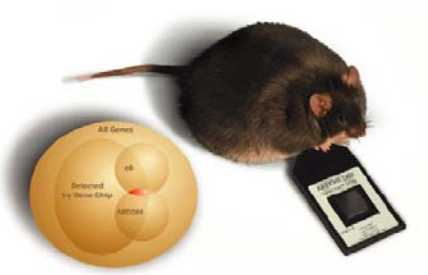


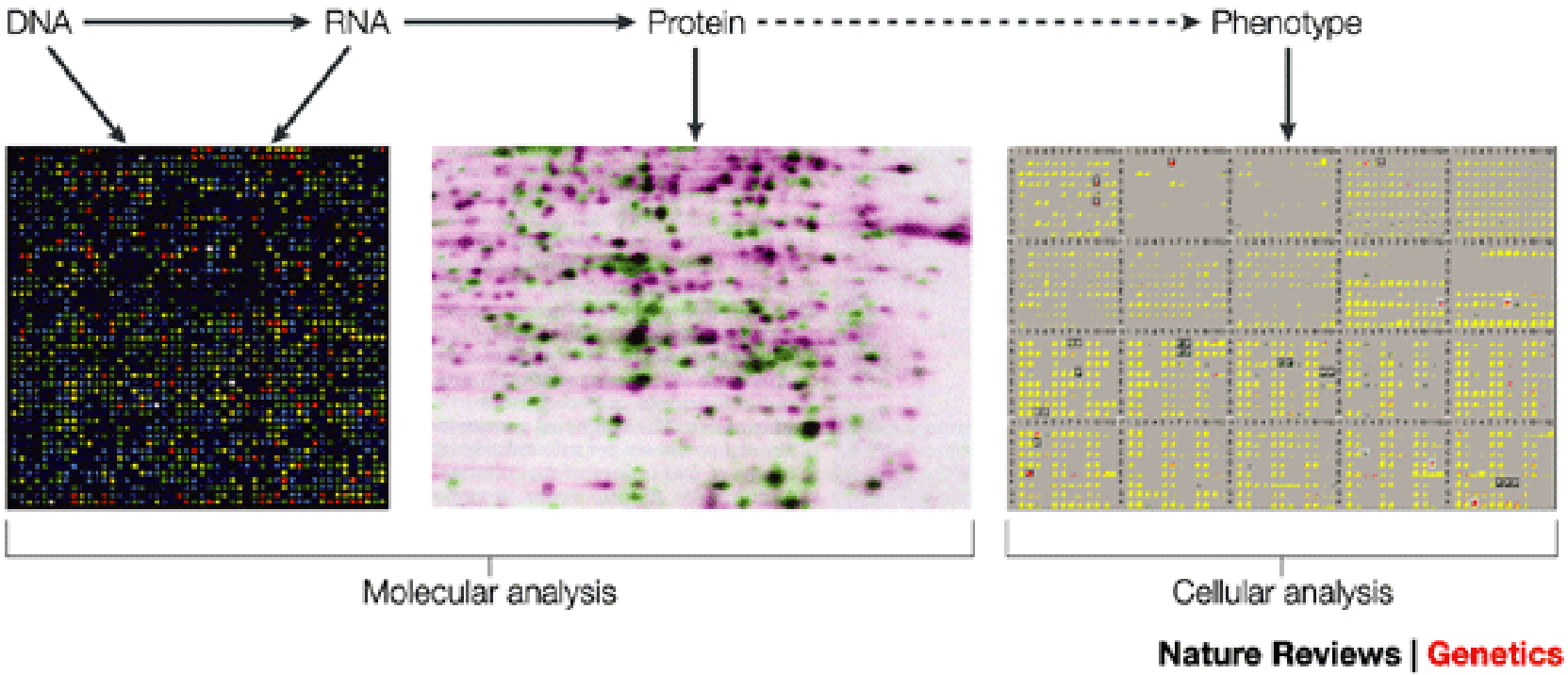
High Throughput Gene Mapping
Brian S. Yandell

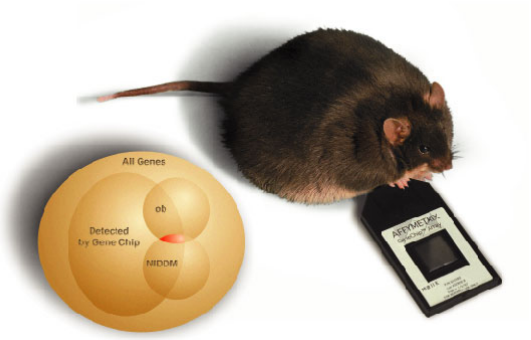
Summer Research Program in Biostatistics

June 2004

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

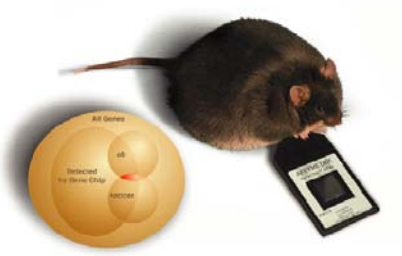
central dogma via microarrays (Bochner 2003)





what can you do?

