Bayesian analysis of microarray traits

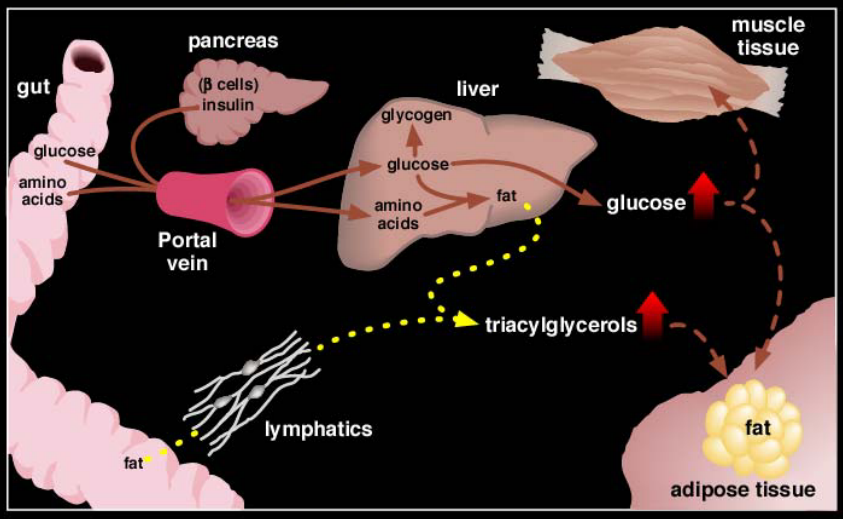
Arabidopsis Microarray Workshop

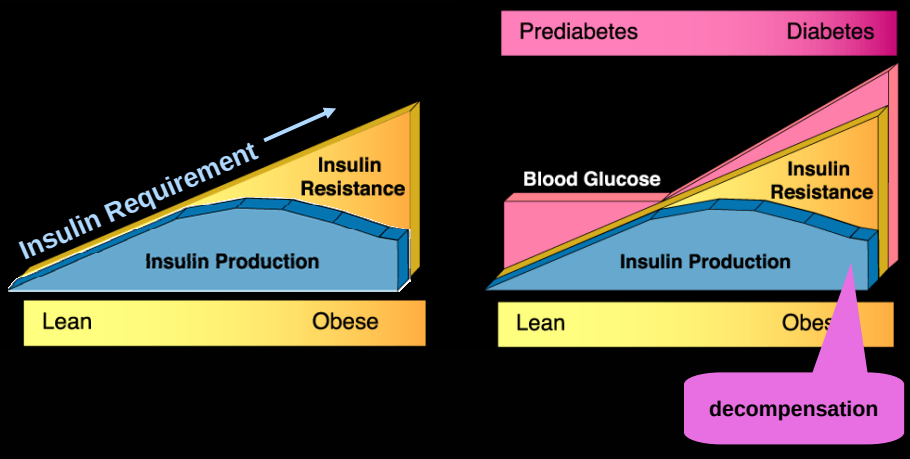
Brian S. Yandell
University of Wisconsin-Madison
www.stat.wisc.edu/~yandell/statgen

