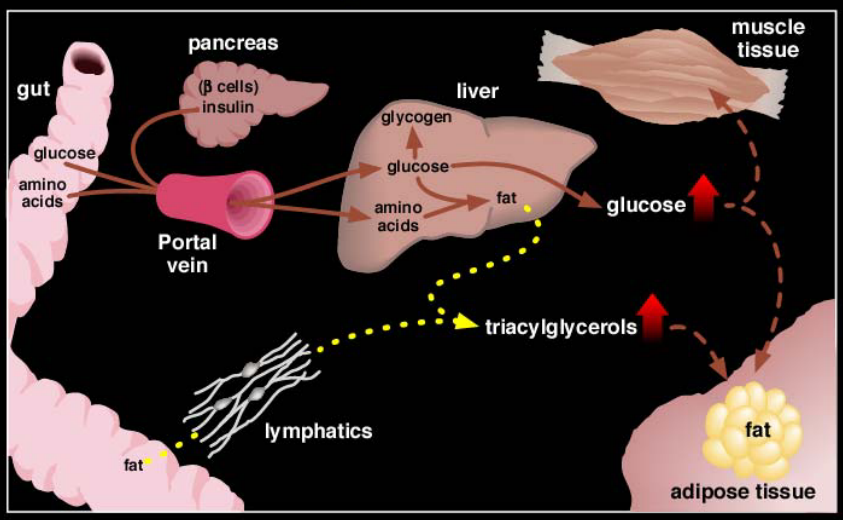
Multiple Traits & Microarrays

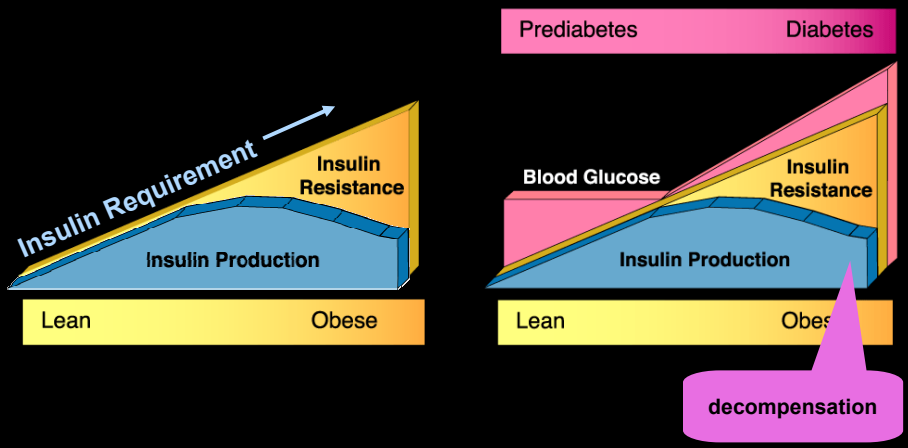
1. why study multiple traits together? 2-10
   - diabetes case study
2. design issues 11-13
   - selective phenotyping
3. why are traits correlated? 14-17
   - close linkage or pleiotropy?
4. modern high throughput 18-31
   - principal components & discriminant analysis
5. graphical models 32-36
   - building causal biochemical networks

1. why study multiple traits together?

