5	**6**	7	**8**	**9**	10	**Count of 8000**
8	**2**	**0**	**1**	**0**	**0**	**2**	**0**	**2**	**1**	**0**	3371
9	*3*	0	1	0	0	2	0	2	1	0	751
7	2	0	1	0	0	2	0	*1*	1	0	377
9	2	0	1	0	0	2	0	2	1	0	218
9	2	0	1	0	0	*3*	0	2	1	0	218
9	2	0	1	0	0	2	0	2	*2*	0	198

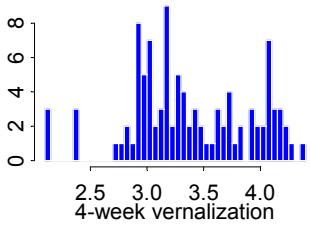
Brassica napus: 1 chromosome

- 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



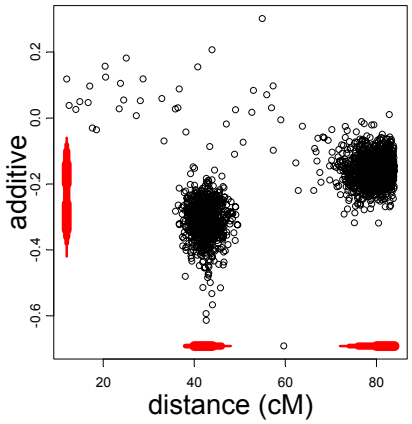




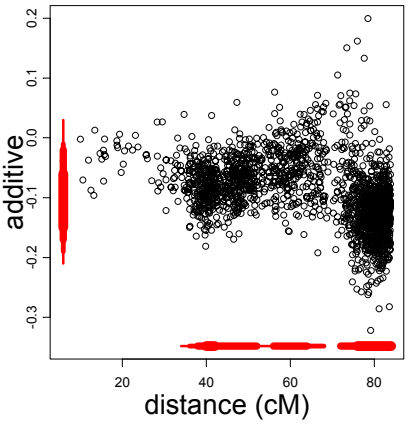
summaries of raw data
joint scatter plots
(identity line)
separate histograms

Brassica credible regions

4-week



8-week



B. napus 8-week vernalization whole genome study

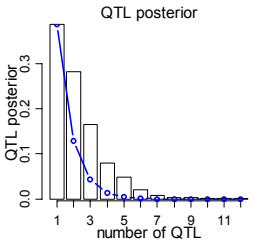
- 108 plants from double haploid
  - similar genetics to backcross: follow 1 gamete
  - parents are Major (biennial) and Stellar (annual)
- 300 markers across genome
  - 19 chromosomes
  - average 6cM between markers
    - median 3.8cM, max 34cM
  - 83% markers genotyped
- phenotype is days to flowering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

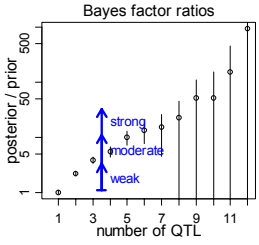
Bayesian model assessment

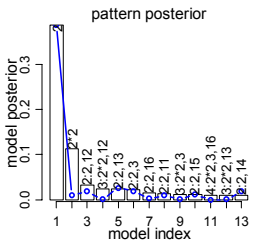
row 1: # QTL
row 2: pattern

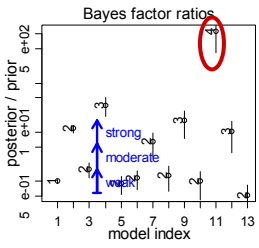
col 1: posterior
col 2: Bayes factor
note error bars on bf

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



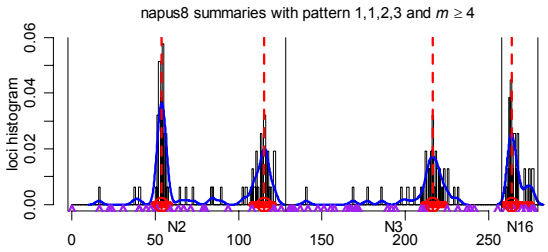




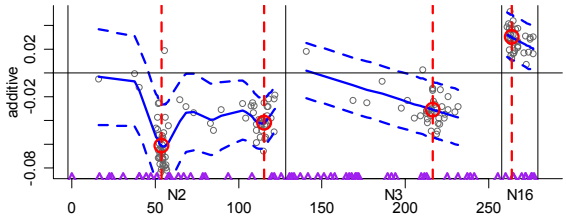


Bayesian estimates of loci & effects

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

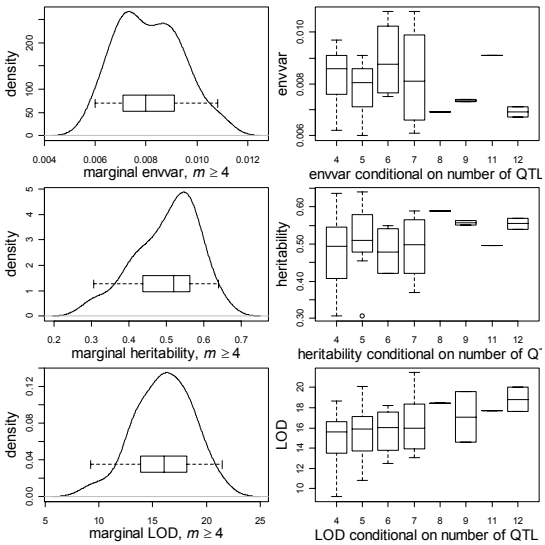


Bayesian model diagnostics

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



shape phenotype in BC study indexed by PC1

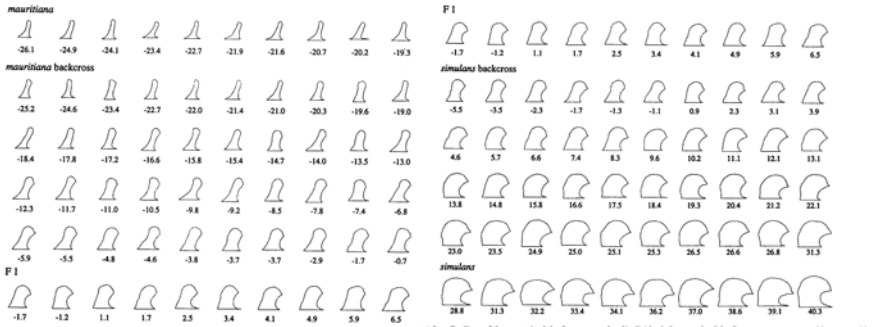


FIGURE 6.—Outlines of the posterior lobe from a sample of individuals from each of the five groups: pure *mauritiana*, *mauritiana* backcross, F_1, *simulans* backcross, and pure *simulans*. Within each group, the outlines are presented in order of their PC1 score (sampled at even intervals from the range of variation). The number below each specimen is its PC1 score. The outlines are drawn to scale with the origin at the centroid of each outline and with all baselines parallel.

Liu et al. (1996) *Genetics*

shape phenotype via PC

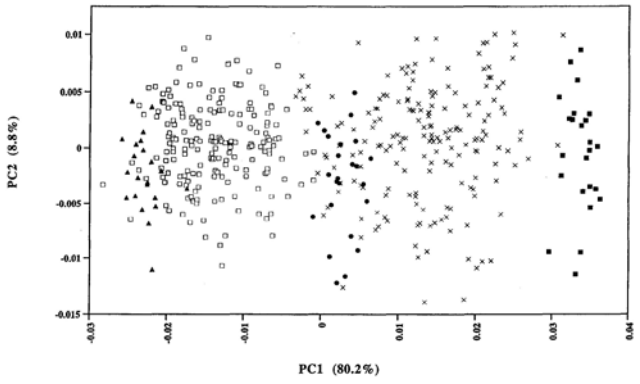


FIGURE 5.—A plot of the first two principal components of the Fourier coefficients from posterior lobe outlines. Many individuals from each of five genotypic classes are represented. Each point represents an average of scores from the left and right sides of an individual (with a few exceptions for which the score is from one side only). The percentage of variation in the Fourier coefficients accounted for by each principal component is given in parentheses. Liu et al. (1996) *Genetics*

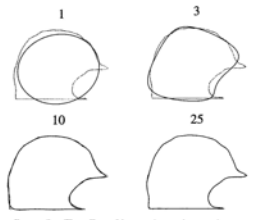


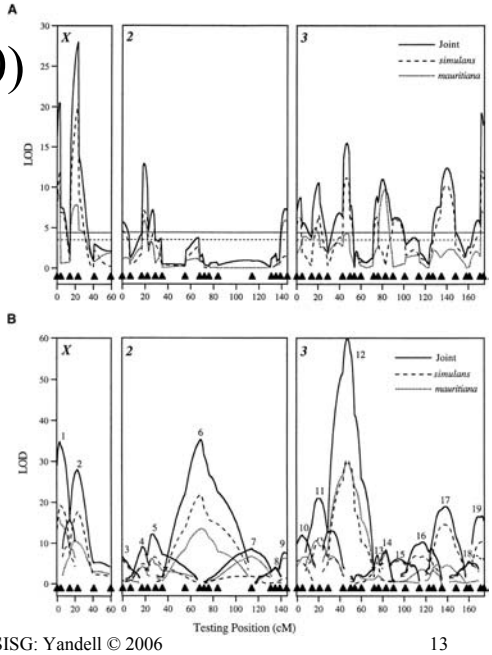
FIGURE 2.—The effect of harmonic number on the accuracy of reconstruction of a posterior lobe outline by elliptical Fourier analysis.