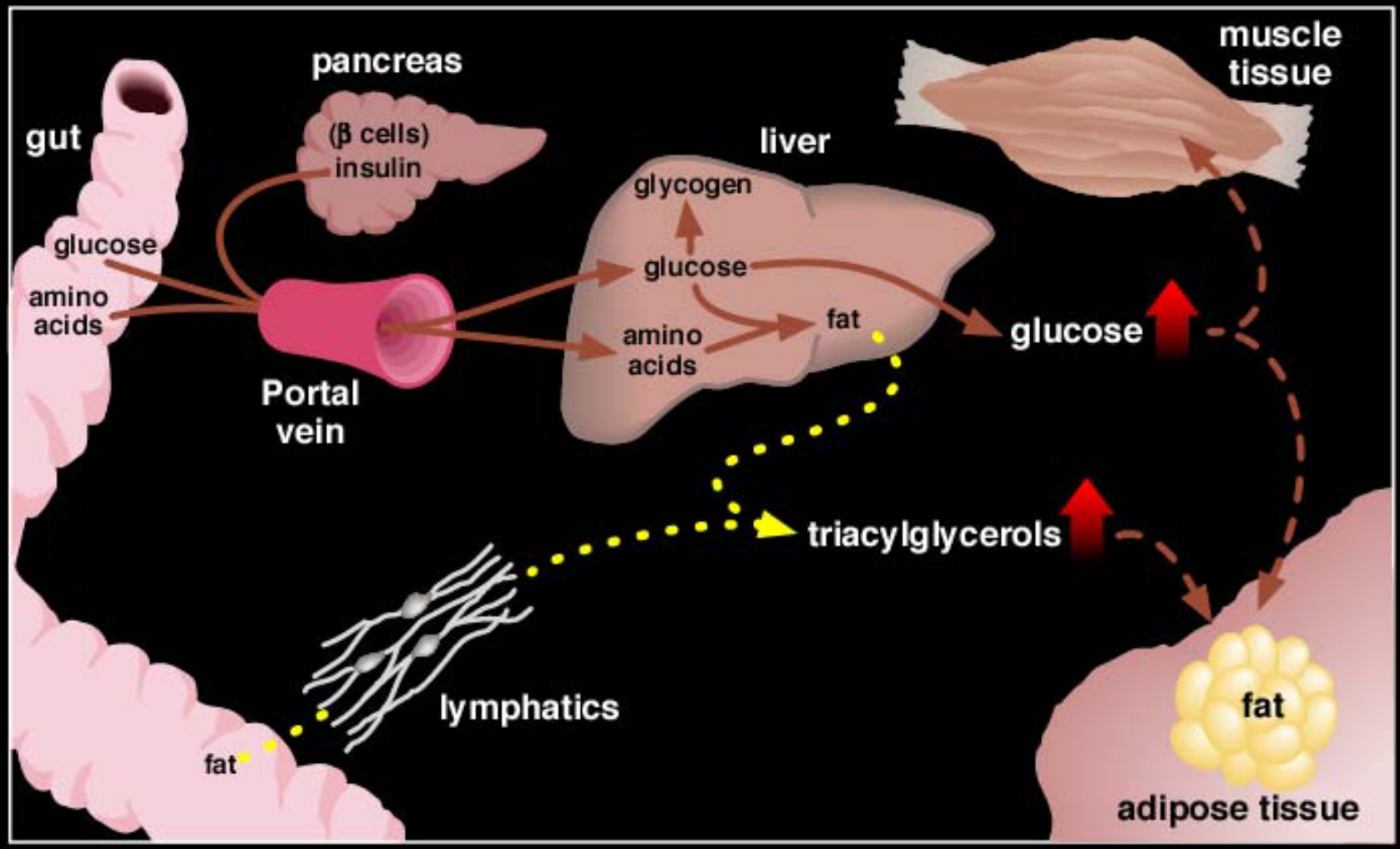
- participate in hierarchy of research teams
  - biostatistics: Yandell, Kendziorski, 3-5 grad students
  - bioinformatics: Attie, Lan (biochem), Craven + 2 CS grad students
  - biochemistry: (optional) weekly interdisciplinary lab meetings
- conduct data analysis of 1-2 large data sets
  - 30,000 responses, 60 individuals, 200 genetic markers
  - learn multivariate statistical & quantitative statistical methods
  - develop innovative graphical summaries
- develop statistical computing tools
  - learn about construction of R libraries and archiving
  - develop new code with potential wide usage
  - transfer research methods to practice through user-friendly code