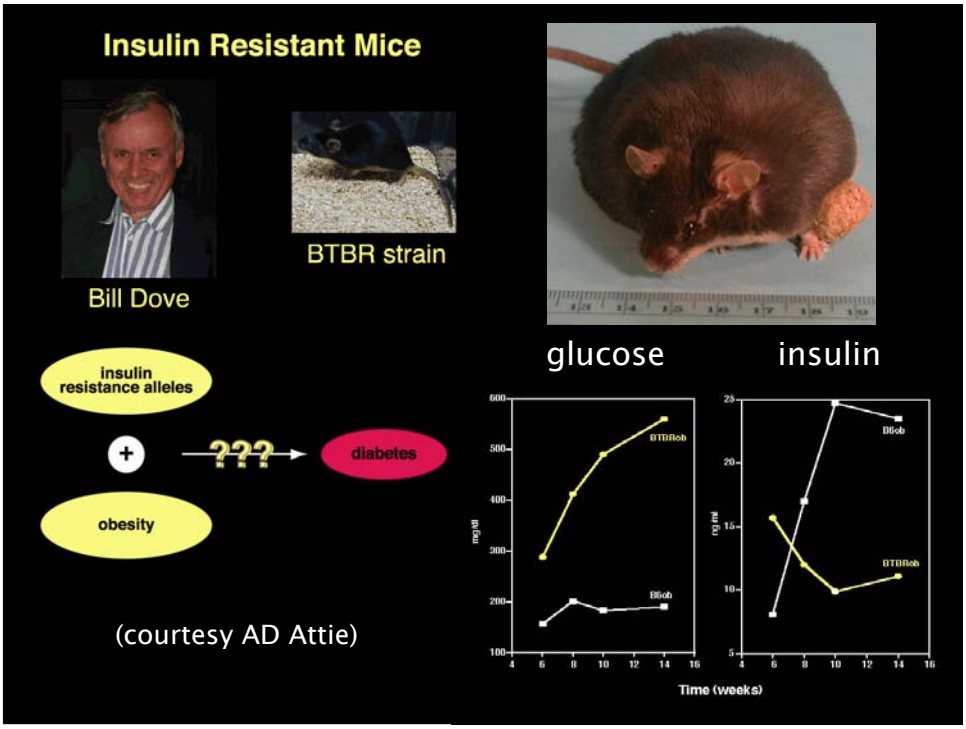
- avoid reductionist approach to biology
  - address physiological/biochemical mechanisms
  - Schmalhausen (1942); Falconer (1952)
- separate close linkage from pleiotropy
  - 1 locus or 2 linked loci?
- identify epistatic interaction or canalization
  - influence of genetic background
- establish QTL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



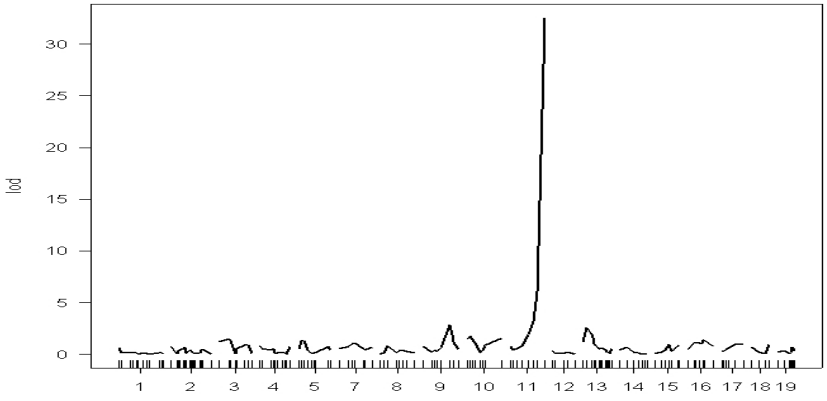
studying diabetes in an F2

- 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

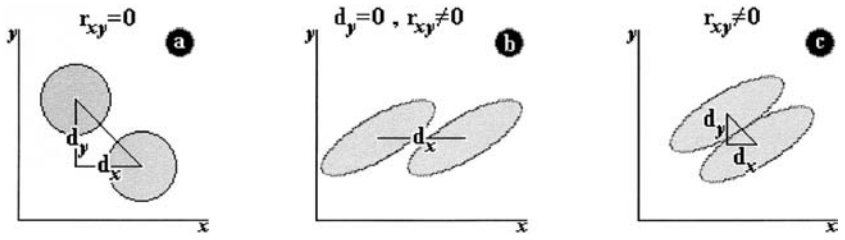
is this relevant to large QTL studies?