studying diabetes in an F2

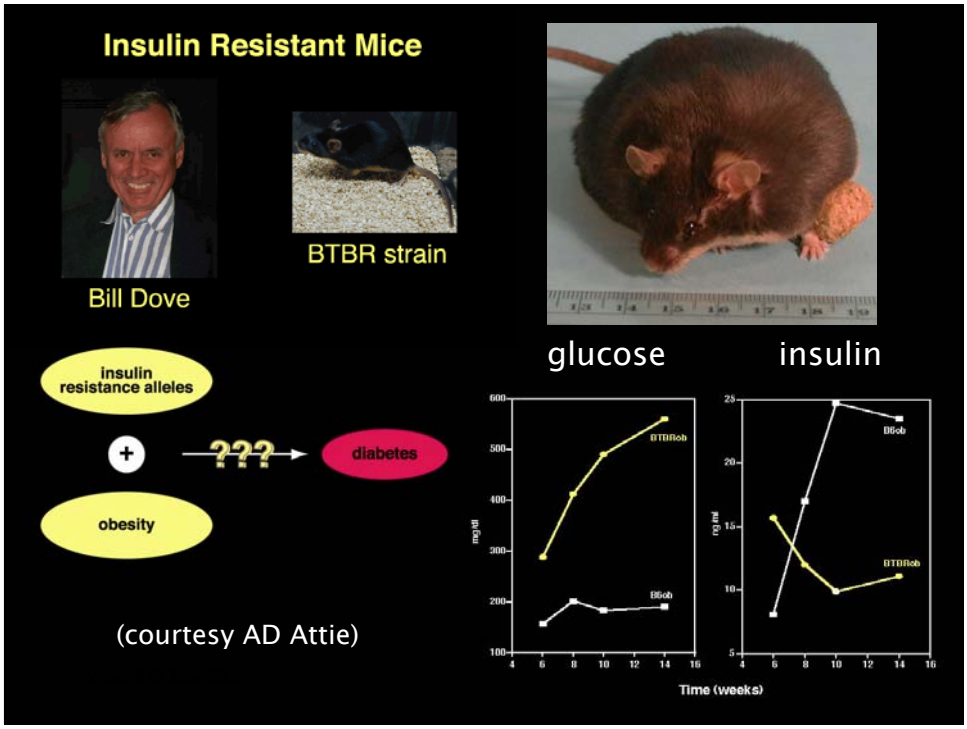
- segregating cross of inbred lines
  - B6.ob x BTBR.ob → F1 → F2
  - selected mice with ob/ob alleles at leptin gene (chr 6)
  - measured and mapped body weight, insulin, glucose at various ages (Stoehr et al. 2000 *Diabetes*)
  - sacrificed at 14 weeks, tissues preserved
- gene expression data
  -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

Type 2 Diabetes Mellitus





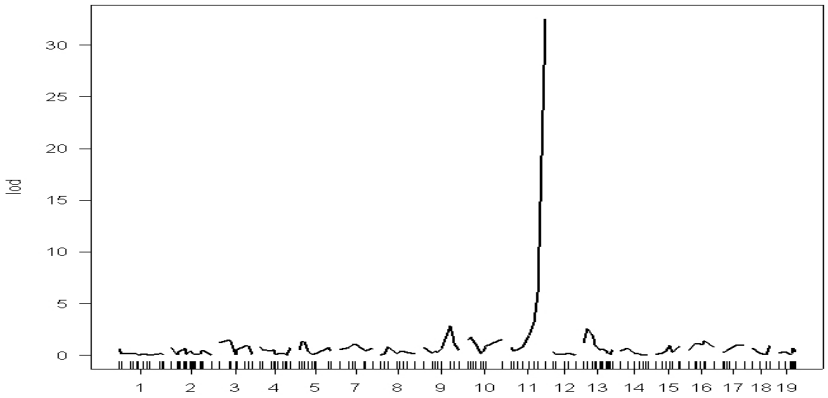
from Unger & Orci *FASEB J.* (2001) 15,312



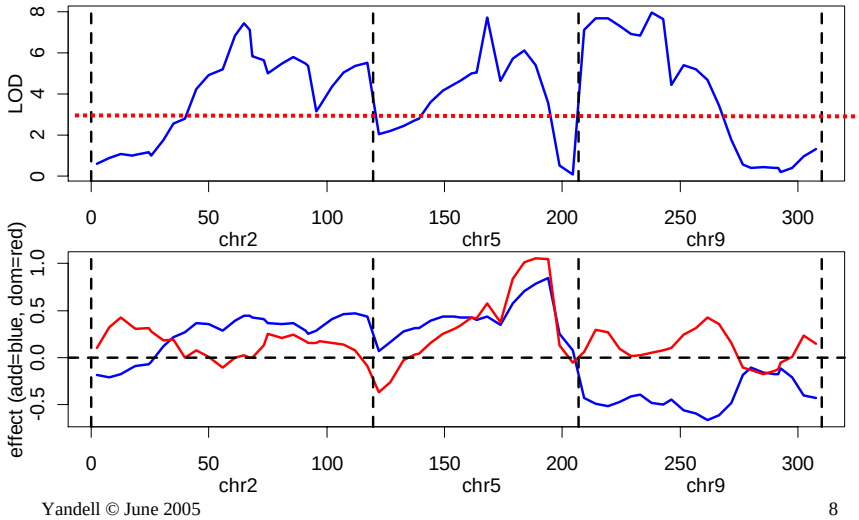
why map gene expression as a quantitative trait?