Zeng et al. (2000) CIM vs. MIM

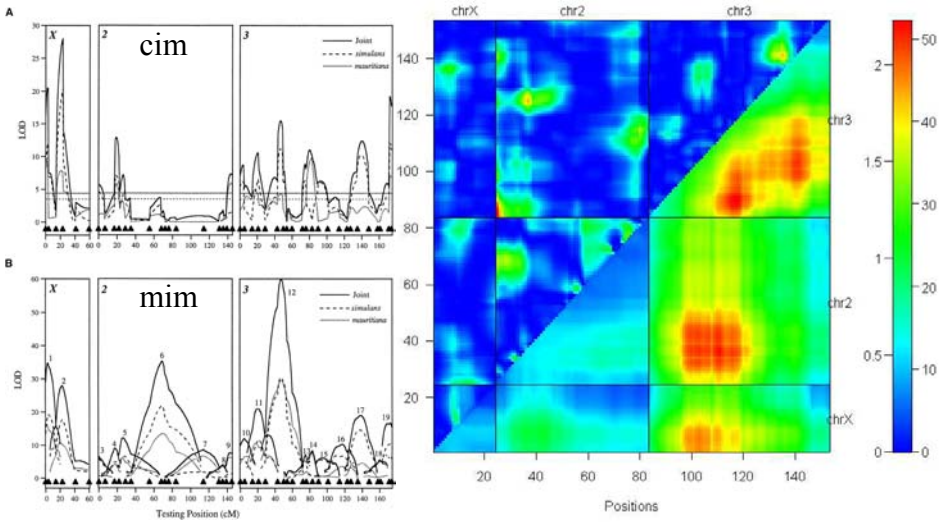
composite interval mapping
(Liu et al. 1996)
narrow peaks
miss some QTL

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

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

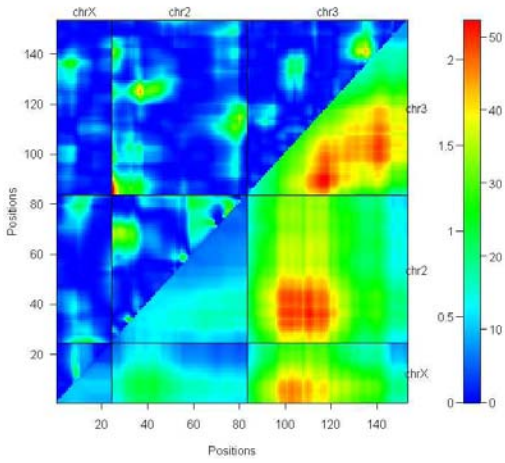


CIM, MIM and IM pairscan

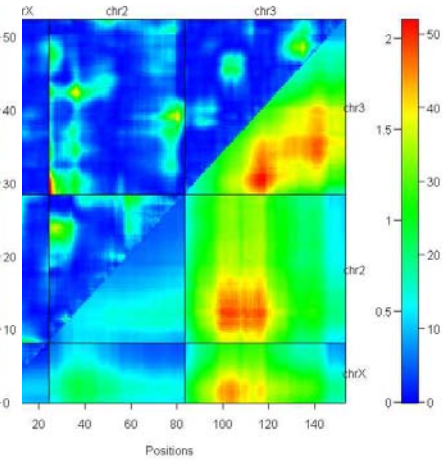


2 QTL + epistasis: IM versus multiple imputation

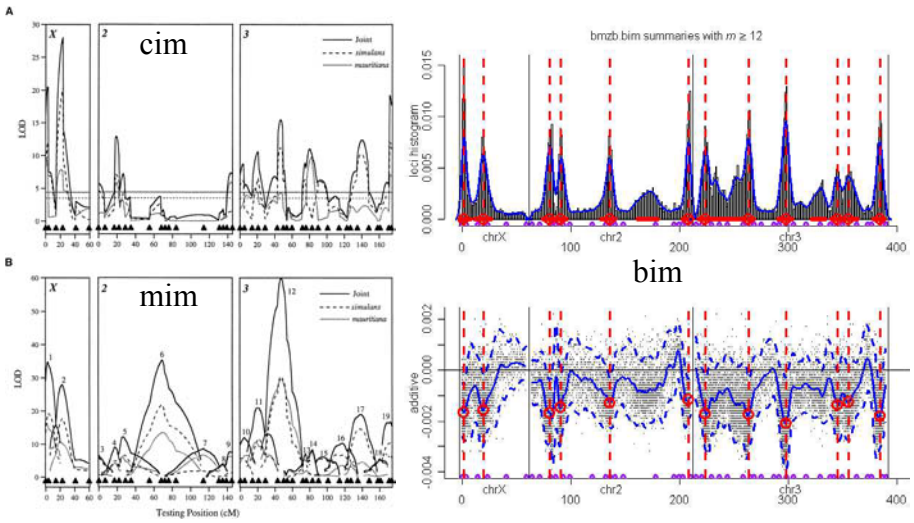
IM pairscan



multiple imputation



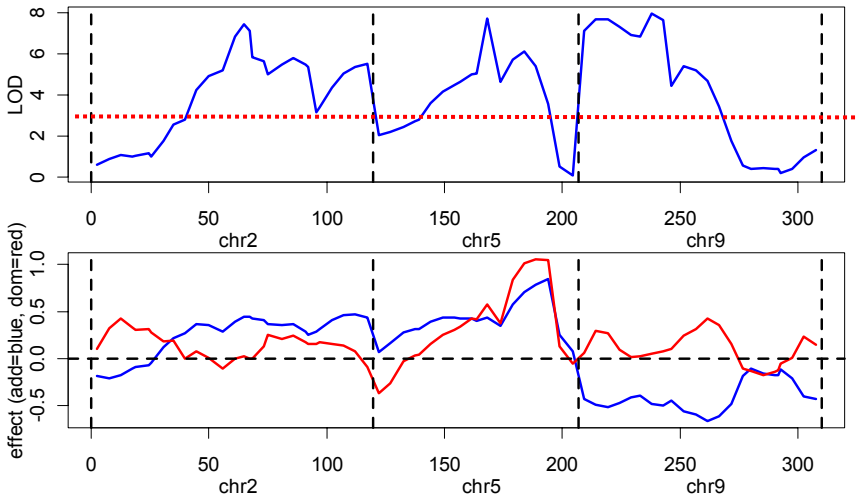
multiple QTL: CIM, MIM and BIM



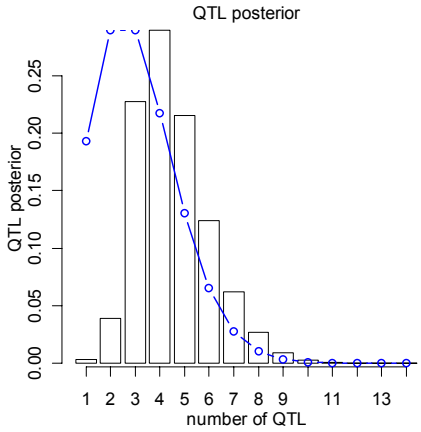
studying diabetes in an F2

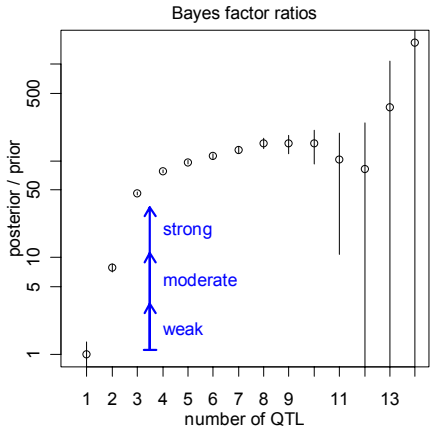
- segregating cross of inbred lines
  - B6.ob x BTBR.ob → F1 → F2
  - selected mice with ob/ob alleles at leptin gene (chr 6)
  - measured and mapped body weight, insulin, glucose at various ages (Stoehr et al. 2000 Diabetes)
  - sacrificed at 14 weeks, tissues preserved
- gene expression data
  - Affymetrix 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 (QTLCart)
SCD1: multiple QTL plus epistasis!