- 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



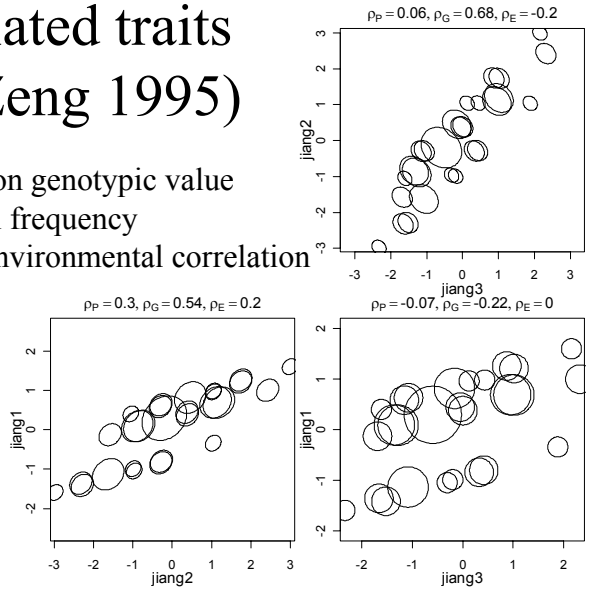
pleiotropy only correlation only both

Korol et al. (2001)

3 correlated traits (Jiang Zeng 1995)

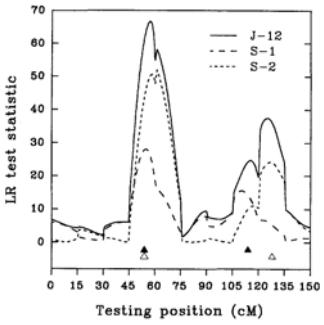
ellipses centered on genotypic value
width for nominal frequency
main axis angle environmental correlation
3 QTL, F2
27 genotypes

note signs of genetic and environmental correlation



pleiotropy or close linkage?

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



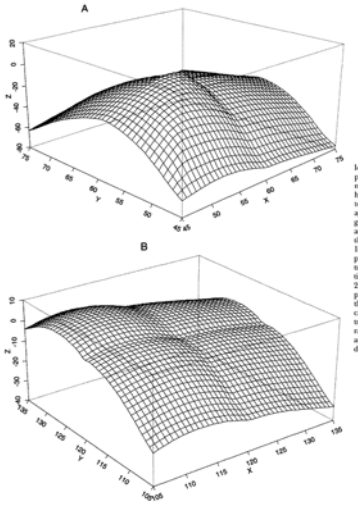


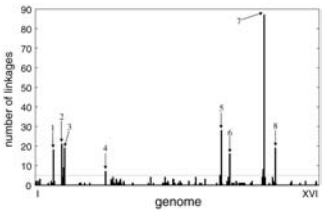
FIGURE 2.—Two-dimensional log-likelihood surfaces (expressed as deviations from the maximum of the log-likelihoods on the diagonal) for the test of pleiotropy *vs.* close linkage are presented for two regions: the region between 45 and 75 cM of Figure 1 (A) and the region between 105 and 135 cM (B). *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.

