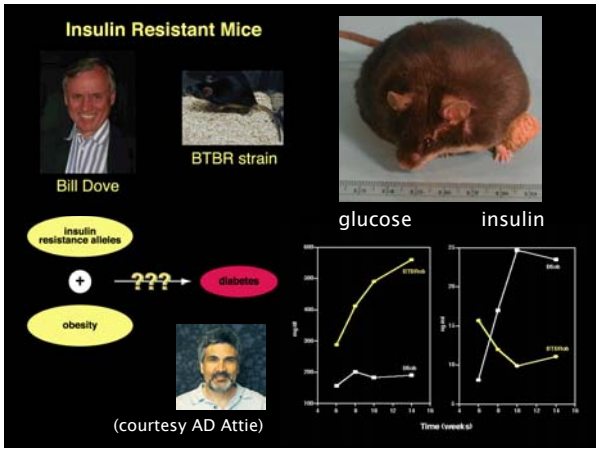
# Inferring Genetic Architecture of Complex Biological Processes

Brian S. Yandell[12], Christina Kendziorski[13], Hong Lan[4],
Elias Chaibub[1], Alan D. Attie[4]
U Alabama Birmingham

[1] Department of Statistics
[2] Department of Horticulture
[3] Department of Biostatistics & Medical Informatics
[4] Department of Biochemistry
University of Wisconsin-Madison

http://www.stat.wisc.edu/~yandell/statgen



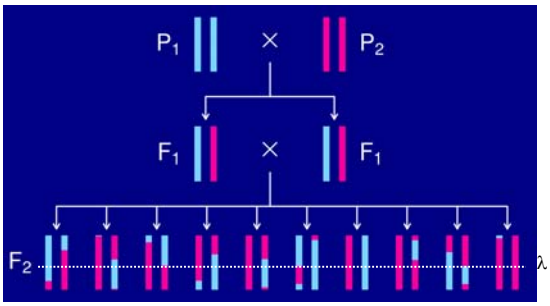



# studying diabetes in an F2

- mouse model: segregating panel from inbred lines
  - B6.ob x BTBR.ob → F1 → F2
  - selected mice with ob/ob alleles at leptin gene (Chr 6)
  - sacrificed at 14 weeks, tissues preserved
- physiological study (Stoehr et al. 2000 *Diabetes*)
  - mapped body weight, insulin, glucose at various ages
- gene expression studies
  - 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
    - U47 A & B chips, RMA normalization
    - design: selective phenotyping (Jin et al. 2004 *Genetics*)

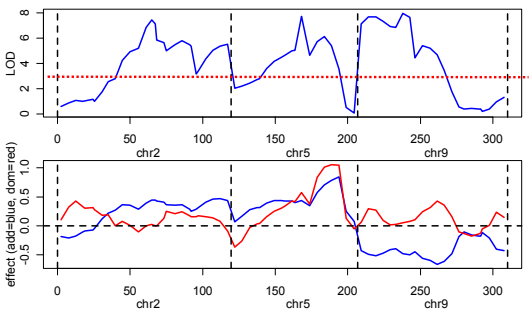


# The intercross (from K Broman)






# mRNA expression as phenotype: interval mapping for SCD1 is complicated



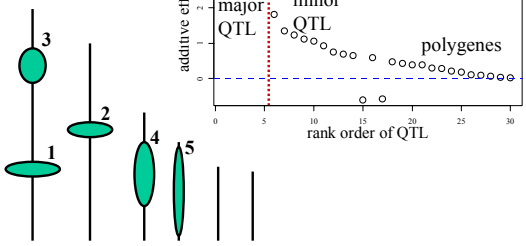



# taking a multiple QTL approach

- 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

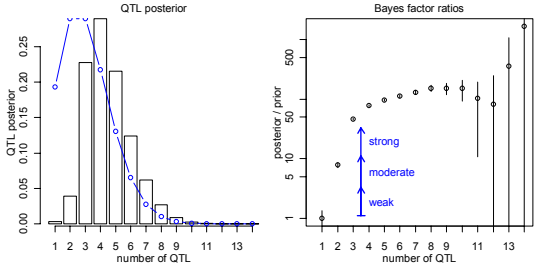
## Pareto diagram of QTL effects

major QTL on linkage map





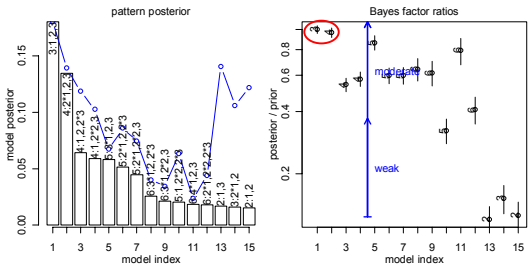

## Bayesian model assessment: number of QTL for SCD1 with R/bim






## Bayesian model assessment

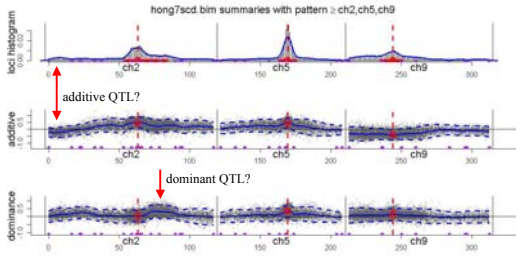
genetic architecture: chromosome pattern