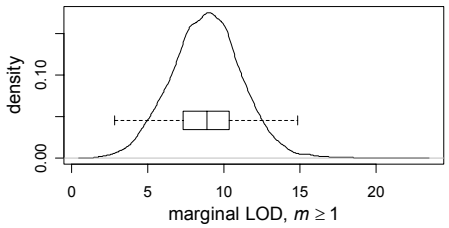


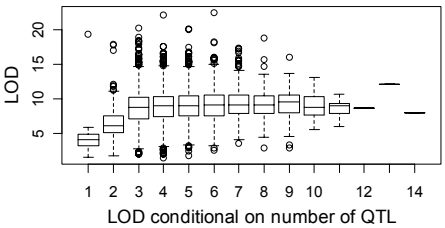
Bayesian model assessment: number of QTL for SCD1

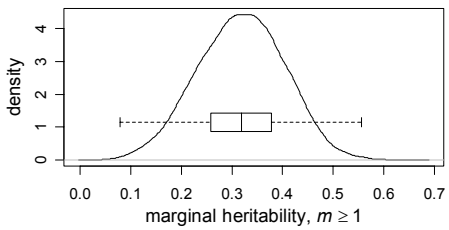


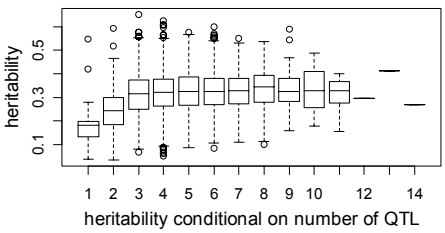


Bayesian LOD and h^2 for SCD1

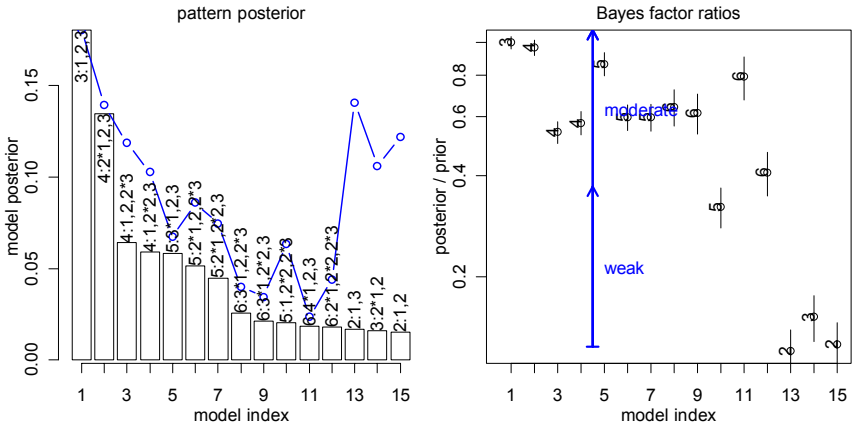






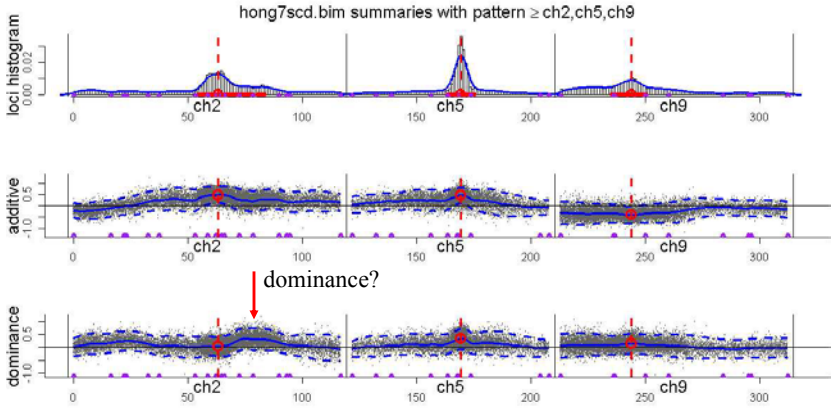


Bayesian model assessment: chromosome QTL pattern for SCD1

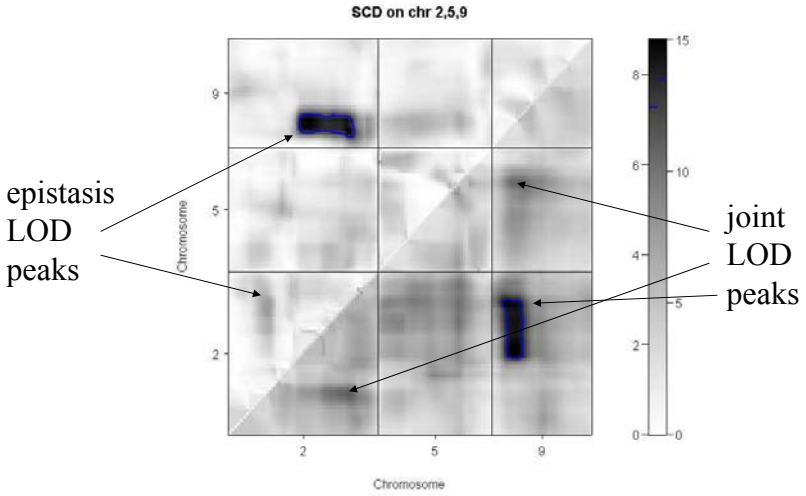


trans-acting QTL for SCD1

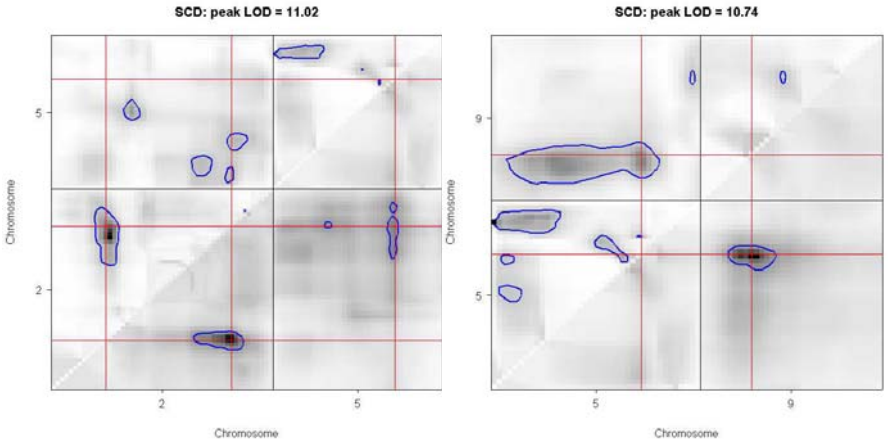
(no epistasis yet: see Yi, Xu, Allison 2003)



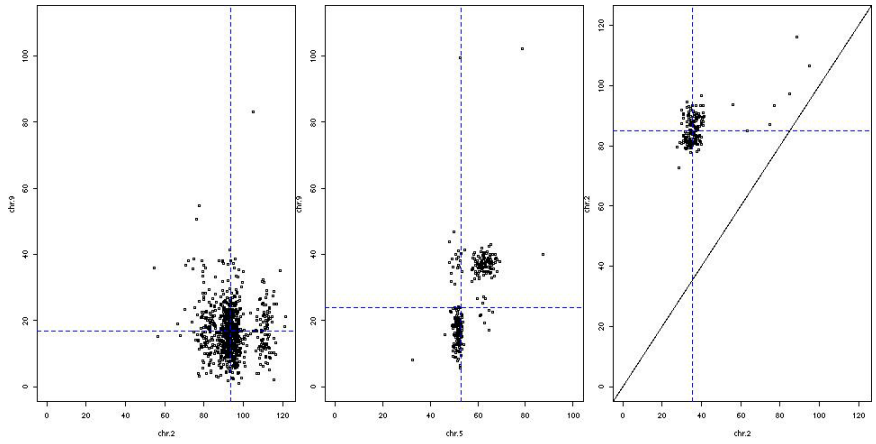
2-D scan: assumes only 2 QTL!