- *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

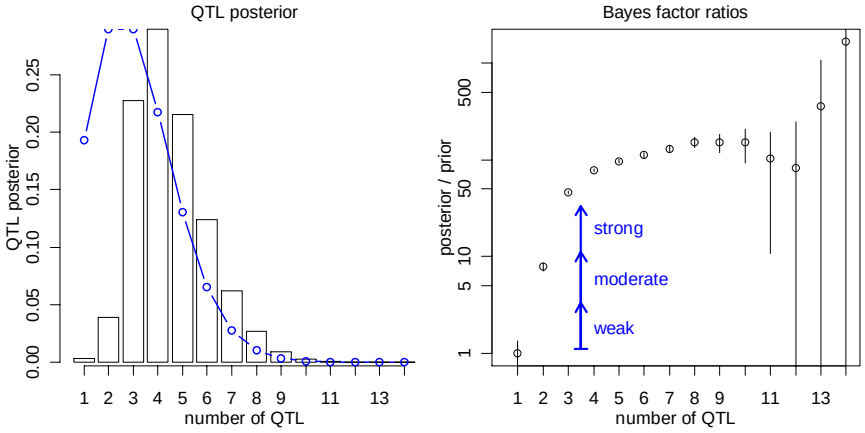
LOD map for PDI: *cis*-regulation (Lan et al. 2003)



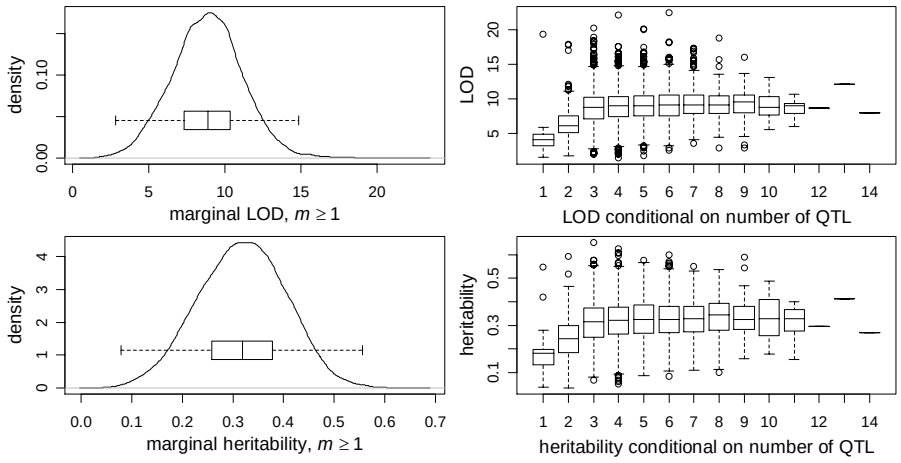
Multiple Interval Mapping (QTLCart) SCD1: multiple QTL plus epistasis!



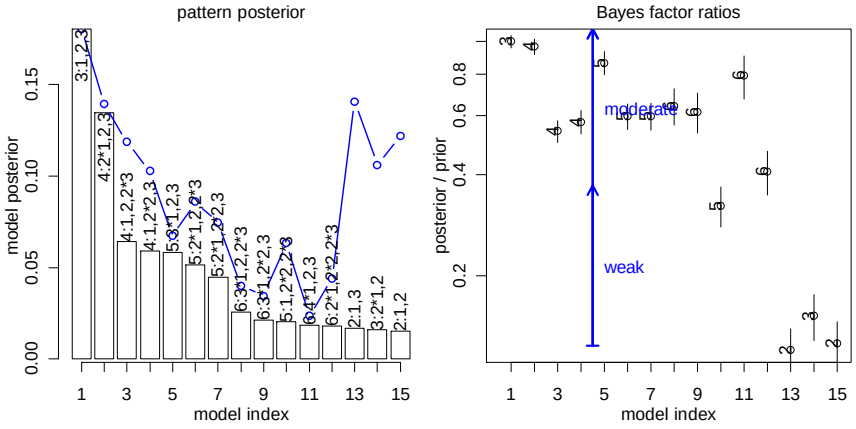
Bayesian model assessment: number of QTL for SCD1