## *trans*-acting QTL for SCD1

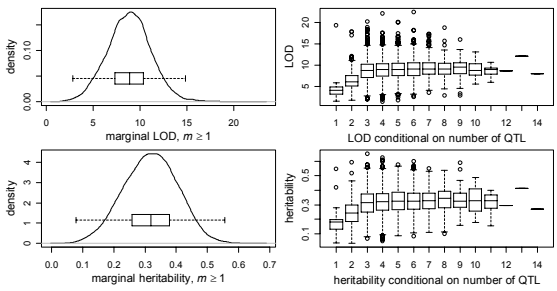
Bayesian model averaging with R/bim





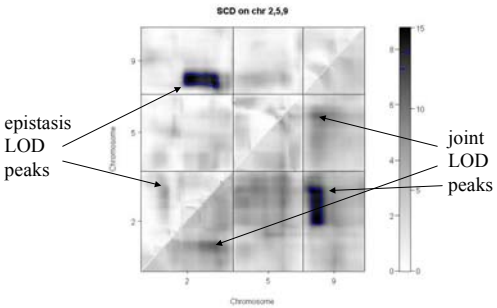

## Bayesian LOD and $h^2$ for SCD1

(summaries from R/bim)

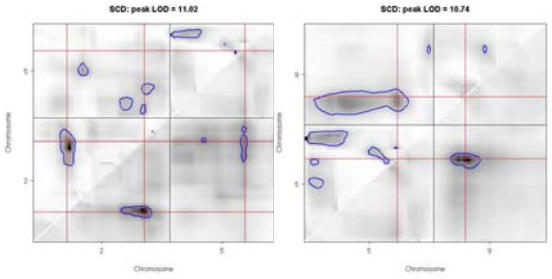





## SCD mRNA expression phenotype 2-D scan for QTL (R/qtl)



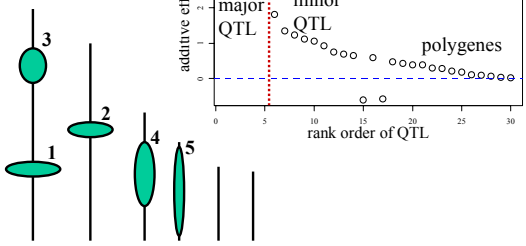
## sub-peaks can be easily overlooked






## hetereogeneity: many genes affect each trait

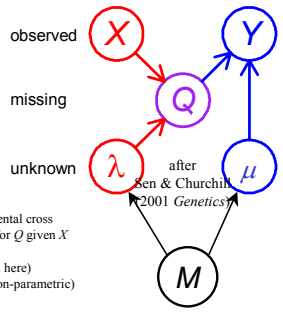
major QTL on linkage map






## interval mapping basics

- observed measurements
  - $Y$ = phenotypic trait
  - $X$ = 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
  - $M$ = genetic architecture
  - $\lambda$ = QT locus (or loci)
  - $\mu$ = phenotype model parameters
- pr($Q$=$q$|$X$,$\lambda$) genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for $Q$ given $X$
- $f(Y|\mu_q)$ phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters $\mu$ (could be non-parametric)






## genetic architecture: heterogeneity

- heterogeneity: many genes can affect phenotype
  - different allelic combinations can yield similar phenotypes
  - multiple genes can affecting phenotype in subtle ways
  - multiple genes can interact (epistasis)
- genetic architecture: model for explained genetic variation
  - loci (genomic regions) that affect trait
  - genotypic effects of loci, including possible epistasis

$M = \{\lambda_1, \lambda_2, \lambda_3, (\lambda_1, \lambda_2)\}$ = 3 loci with epistasis between two

$\mu_q = \beta_0 + \beta_{q1} + \beta_{q2} + \beta_{q3} + \beta_{q(1,2)}$ = linear model for genotypic mean

$\lambda = (\lambda_1, \lambda_2, \lambda_3)$ = loci in model $M$
$q = (q_1, q_2, q_3)$ = possible genotype at loci $\lambda$
$Q = (Q_1, Q_2, Q_3)$ = genotype for each individual at loci $\lambda$