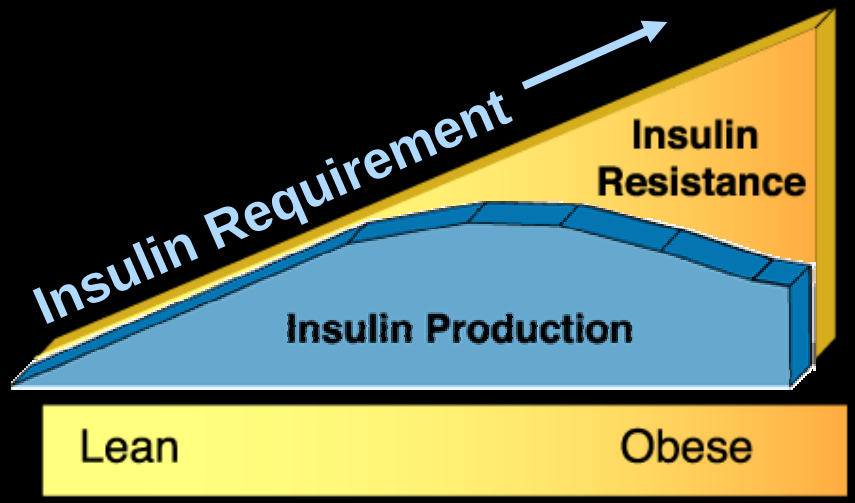


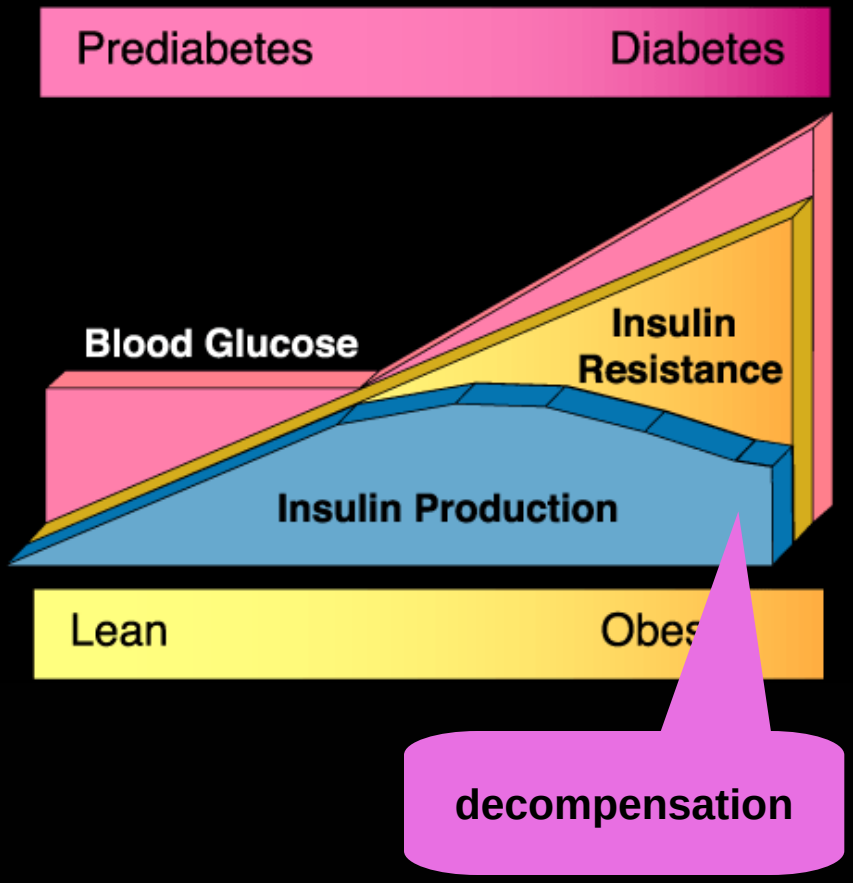
studying diabetes in an F2

- 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 progress)

Type 2 Diabetes Mellitus



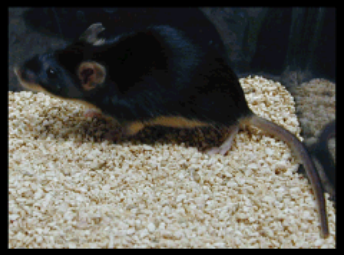




from Unger & Orci *FASEB J.* (2001) 15,312

Insulin Resistant Mice

Bill Dove



BTBR strain

insulin resistance alleles

+

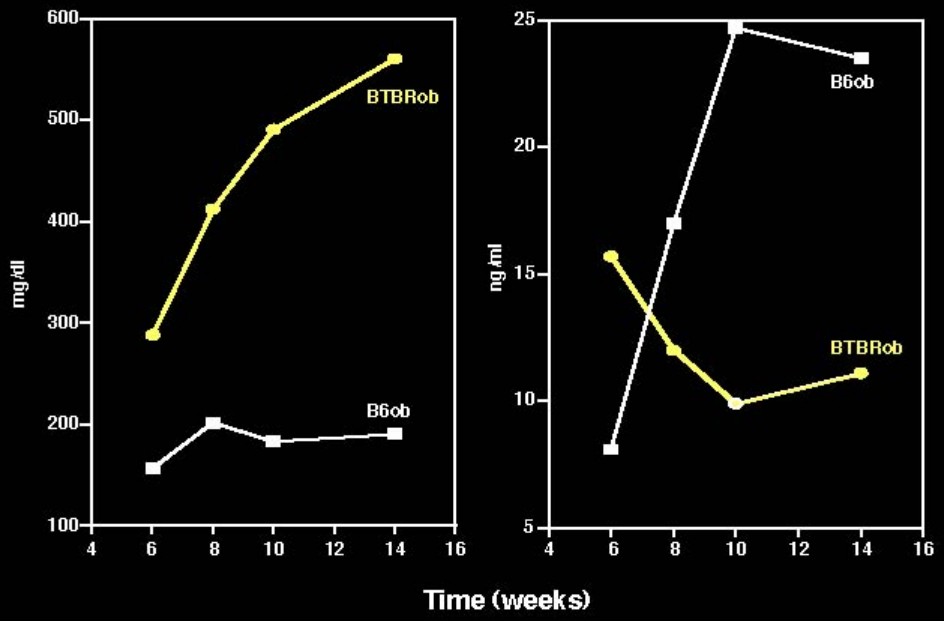
obesity

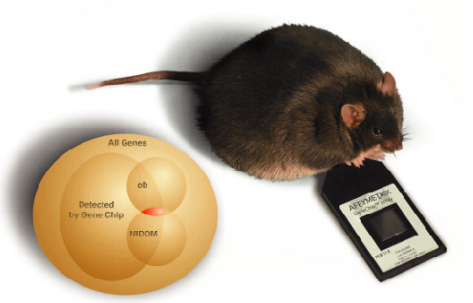
??? → diabetes

(courtesy AD Attie)