Bayesian LOD and h^2 for SCD1

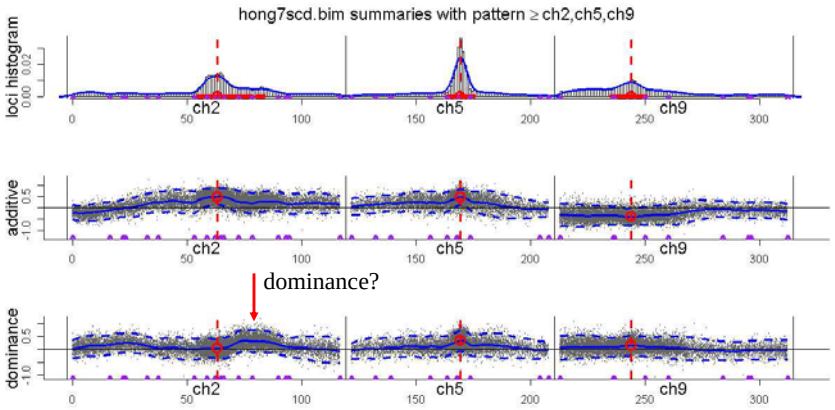


Bayesian model assessment: chromosome QTL pattern for SCD1

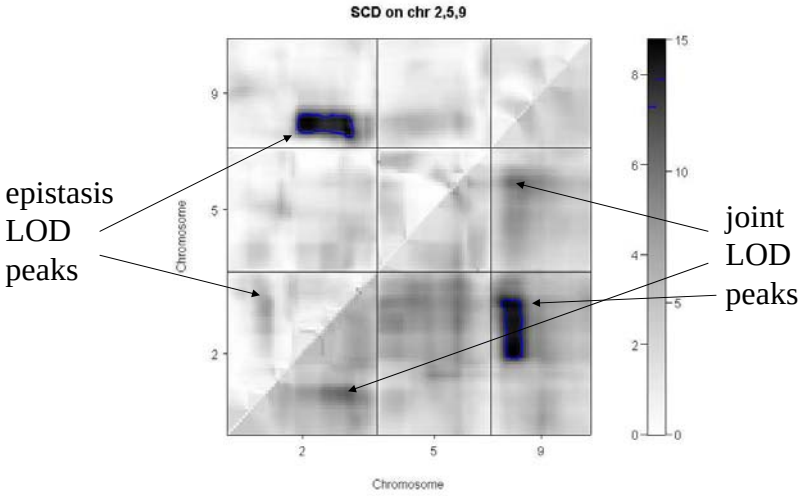


trans-acting QTL for SCD1

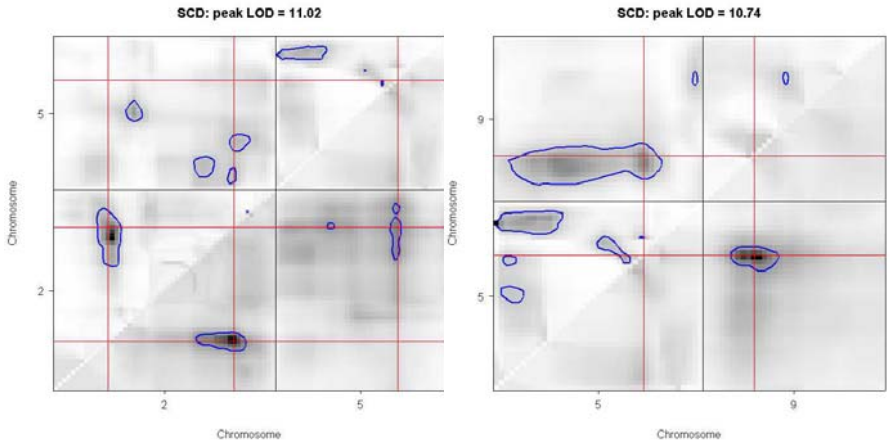
(no epistasis yet: see Yi, Xu, Allison 2003)