## multiple QTL interval mapping

- genotypic mean depends on model $M$
  $\mu_q = \beta_0 + \Sigma_{\{q \text{ in } M\}} \beta_{qk}$
- interval mapping between flanking markers
  $f(Y \mid X, M) = \Sigma_q f(Y \mid \mu_q) f(Q = q \mid X, \lambda)$
- model selection
  - choice of distribution: $f$ is normal
  - sample many possible architectures
  - compare based on Bayes factors (BIC)



# 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



# 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 $\mu$

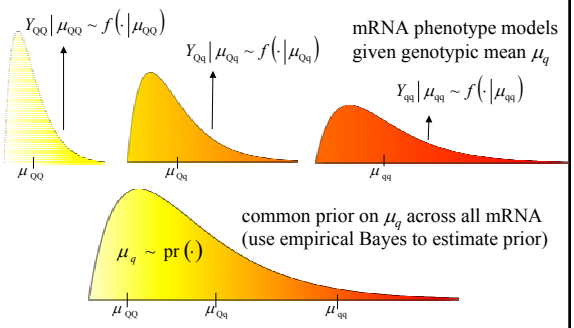
  $\text{pr}(Y|Q,\text{ not } H) = f_0(Y) = \int f(Y|\mu)\ \text{pr}(\mu)\ d\mu$ if not $H$

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



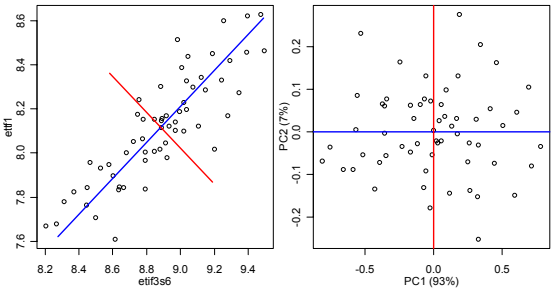



# expression meta-traits: pleiotropy

- 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

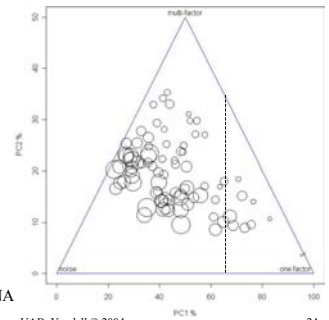
Affy chips on 60 mice
~40,000 mRNA

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

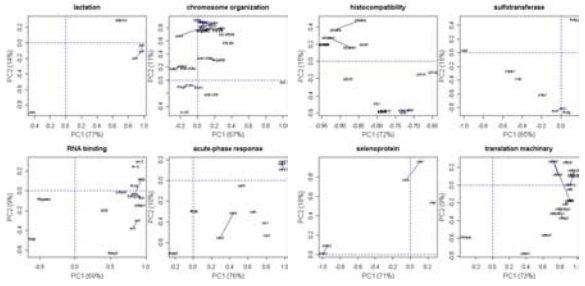
1500+ organized in
85 functional groups
2-35 mRNA / group

which are interesting?
examine PC1, PC2

circle size = # unique mRNA



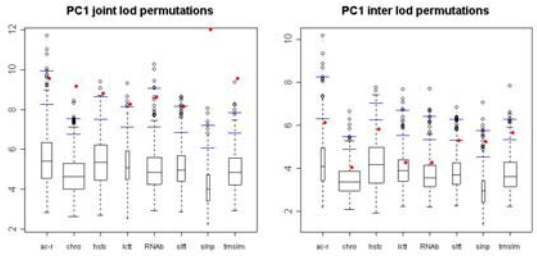
# factor loadings for PC1&2






# how well does PC1 do?

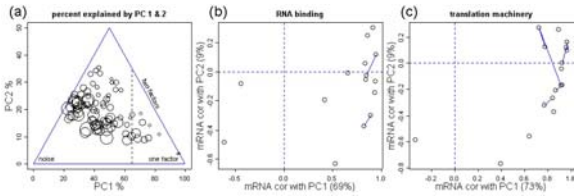
lod peaks for 2 QTL at best pair of chr
data (red) vs. 500 permutations (boxplots)