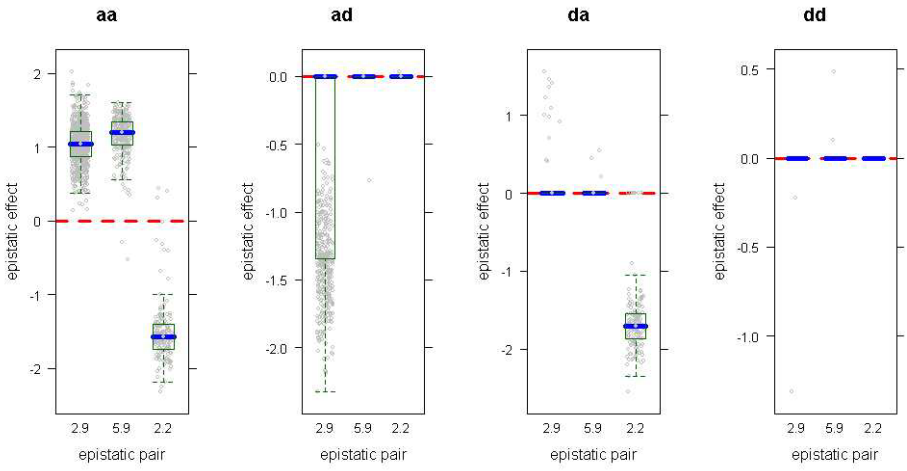
sub-peaks can be easily overlooked!



epistatic model fit



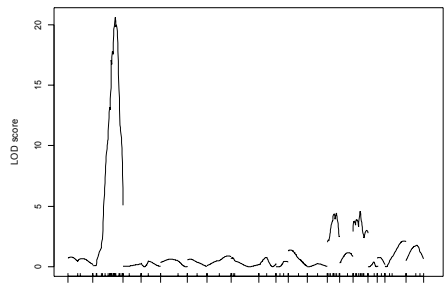
Cockerham epistatic effects



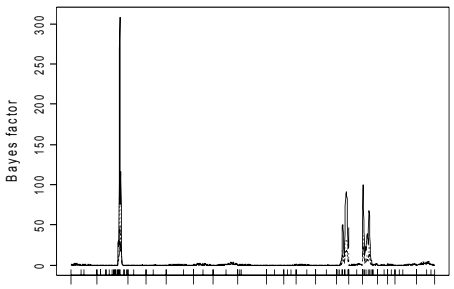
obesity in CAST/Ei BC onto M16i

- 421 mice (Daniel Pomp)
  - (213 male, 208 female)
- 92 microsatellites on 19 chromosomes
  - 1214 cM map
- subcutaneous fat pads
  - pre-adjusted for sex and dam effects
- Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005) *Genetics* (in press)

non-epistatic analysis

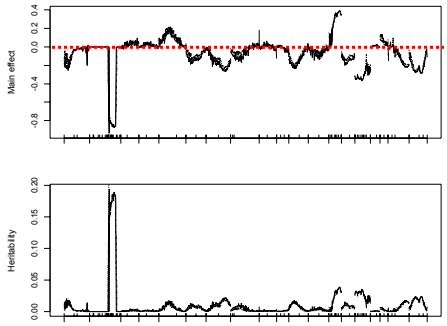


single QTL LOD profile

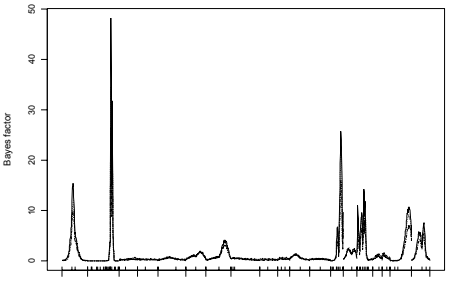


multiple QTL
Bayes factor profile

posterior profile of main effects in epistatic analysis

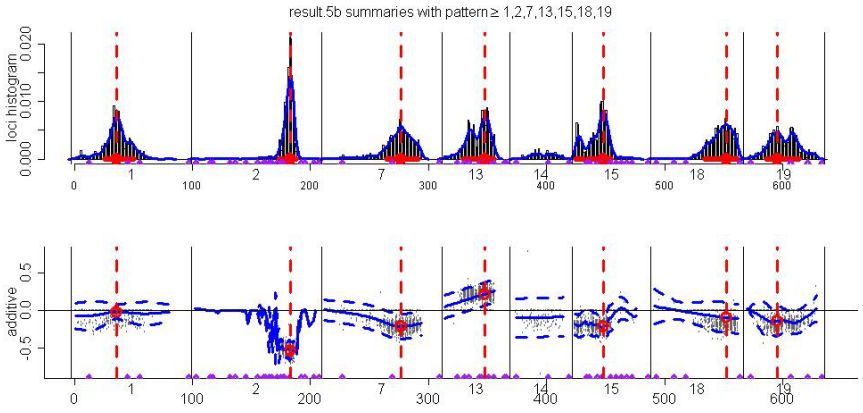


main effects & heritability profile

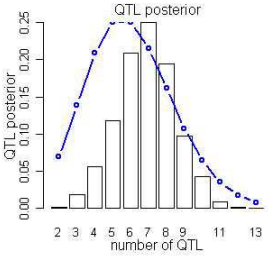


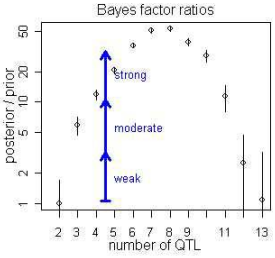
Bayes factor profile

posterior profile of main effects in epistatic analysis



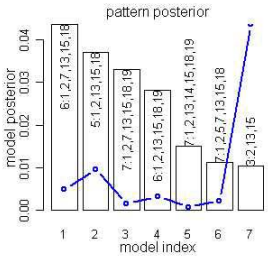
model selection via Bayes factors for epistatic model