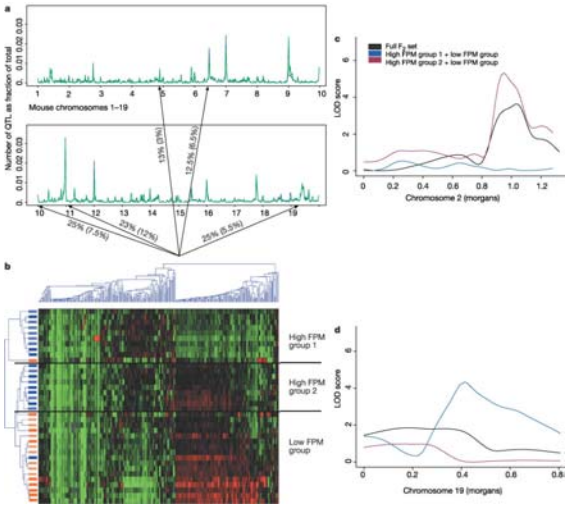
4. modern high throughput biology

- 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

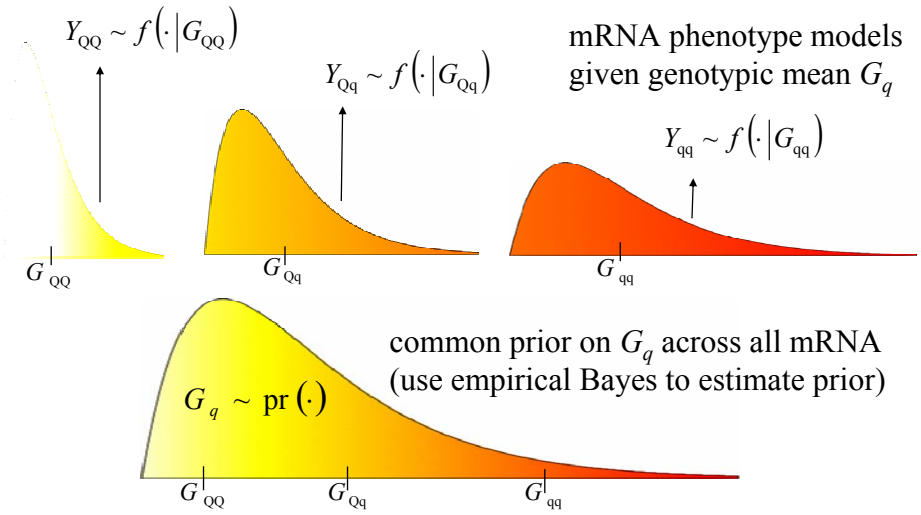
  $\text{pr}(Y|Q, \text{not } H) = f_0(Y) = \int f(Y|G)\ \text{pr}(G)\ dG$ if not H

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

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

hierarchical model for expression phenotypes

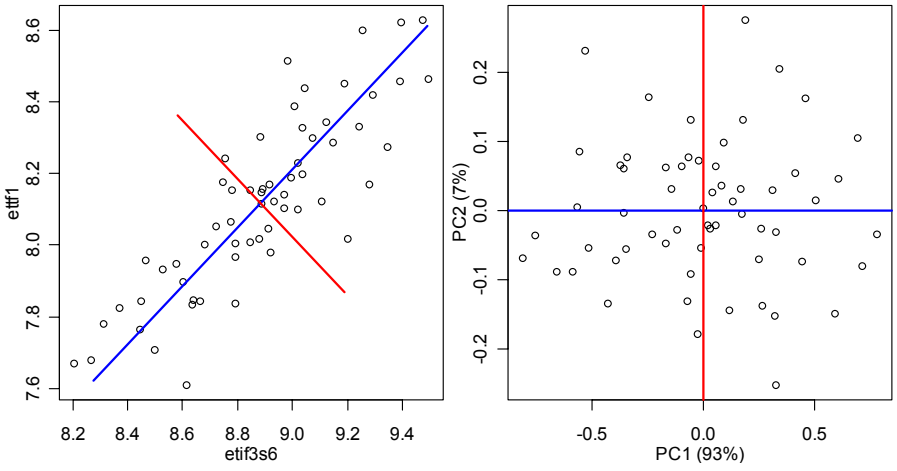
(EB arrays: Christina Kendziorski)