glucose

insulin

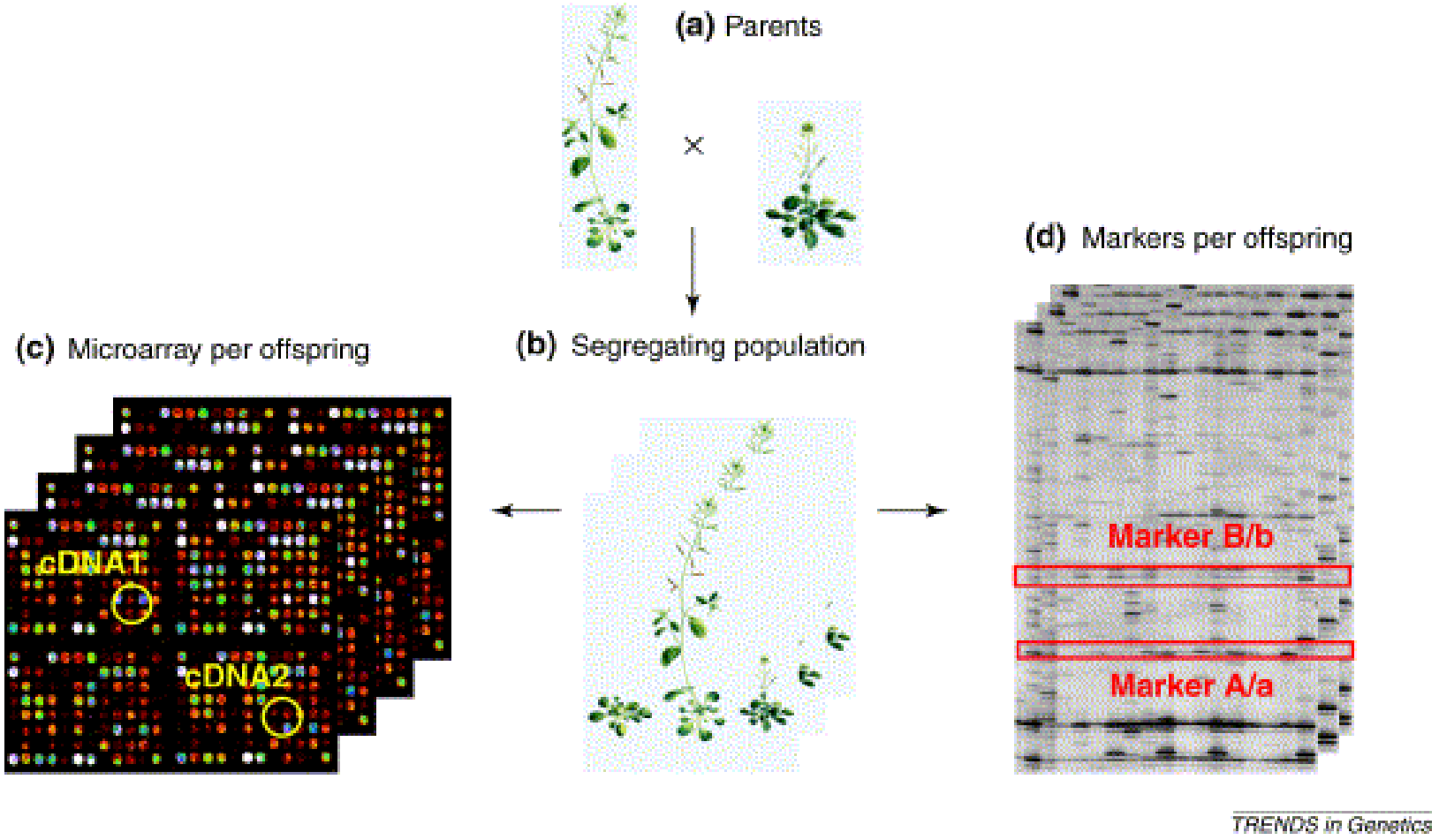




why map gene expression as a quantitative trait?

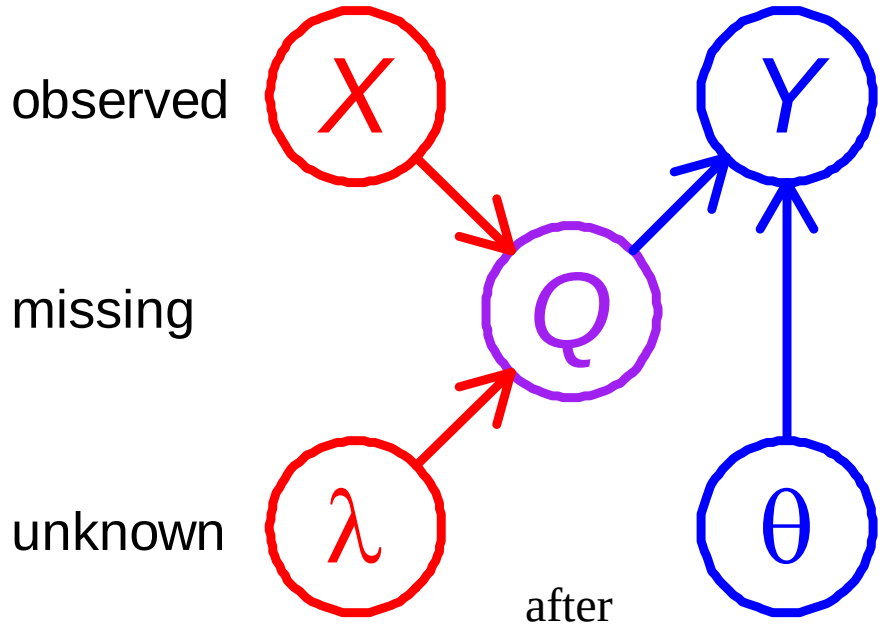
- *cis*- or *trans*-action?
  - does gene control its own expression?
  - or is it influenced by one or more other genomic regions?
  - evidence for both modes (Brem et al. 2002 Science)
- simultaneously measure all mRNA in a tissue
  - ~5,000 mRNA active per cell on average
  - ~30,000 genes in genome
  - 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

idea of mapping microarrays (Jansen Nap 2001)



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
  - λ = QT locus (or loci)
  - θ = phenotype model parameters
  - m = number of QTL
- pr($Q|X,\lambda,m$) recombination model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for Q given X
- pr($Y|Q,\theta,m$) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters θ (could be non-parametric)