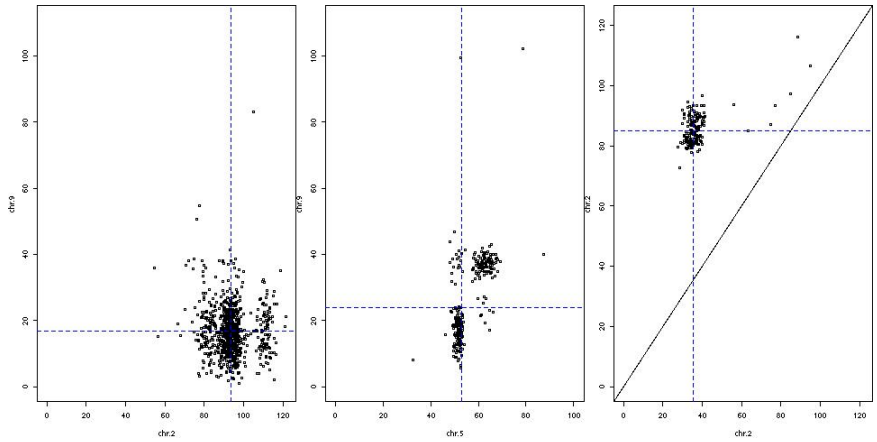
2-D scan: assumes only 2 QTL!



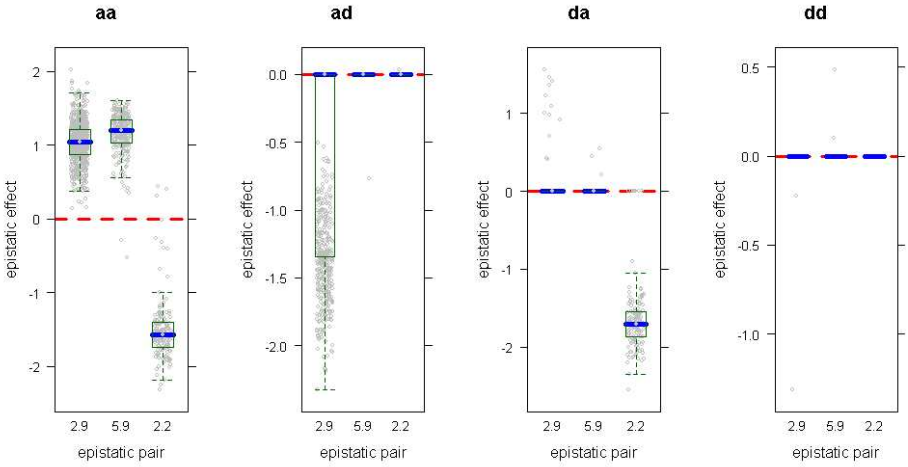
sub-peaks can be easily overlooked!



epistatic model fit



Cockerham epistatic effects



our Bayesian QTL 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)

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 μ

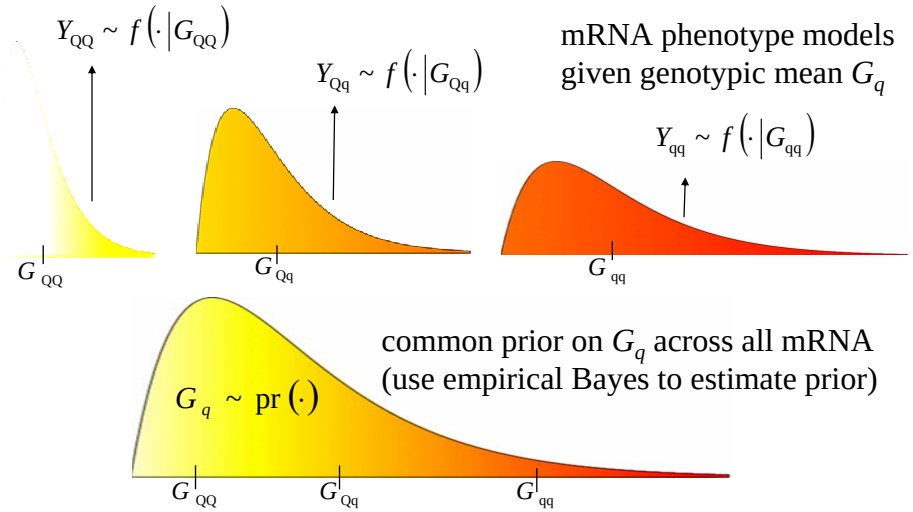
  $\text{pr}(Y|Q, \text{not } H) = f_0(Y) = \int f(Y|G)\ \text{pr}(G)\ dG$ if not H