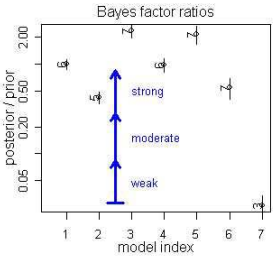




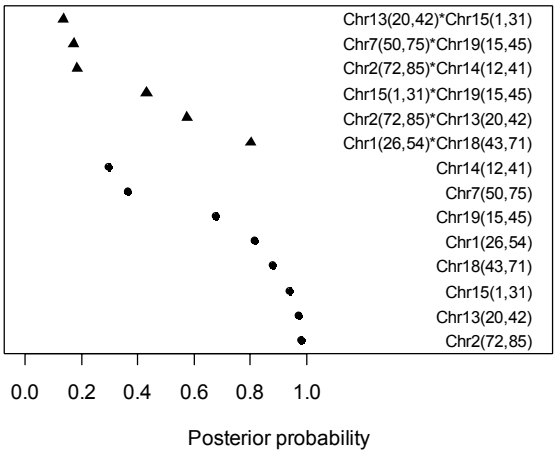
number of QTL

QTL pattern

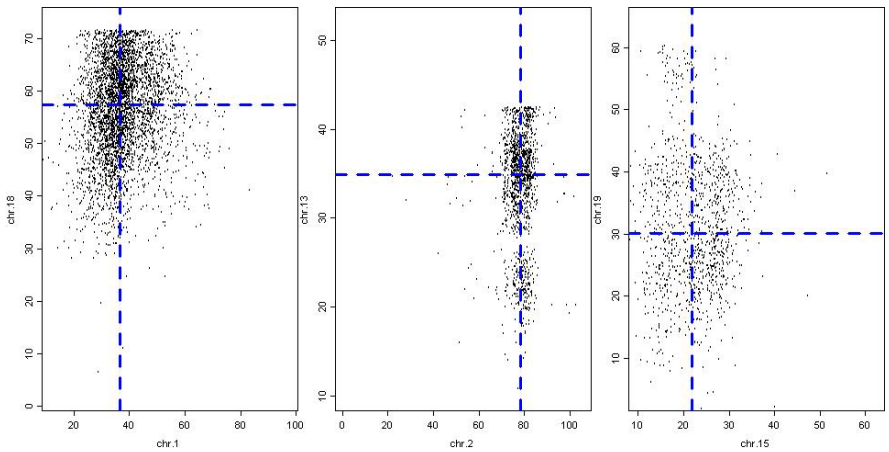




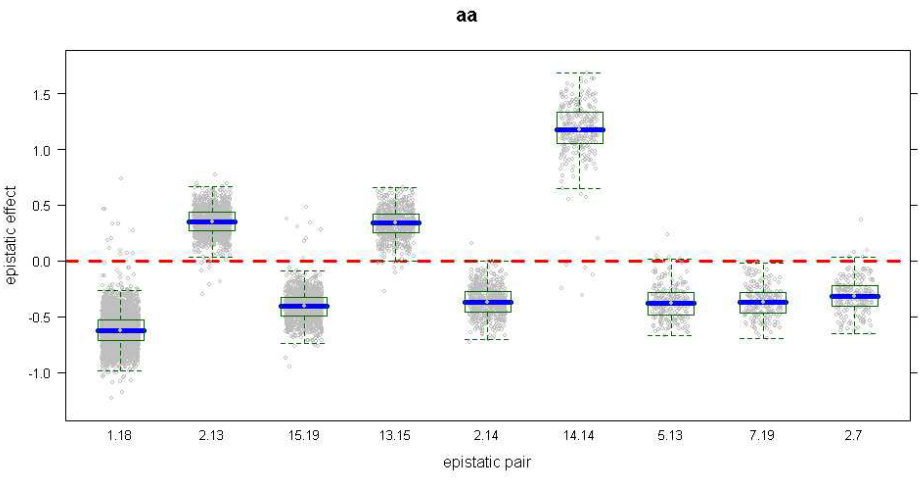
posterior probability of effects



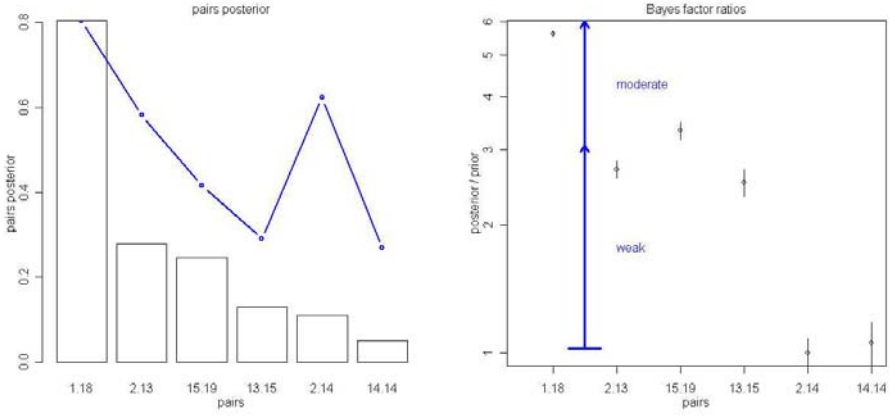
scatterplot estimates of epistatic loci



stronger epistatic effects



model selection for pairs



our RJ-MCMC software

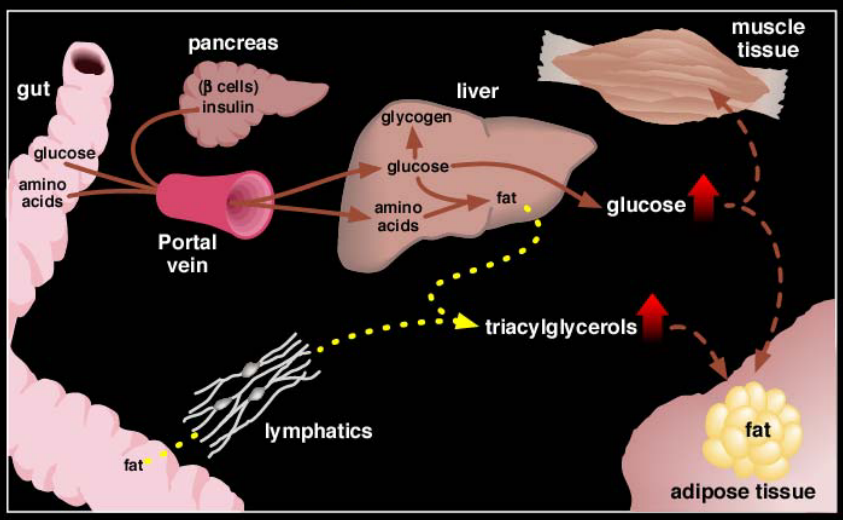
- R: www.r-project.org
  - freely available statistical computing application R
  - library(bim) builds on Broman's library(qtl)
- QTLCart: statgen.ncsu.edu/qtlcart
  - Bmapqtl incorporated into QTLCart (S Wang 2003)
- www.stat.wisc.edu/~yandell/qtl/software/bmqtl
- R/bim
  - initially designed by JM Satagopan (1996)
  - major revision and extension by PJ Gaffney (2001)
    - whole genome, multivariate and long range updates
    - speed improvements, pre-burnin
  - built as official R library (H Wu, Yandell, Gaffney, CF Jin 2003)
- R/bmqtl
  - collaboration with N Yi, H Wu, GA Churchill
  - initial working module: Winter 2005
  - improved module and official release: Summer/Fall 2005
  - major NIH grant (PI: Yi)

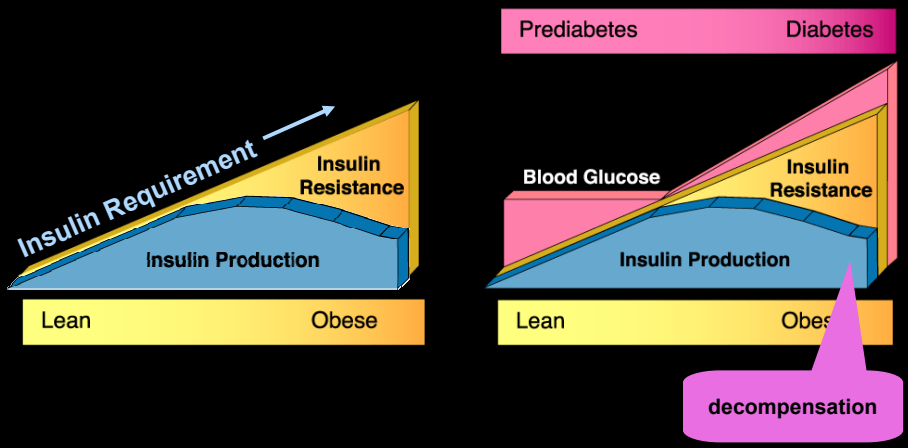
Multiple Traits & Microarrays

1. why study multiple traits together?

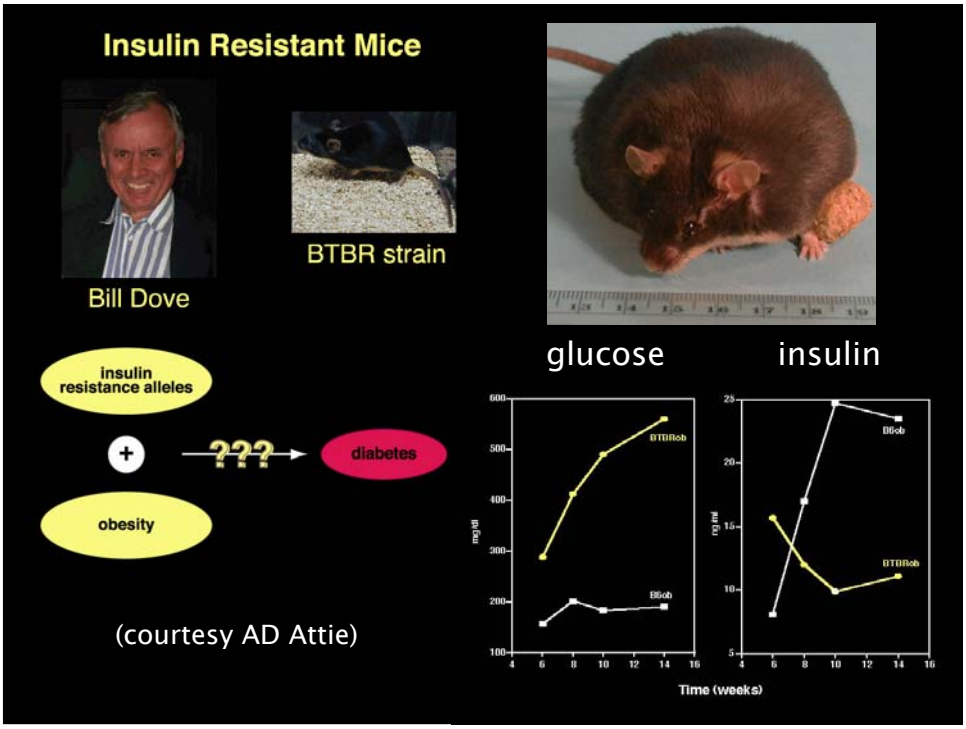
- avoid reductionist approach to biology
  - address physiological/biochemical mechanisms
  - Schmalhausen (1942); Falconer (1952)
- separate close linkage from pleiotropy
  - 1 locus or 2 linked loci?
- identify epistatic interaction or canalization
  - influence of genetic background
- establish QTL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

Type 2 Diabetes Mellitus





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



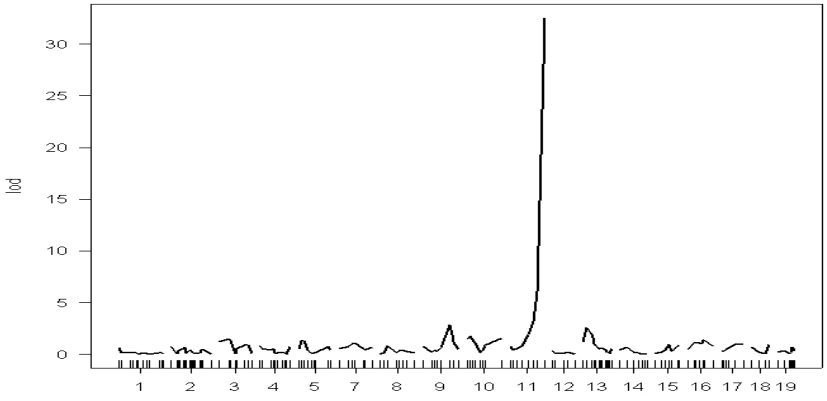
studying diabetes in an F2

- 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*)