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

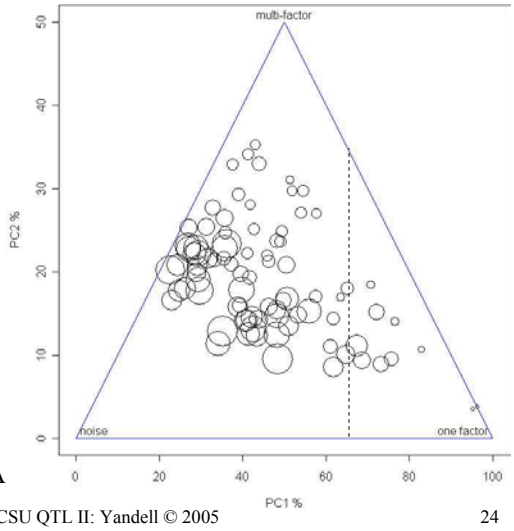
Affy chips on 60 mice
~40,000 mRNA

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

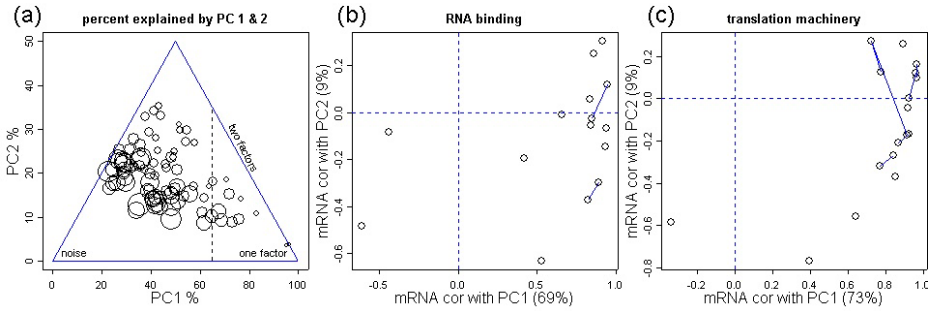
1500+ organized in
85 functional groups
2-35 mRNA / group

which are interesting?
examine PC1, PC2

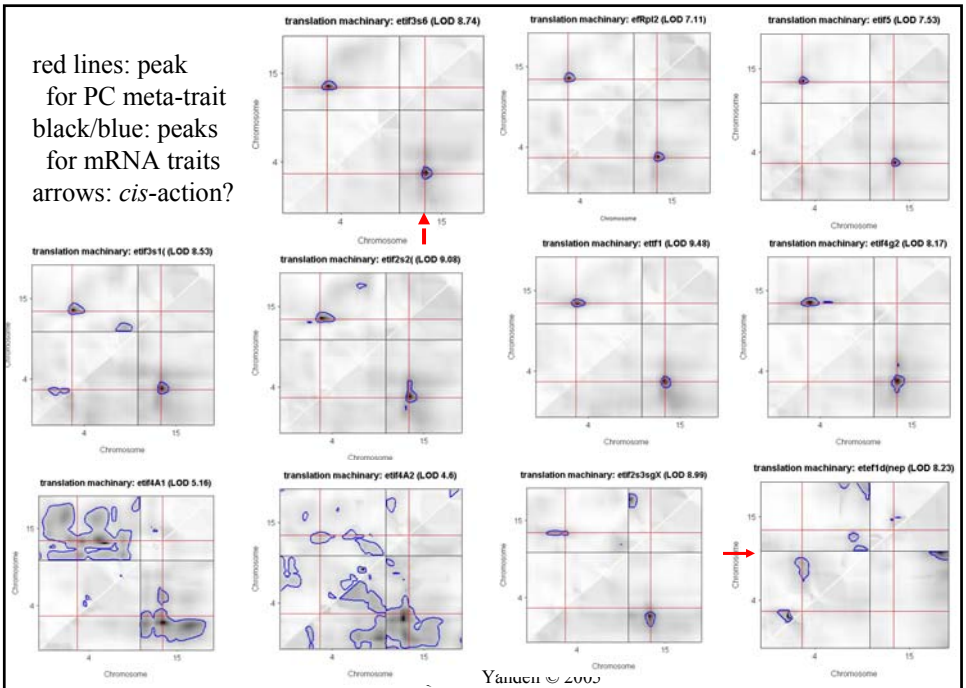
circle size = # unique mRNA



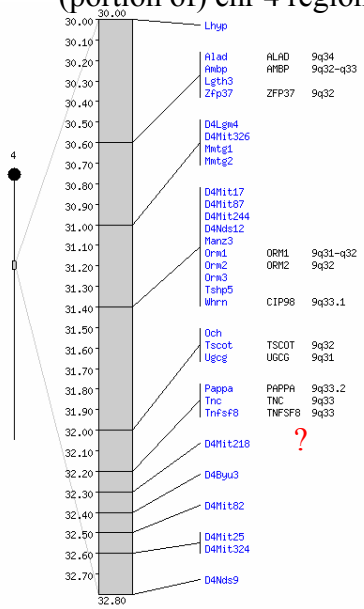
84 PC meta-traits by functional group focus on 2 interesting groups