blue bars at 1%, 5%; width proportional to group size

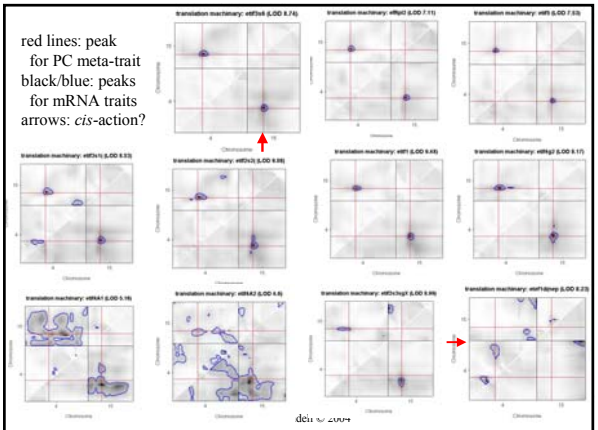


# 84 PC meta-traits by functional group
# focus on 2 interesting groups





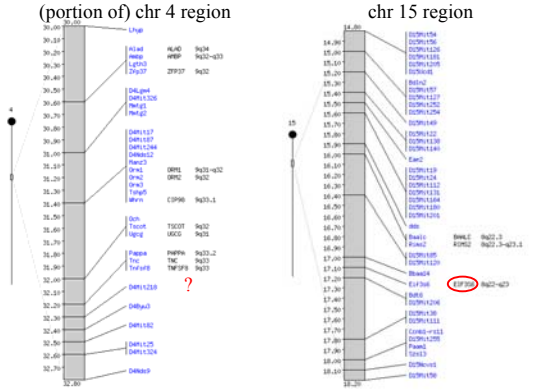

red lines: peak
for PC meta-trait
black/blue: peaks
for mRNA traits
arrows: *cis*-action?





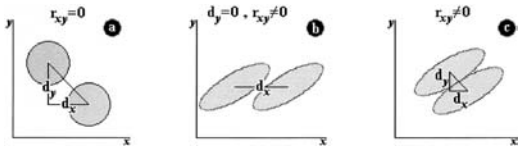
(portion of) chr 4 region

chr 15 region





# DA meta-traits: separate pleiotropy from environmental correlation



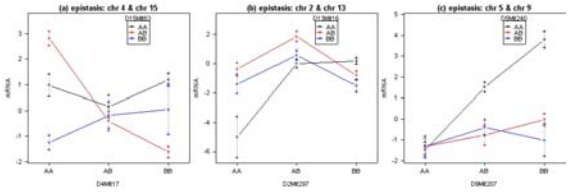

pleiotropy only

environmental correlation only

both
Korol et al. (2001)

# 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



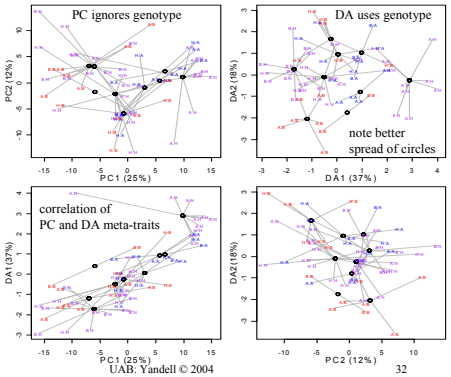


# comparison of PC and DA meta-traits on 1500+ mRNA traits

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

PC captures spread without genotype

DA creates best separation by genotype

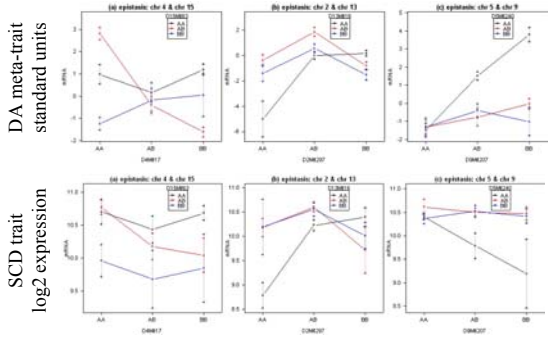





# relating meta-traits to mRNA traits

DA meta-trait standard units

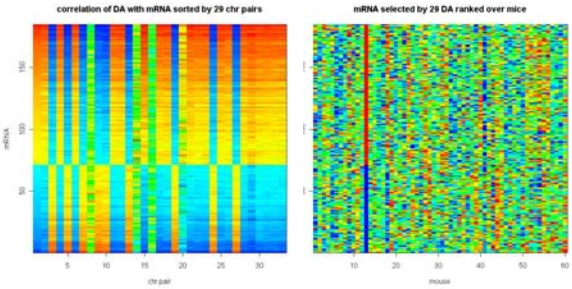
SCD trait log2 expression





# 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 $\mu$
  $\text{pr}(\underline{Y} \mid Q, G) = f_1(\underline{Y} \mid Q, G) = \prod q\ f_0(\underline{Yq} \mid G)$
  $f_0(\underline{Y}_q \mid G) = \int f(\underline{Y}_q \mid \underline{\mu}, G)\ \text{pr}(\underline{\mu})\ d\underline{\mu}$
- 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)

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