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

- emphasize additive effects in F2
  - F2 design: 1QQ:2Qq:1qq
  - best design for additive only: 1QQ:1Qq
  - drop heterozygotes (Qq)
  - reduce sample size by half with no power loss
- emphasize general effects in F2
  - best design: 1QQ:1Qq:1qq
  - drop half of heterozygotes (25% reduction)
- multiple loci
  - same idea but care is needed
  - drop 7/16 of sample for two unlinked loci

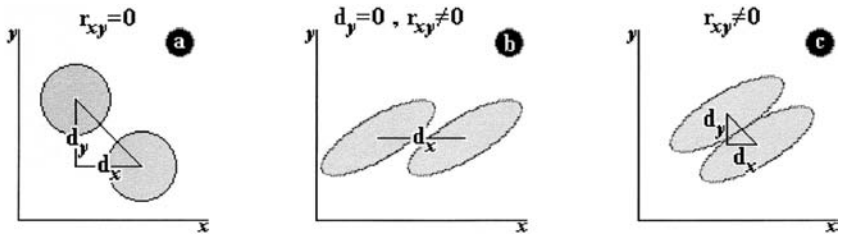
is this relevant to large QTL studies?

- 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

interplay of pleiotropy & correlation



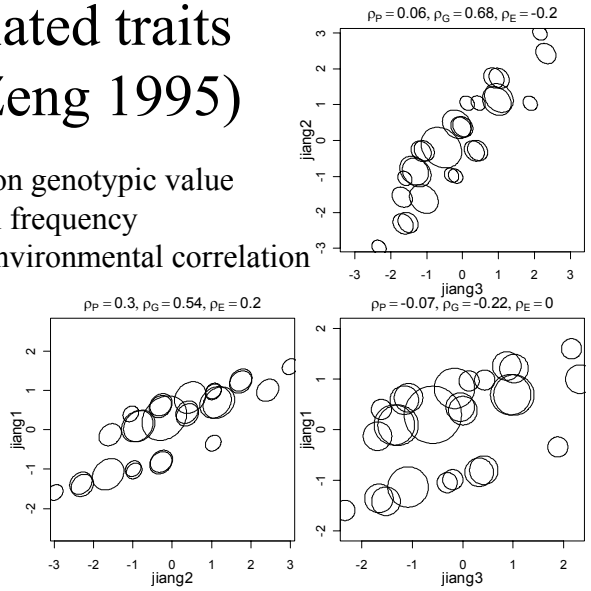
pleiotropy only correlation only both

Korol et al. (2001)

3 correlated traits (Jiang Zeng 1995)

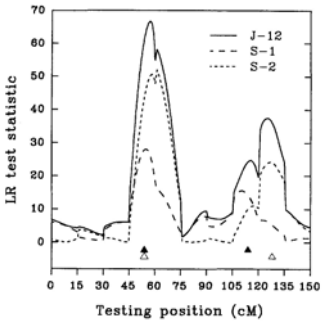
ellipses centered on genotypic value
width for nominal frequency
main axis angle environmental correlation
3 QTL, F2
27 genotypes

note signs of genetic and environmental correlation



pleiotropy or close linkage?

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



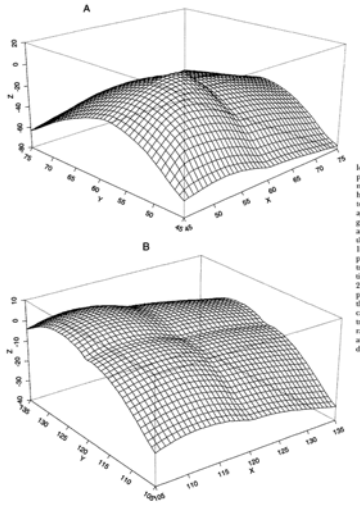


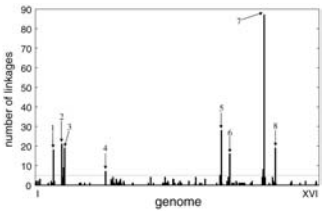
FIGURE 2.—Two-dimensional log-likelihood surfaces (expressed as deviations from the maximum of the log-likelihoods on the diagonal) for the test of pleiotropy *vs.* close linkage are presented for two regions: the region between 45 and 75 cM of Figure 1 (A) and the region between 105 and 135 cM (B).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

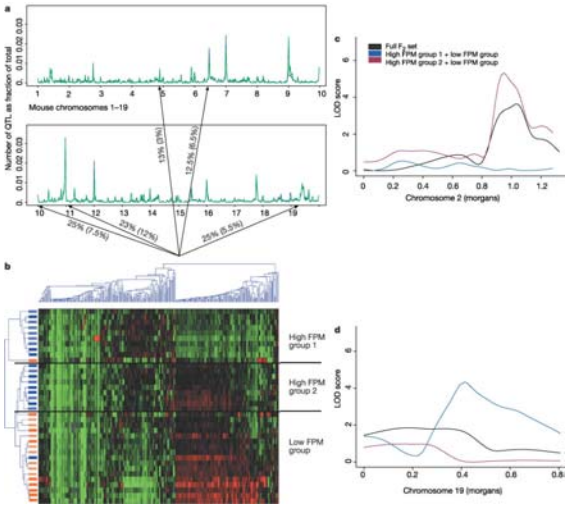
4. modern high throughput biology

- measuring the molecular dogma of biology
  - DNA → RNA → protein → metabolites
  - measured one at a time only a few years ago
- massive array of measurements on whole systems ("omics")
  - thousands measured per individual (experimental unit)
  - all (or most) components of system measured simultaneously
    - whole genome of DNA: genes, promoters, etc.
    - all expressed RNA in a tissue or cell
    - all proteins
    - all metabolites
- systems biology: focus on network interconnections
  - chains of behavior in ecological community
  - underlying biochemical pathways
- genetics as one experimental tool
  - perturb system by creating new experimental cross
  - each individual is a unique mosaic

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

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





finding heritable traits

(from Christina Kendziorski)

- reduce 30,000 traits to 300-3,000 heritable traits
- probability a trait is heritable

  $\text{pr}(H|Y,Q) = \text{pr}(Y|Q,H)\ \text{pr}(H|Q) / \text{pr}(Y|Q)$ Bayes rule

  $\text{pr}(Y|Q) = \text{pr}(Y|Q,H)\ \text{pr}(H|Q) + \text{pr}(Y|Q, \text{not } H)\ \text{pr}(\text{not } H|Q)$
- phenotype averaged over genotypic mean μ

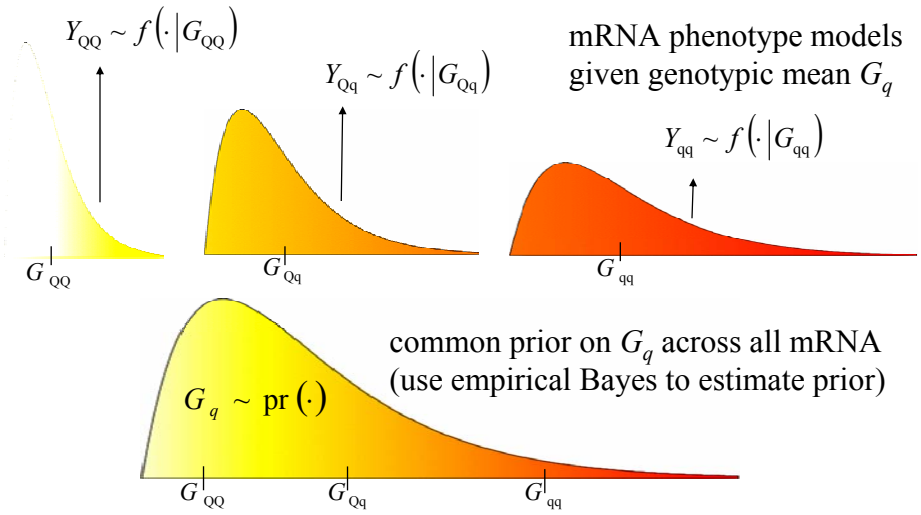
  $\text{pr}(Y|Q, \text{not } H) = f_0(Y) = \int f(Y|G)\ \text{pr}(G)\ dG$ if not H