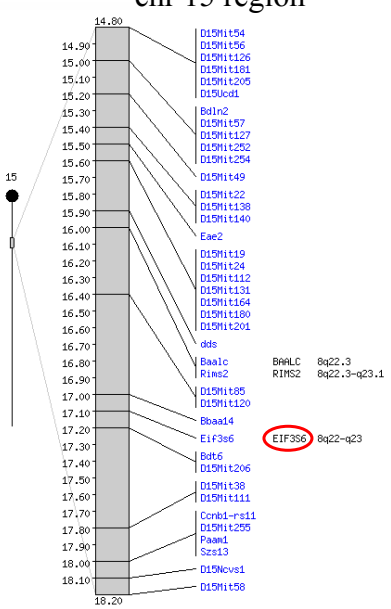
red lines: peak for PC meta-trait
black/blue: peaks for mRNA traits
arrows: *cis*-action?



(portion of) chr 4 region



chr 15 region



interaction plots for DA meta-traits

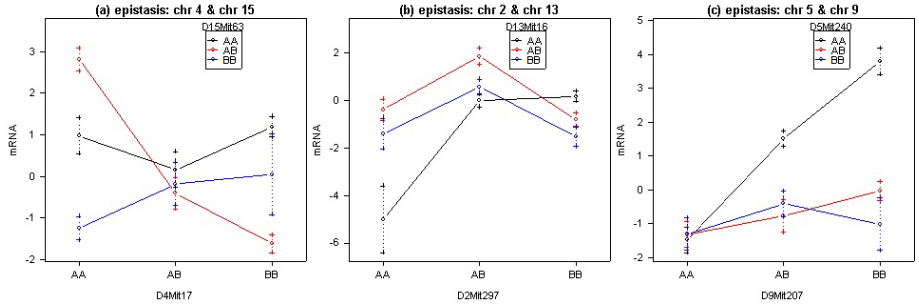
DA for all pairs of markers:

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

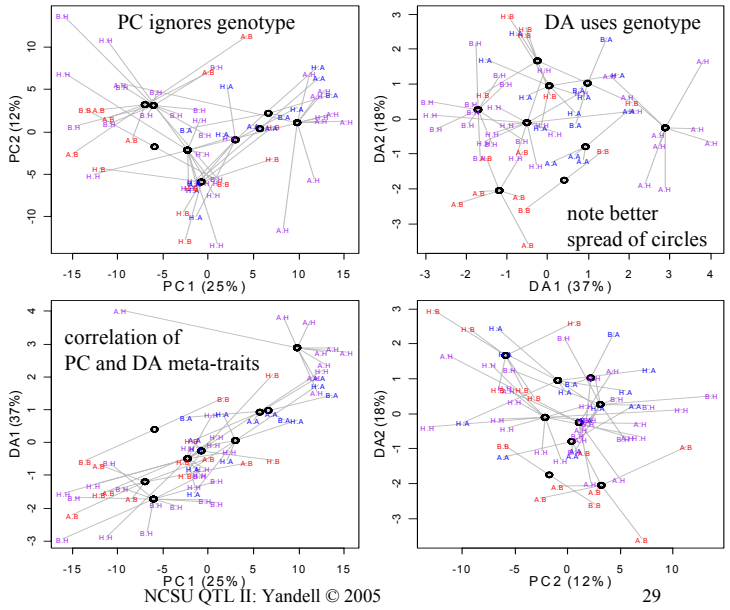


comparison of PC and DA meta-traits on 1500+ mRNA traits

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

PC captures spread without genotype

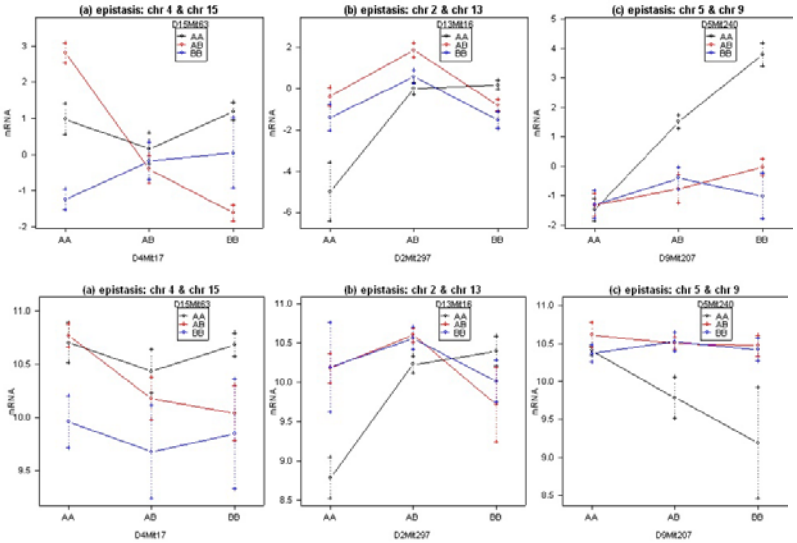
DA creates best separation by genotype



relating meta-traits to mRNA traits

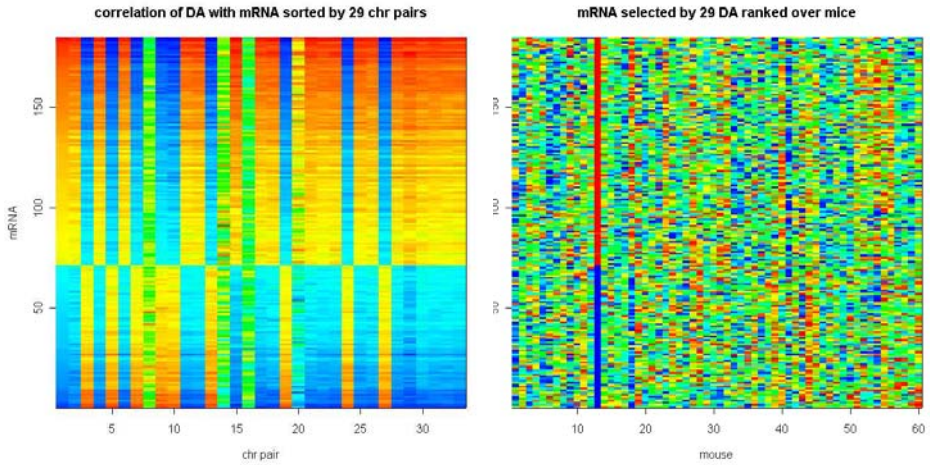
DA meta-trait standard units

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

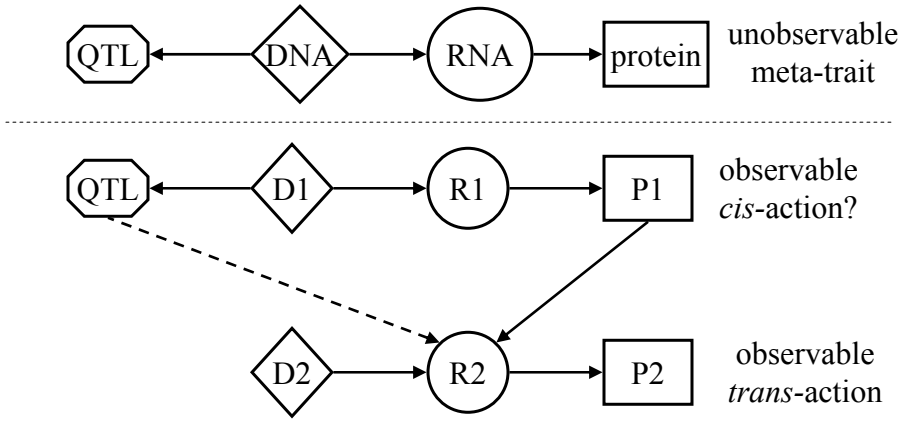
  – best model = putative biochemical pathway
- parallel biochemical investigation

  – 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