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



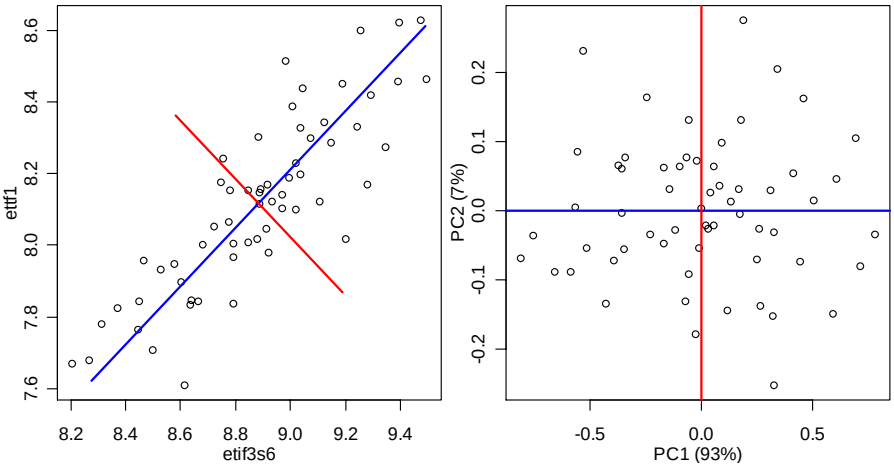
why study multiple traits together?

- 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

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

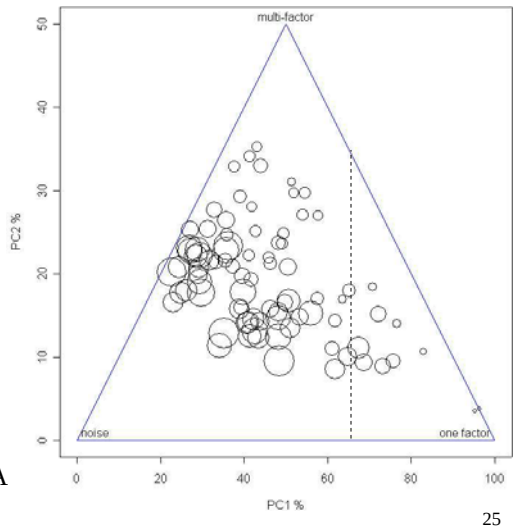
Affy chips on 60 mice
~40,000 mRNA

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

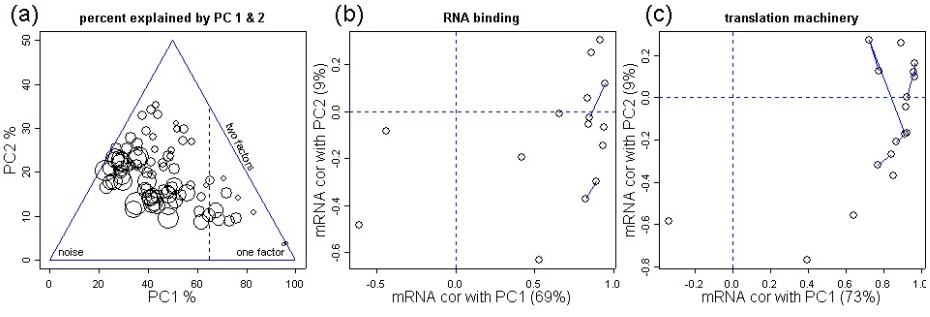
1500+ organized in
85 functional groups
2-35 mRNA / group