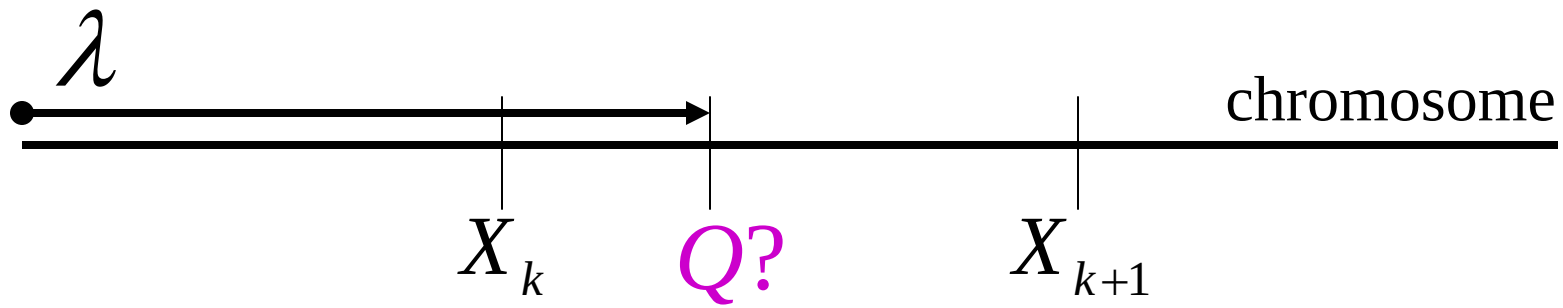


after
Sen Churchill (2001)

recombination model pr($Q|X,\lambda$)

- locus λ is distance along linkage map
  - identifies flanking marker region
- flanking markers provide good approximation
  - map assumed known from earlier study
  - inaccuracy slight using only flanking markers
    - extend to next flanking markers if missing data
  - could consider more complicated relationship
    - but little change in results

$$\text{pr}(Q|X,\lambda) = \text{pr}(\text{geno} \mid \text{map, locus}) \approx \text{pr}(\text{geno} \mid \text{flanking markers, locus})$$



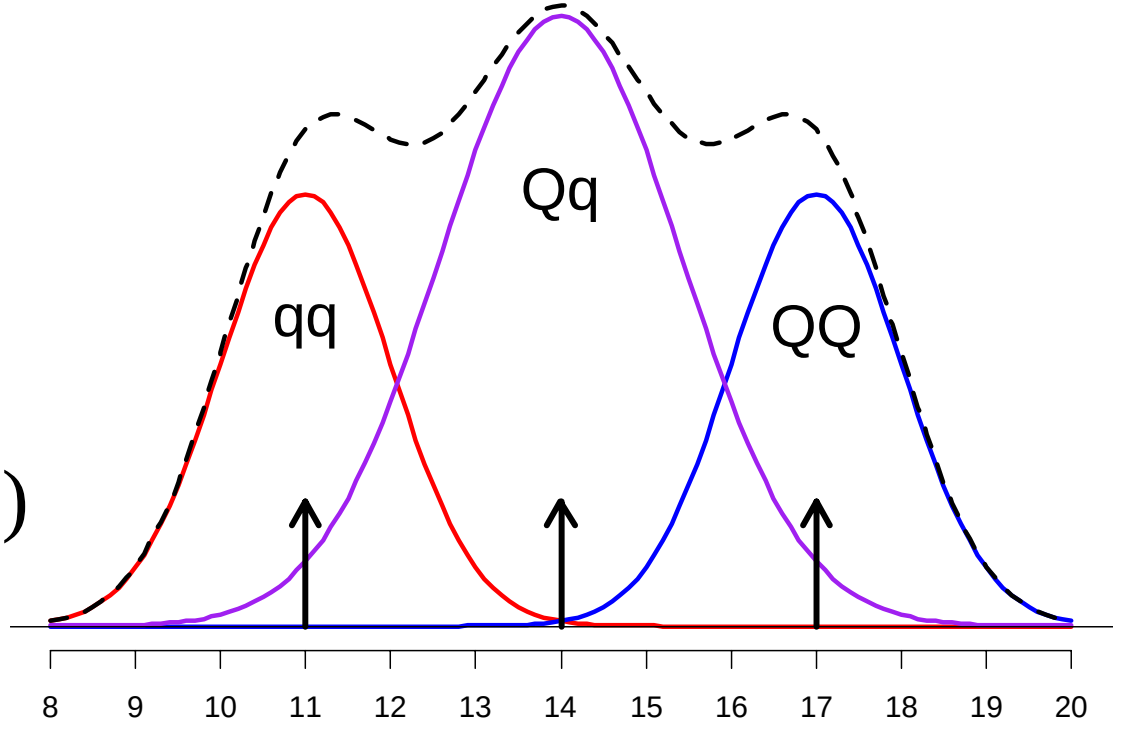
idealized phenotype model

- trait = mean + additive + error
- trait = effect_of_geno + error
- pr(trait | geno, effects)

$$Y = G_Q + E$$

$$\text{pr}(Y \mid Q, \theta) =$$

$$\text{normal}(G_Q, \sigma^2)$$



interval mapping objective

- likelihood mixes over genotypes Q

  $L(\lambda,\theta|Y) = \text{product}_i\ [\text{sum}_Q\ \text{pr}(Q|X_i,\lambda)\ \text{pr}(Y_i|Q,\theta)]$

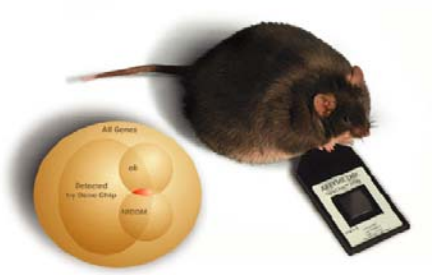
  – maximize likelihood to estimate loci & effects

  – LOD = log10($L(\lambda,\theta|Y)$ / null likelihood)