  $\text{pr}(Y|Q, H) = f_1(Y|Q) = \prod_q f_0(Y_q)$ if heritable

  $Y_q = \{Y_i \mid Q_i = q\}$ = trait values with genotype $Q=q$

hierarchical model for expression phenotypes

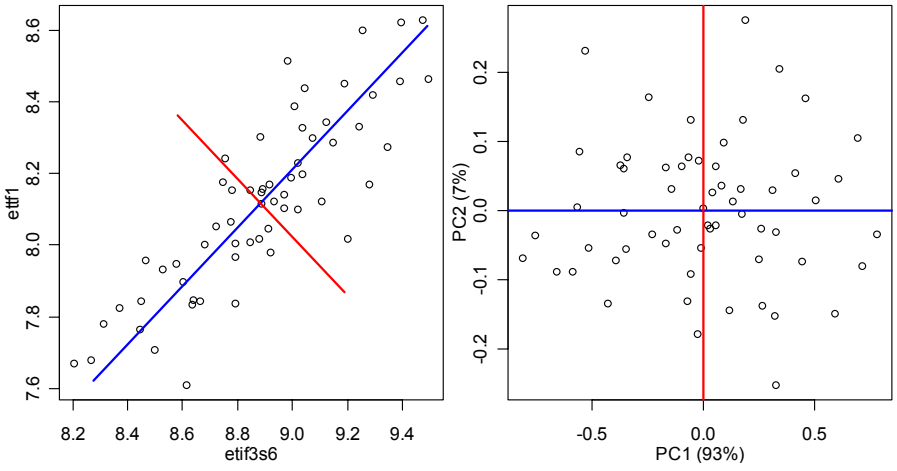
(EB arrays: Christina Kendziorski)



expression meta-traits: pleiotropy

- reduce 3,000 heritable traits to 3 meta-traits(!)
- what are expression meta-traits?
  - pleiotropy: a few genes can affect many traits
    - transcription factors, regulators
  - weighted averages: $Z = YW$
    - principle components, discriminant analysis
- infer genetic architecture of meta-traits
  - model selection issues are subtle
    - missing data, non-linear search
    - what is the best criterion for model selection?
  - time consuming process
    - heavy computation load for many traits
    - subjective judgement on what is best

PC for two correlated mRNA



PC across microarray functional groups

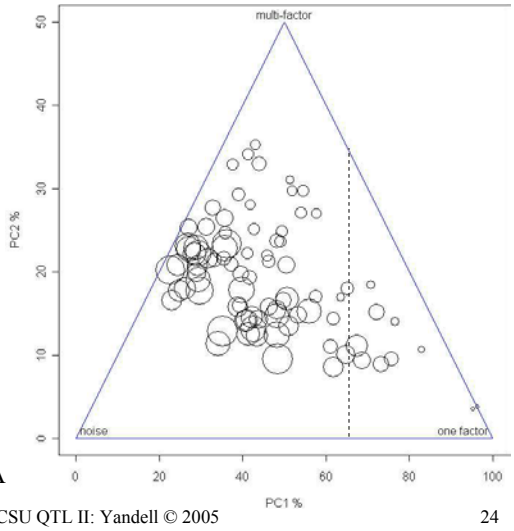
Affy chips on 60 mice
~40,000 mRNA

2500+ mRNA show DE
(via EB arrays with
marker regression)

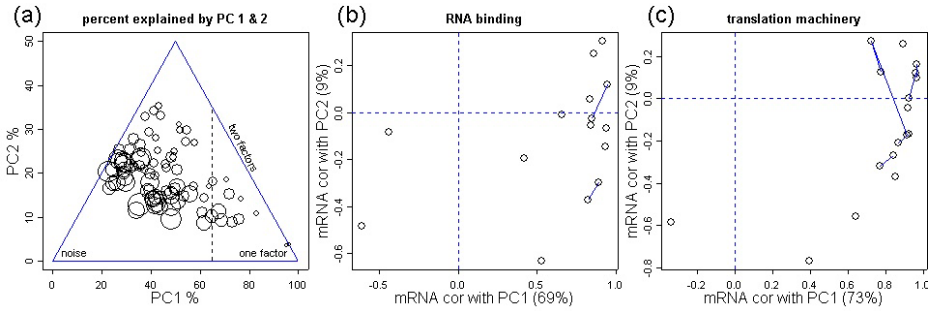
1500+ organized in
85 functional groups
2-35 mRNA / group

which are interesting?
examine PC1, PC2

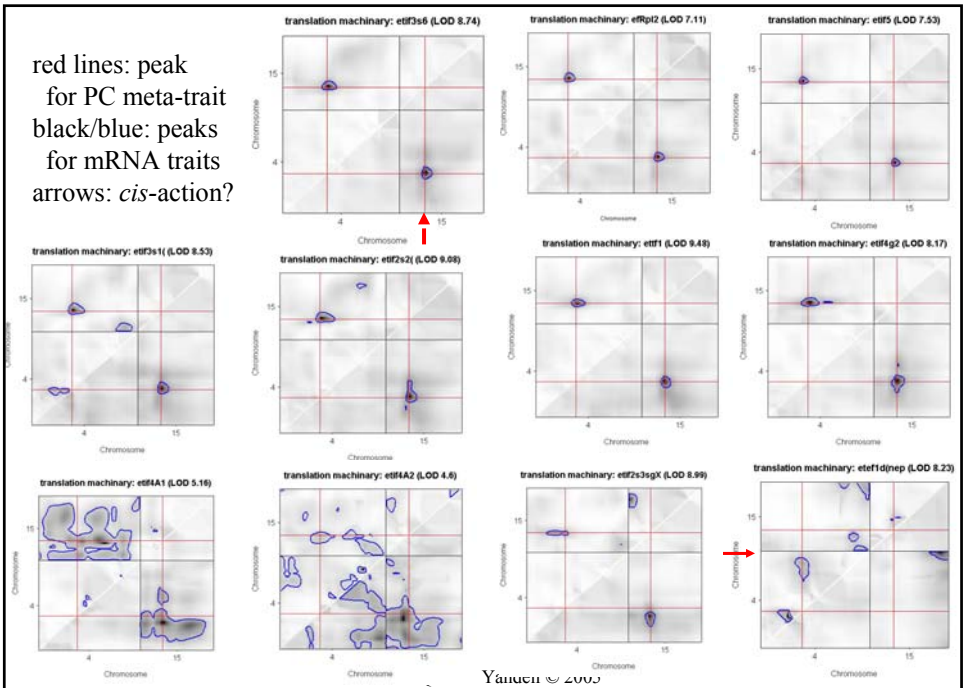
circle size = # unique mRNA



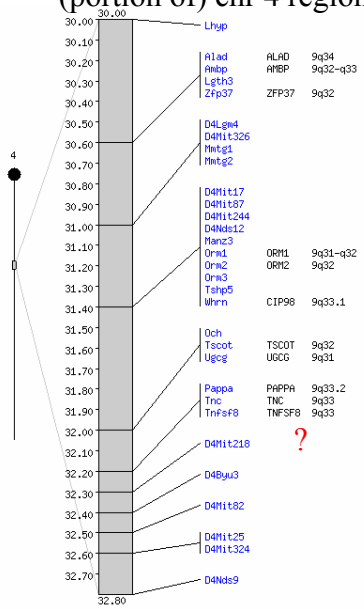
84 PC meta-traits by functional group focus on 2 interesting groups



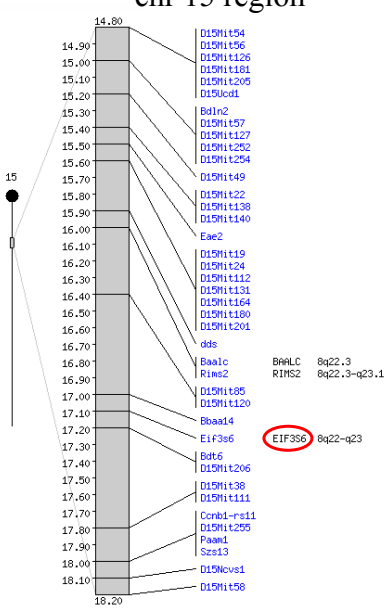
red lines: peak for PC meta-trait
black/blue: peaks for mRNA traits
arrows: *cis*-action?