which are interesting?
examine PC1, PC2

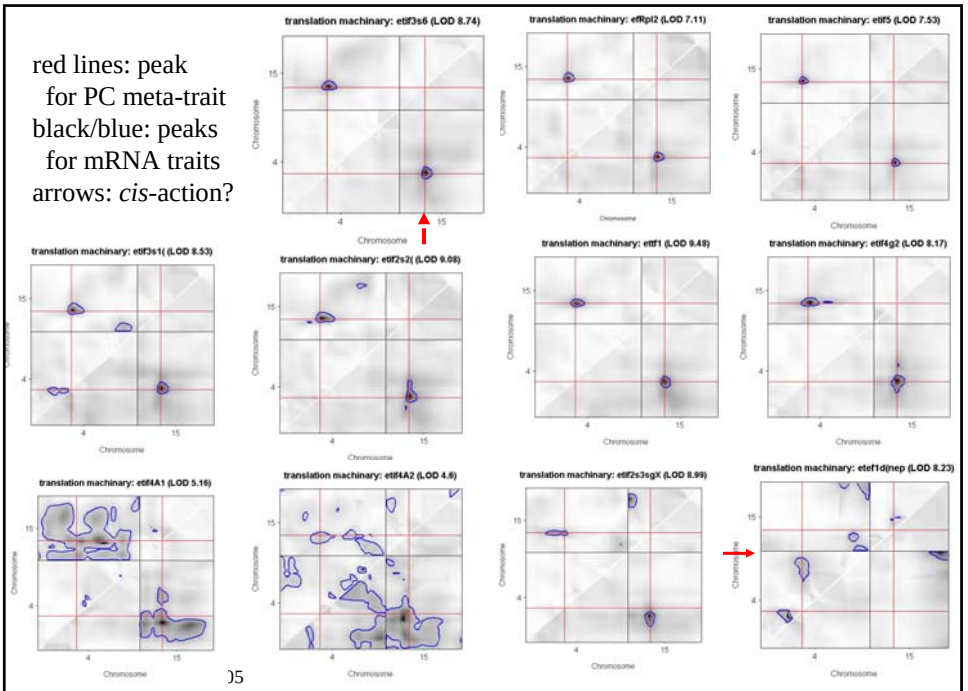
circle size = # unique mRNA



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

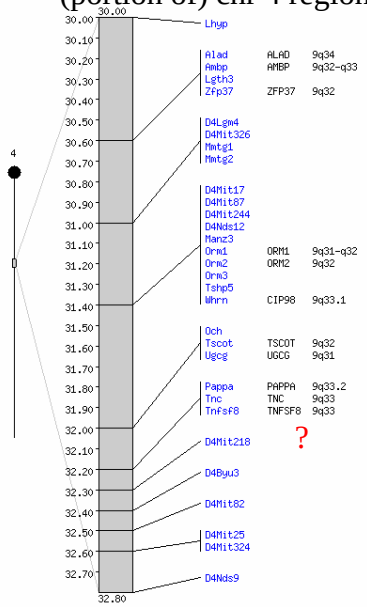


red lines: peak
for PC meta-trait
black/blue: peaks
for mRNA traits
arrows: *cis*-action?