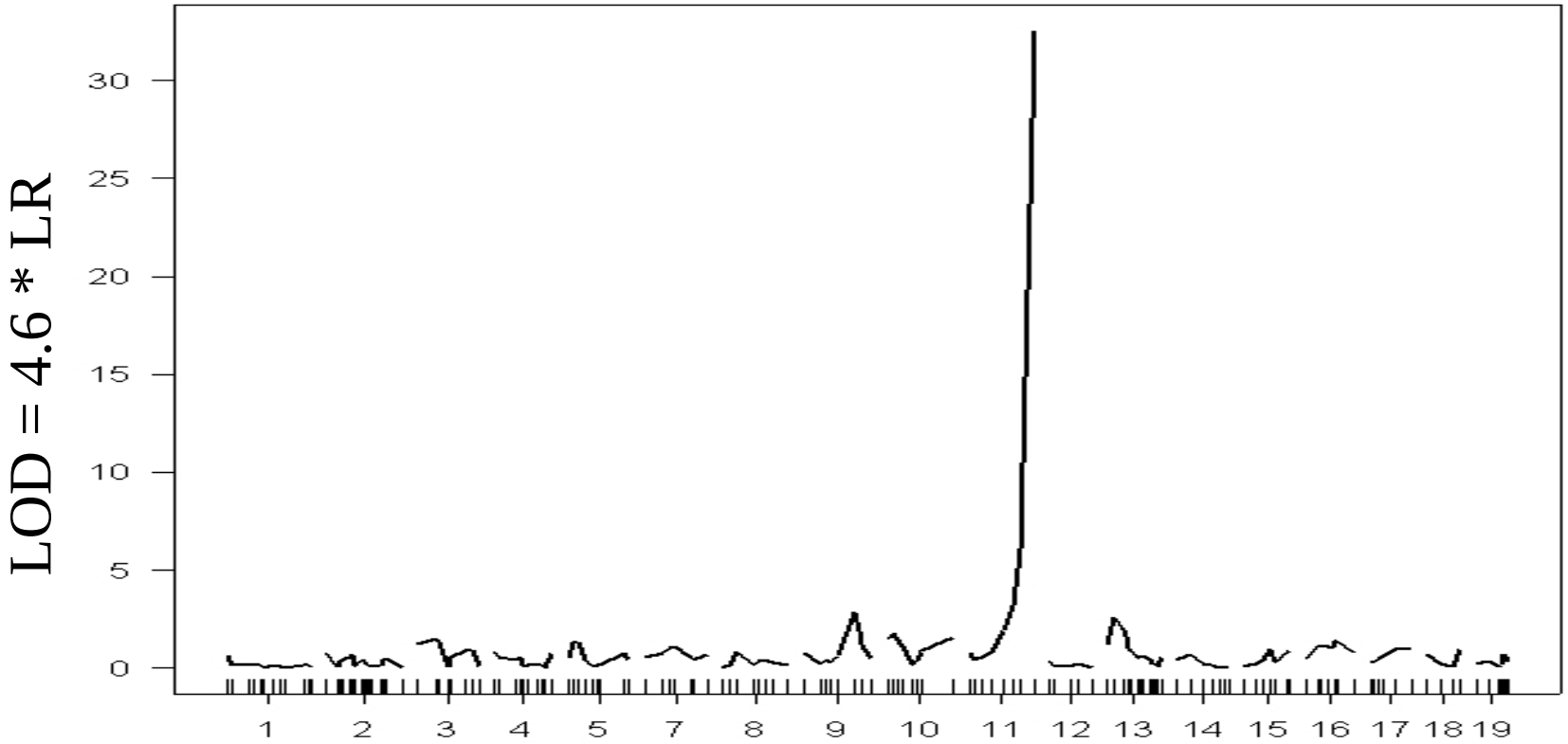
- Bayesian posterior samples Q as missing data

  $\text{pr}(\lambda,Q,\theta|Y,X) = \text{pr}(\lambda,\theta)\ \text{product}_i\ \text{pr}(Q_i|X_i,\lambda)\ \text{pr}(Y_i|Q_i,\theta)$