(portion of) chr 4 region



chr 15 region



interaction plots for DA meta-traits

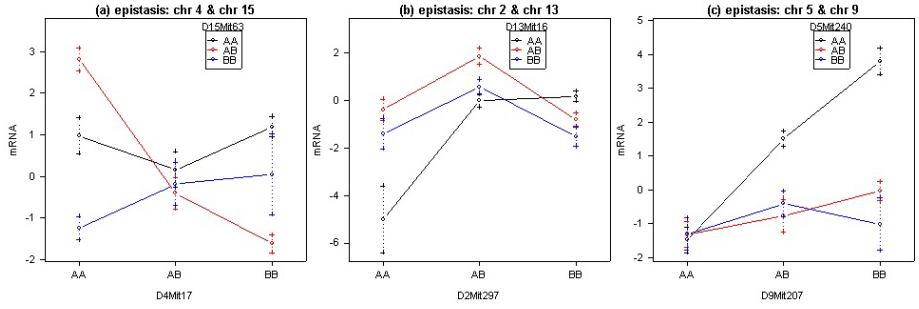
DA for all pairs of markers:

separate 9 genotypes based on markers

(a) same locus pair found with PC meta-traits

(b) Chr 2 region interesting from biochemistry (Jessica Byers)

(c) Chr 5 & Chr 9 identified as important for insulin, SCD

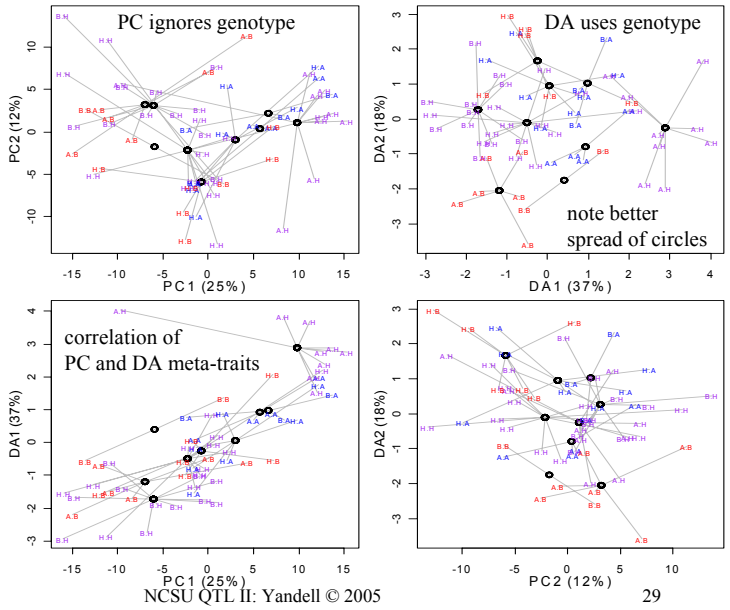


comparison of PC and DA meta-traits on 1500+ mRNA traits

genotypes from Chr 4/Chr 15 locus pair (circle=centroid)

PC captures spread without genotype

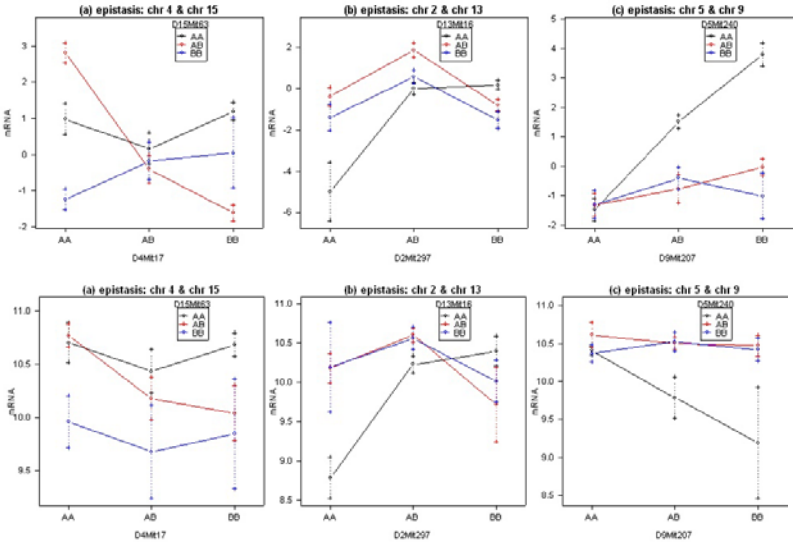
DA creates best separation by genotype



relating meta-traits to mRNA traits

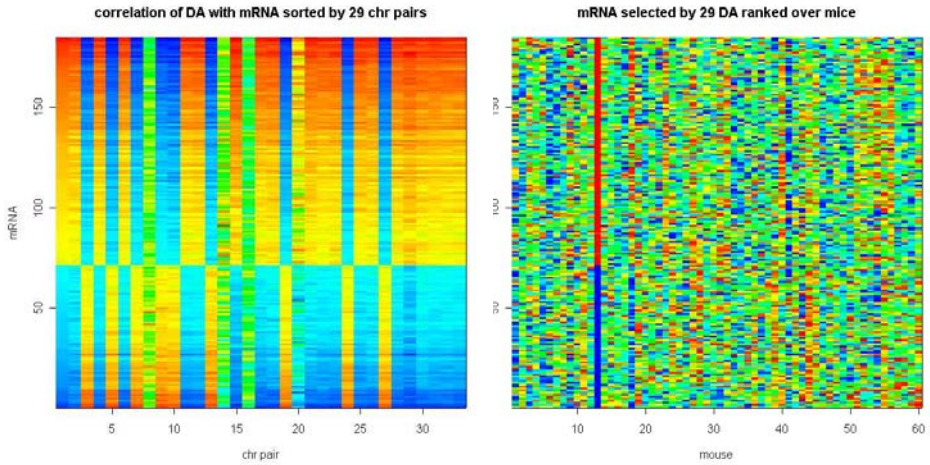
DA meta-trait standard units

SCD trait log2 expression



DA: a cautionary tale

(184 mRNA with |cor| > 0.5; mouse 13 drives heritability)



building graphical models

- infer genetic architecture of meta-trait
  - $E(Z \mid Q, M) = \mu_q = \beta_0 + \sum_{\{q \text{ in } M\}} \beta_{qk}$
- find mRNA traits correlated with meta-trait
  - $Z \approx \underline{YW}$ for modest number of traits $\underline{Y}$
- extend meta-trait genetic architecture
  - $\underline{M}$ = genetic architecture for $\underline{Y}$
  - expect subset of QTL to affect each mRNA
  - may be additional QTL for some mRNA

posterior for graphical models

•posterior for graph given multivariate trait & architecture

$\text{pr}(G \mid \underline{Y}, Q, \underline{M}) = \text{pr}(\underline{Y} \mid Q, G)\ \text{pr}(G \mid \underline{M}) / \text{pr}(\underline{Y} \mid Q)$

–$\text{pr}(G \mid \underline{M})$ = prior on valid graphs given architecture

•multivariate phenotype averaged over genotypic mean μ

$\text{pr}(\underline{Y} \mid Q, G) = f_1(\underline{Y} \mid Q, G) = \prod q\ f_0(\underline{Y}q \mid G)$

$f_0(\underline{Y}_q \mid G) = \int f(\underline{Y}_q \mid \underline{\mu}, G)\ \text{pr}(\underline{\mu})\ \text{d}\underline{\mu}$

•graphical model G implies correlation structure on $\underline{Y}$

•genotype mean prior assumed independent across traits

$\text{pr}(\underline{\mu}) = \prod_t \text{pr}(\mu_t)$

from graphical models to pathways

- build graphical models

  QTL → RNA1 → RNA2

  – class of possible models

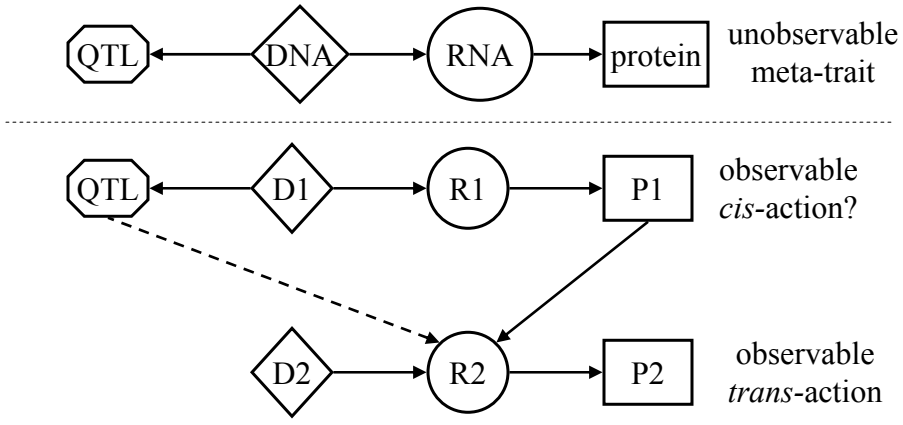
  – best model = putative biochemical pathway
- parallel biochemical investigation

  – candidate genes in QTL regions

  – laboratory experiments on pathway components

graphical models (with Elias Chaibub)

$f_1(\underline{Y} \mid Q, G=g) = f_1(Y_1 \mid Q)\ f_1(Y_2 \mid Q, Y_1)$



summary

- expression QTL are complicated
  - need to consider multiple interacting QTL
- coherent approach for high-throughput traits
  - identify heritable traits
  - dimension reduction to meta-traits
  - mapping genetic architecture
  - extension via graphical models to networks
- many open questions
  - model selection
  - computation efficiency
  - inference on graphical models