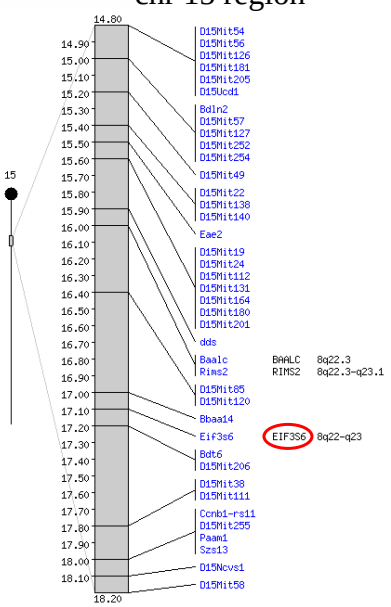


)5

(portion of) chr 4 region



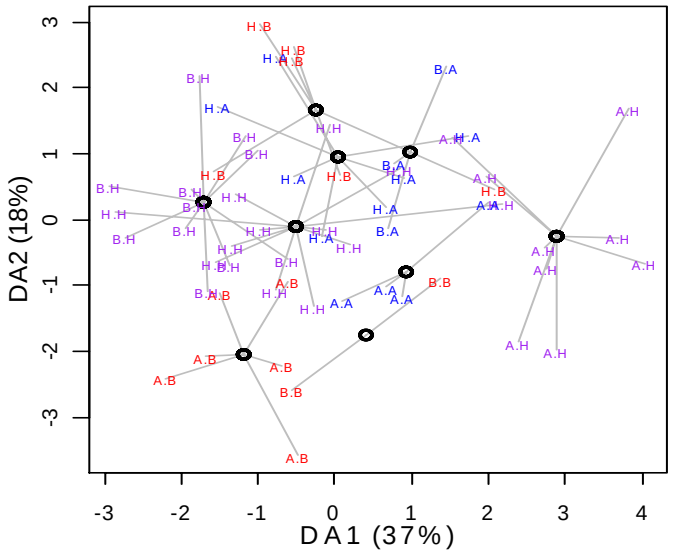
chr 15 region



DA meta-traits on 1500+ mRNA traits

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

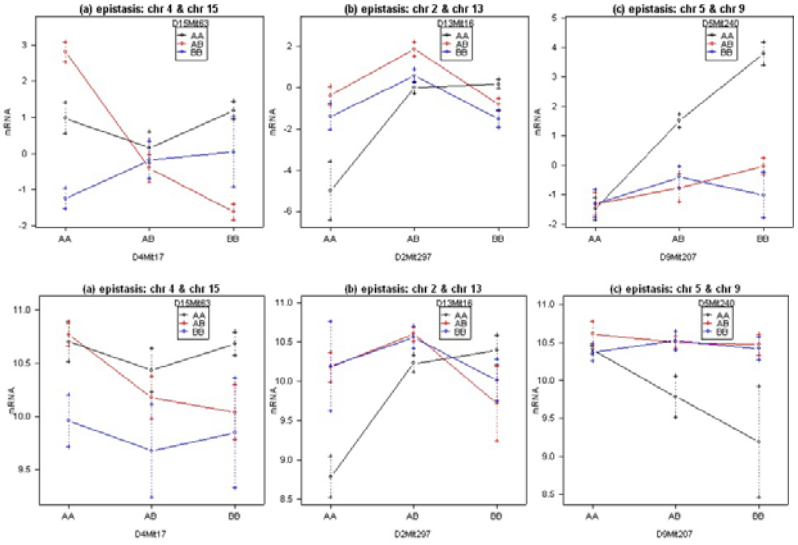
DA creates best separation by genotype



relating meta-traits to mRNA traits

DA meta-trait standard units

SCD trait log2 expression



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

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

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

•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 $\rightarrow$ RNA1 $\rightarrow$ RNA2

  – class of possible models

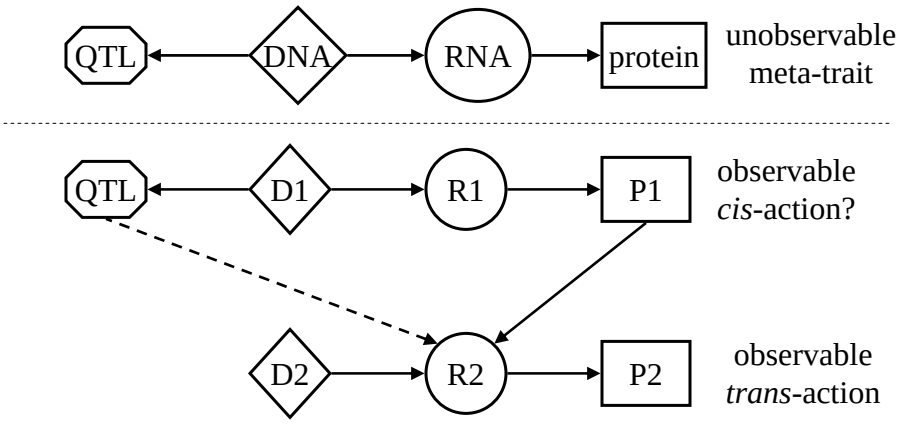
  – best model = putative biochemical pathway
- parallel biochemical investigation

  – 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