  – average over unknown Q to study loci & effects

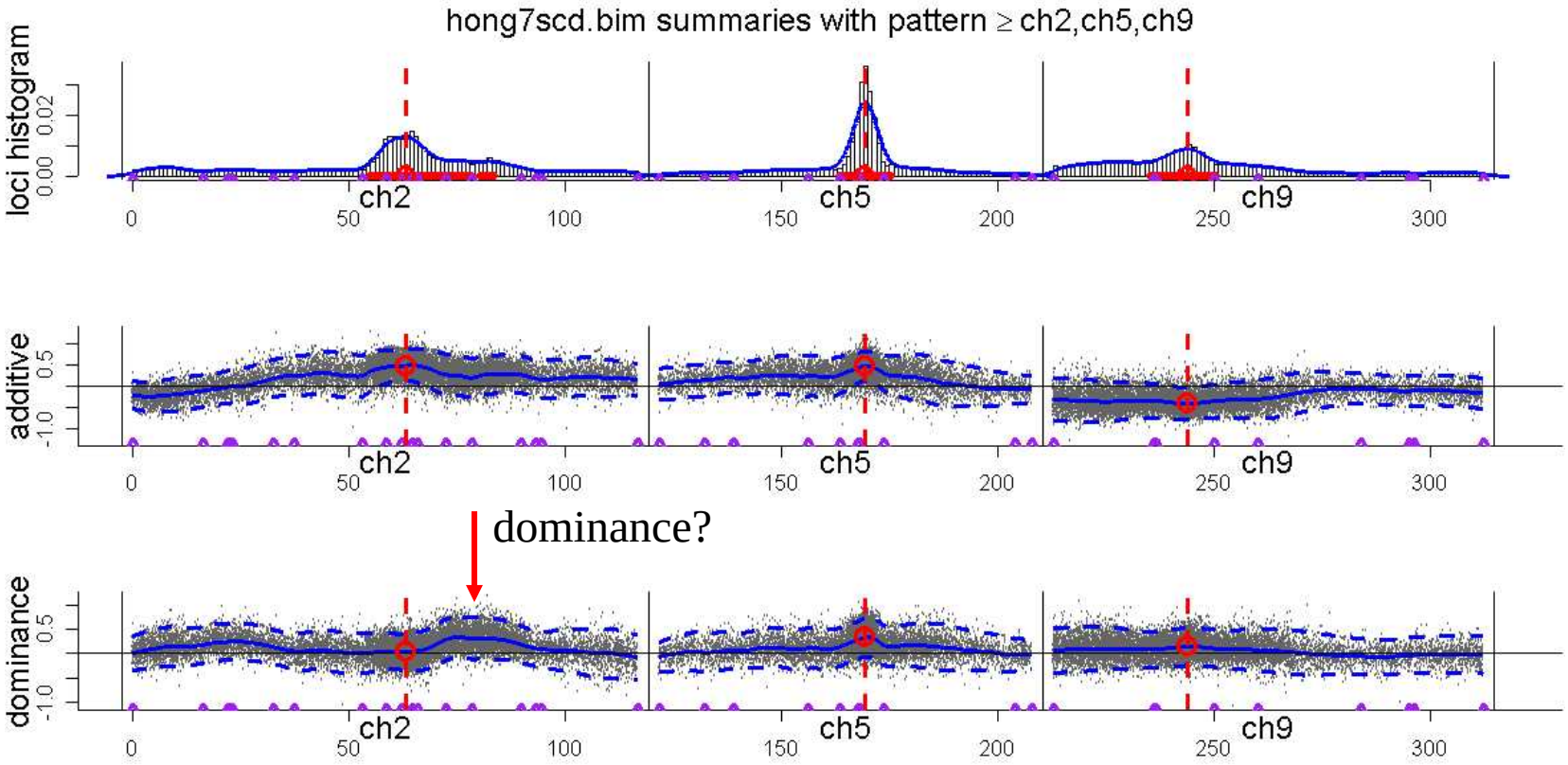


simple LOD map for PDI: *cis*-regulation (Lan et al. 2003)

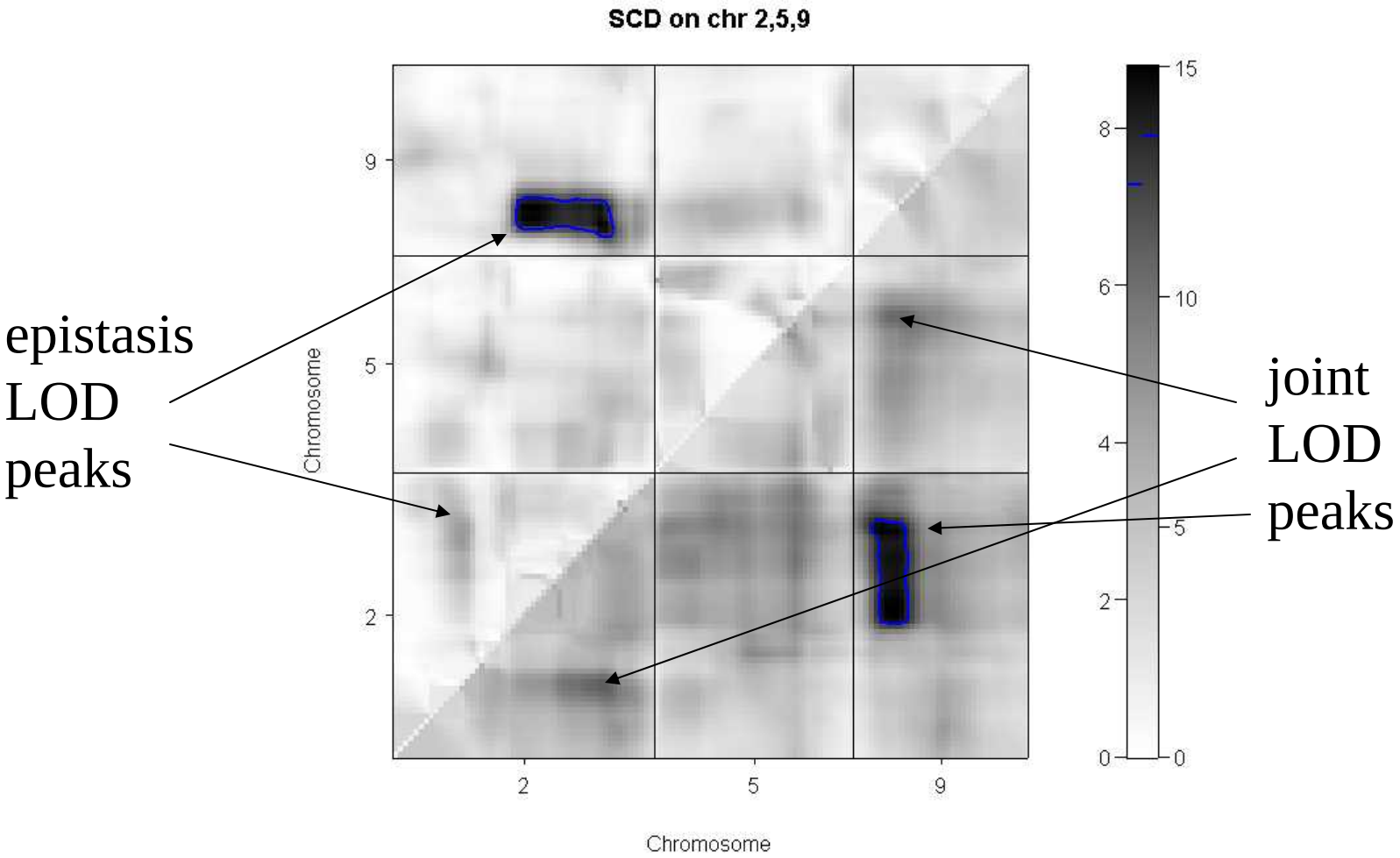


complicated *trans*-action for SCD1

(3-4 gene regions influence expression of SCD1)



statistical interaction for SCD1



multiple QTL phenotype model

- phenotype affected by genotype & environment

  $\text{pr}(Y|Q,\theta) \sim N(G_Q, \sigma^2)$

  $Y = G_Q$ + environment

- partition genotypic mean into QTL effects

  $G_Q = \mu + \beta_1(Q) + \ldots + \beta_m(Q) + \beta_{12}(Q) + \ldots$

  G_Q = mean + main effects + epistatic interactions

- general form of QTL effects for model M

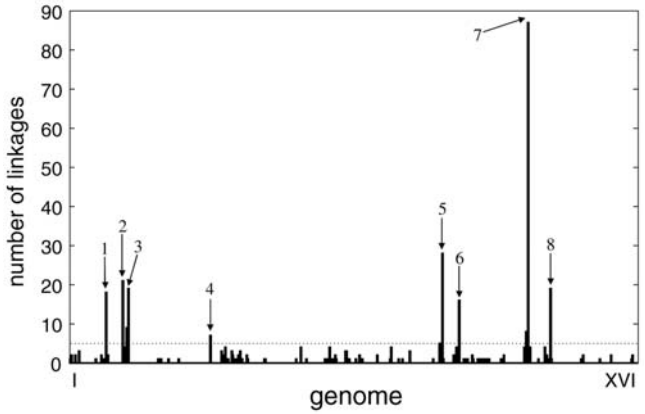
  $G_Q = \mu + \text{sum}_{j \text{ in } M} \beta_j(Q)$

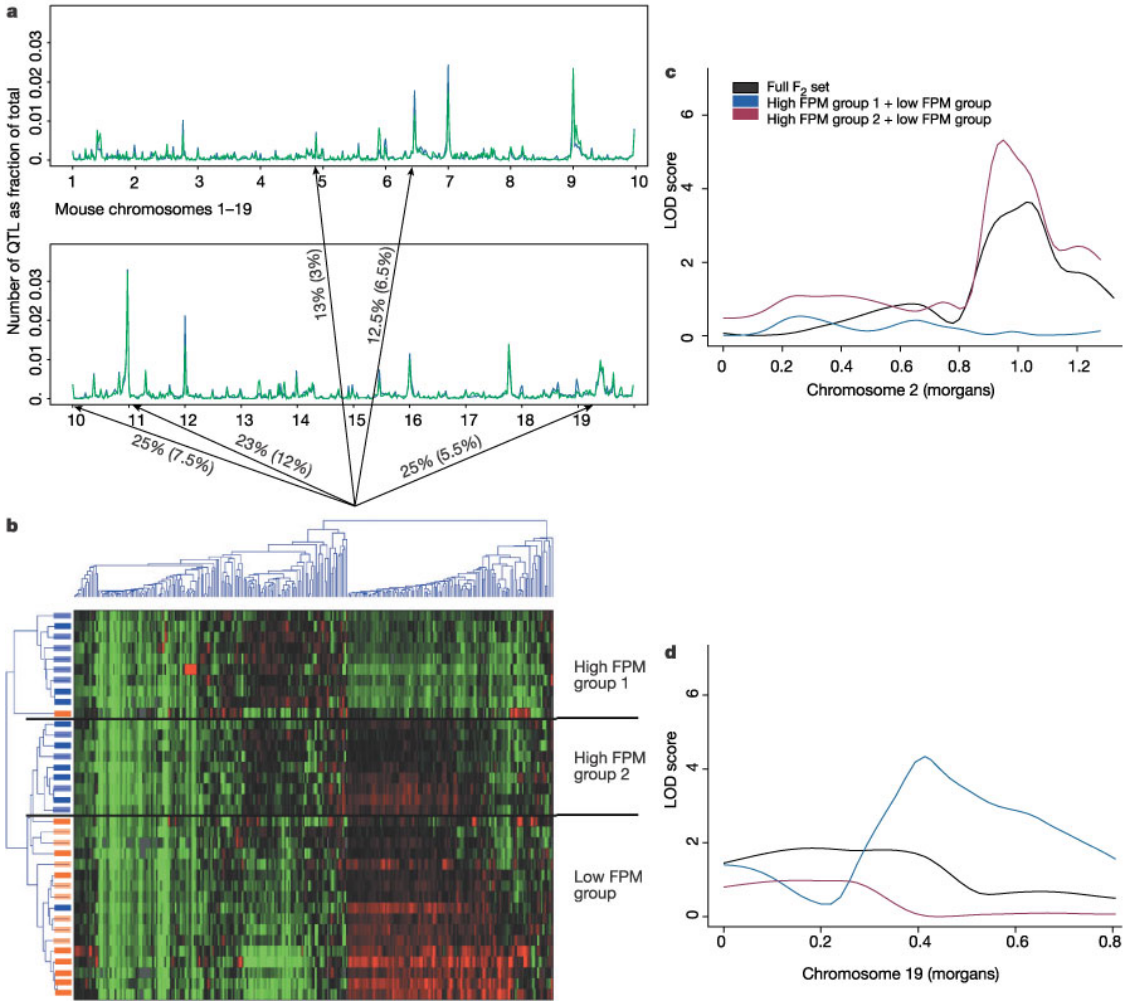
  $|M|$ = number of terms in model $M < 2^m$

$60,000
experiment

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

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

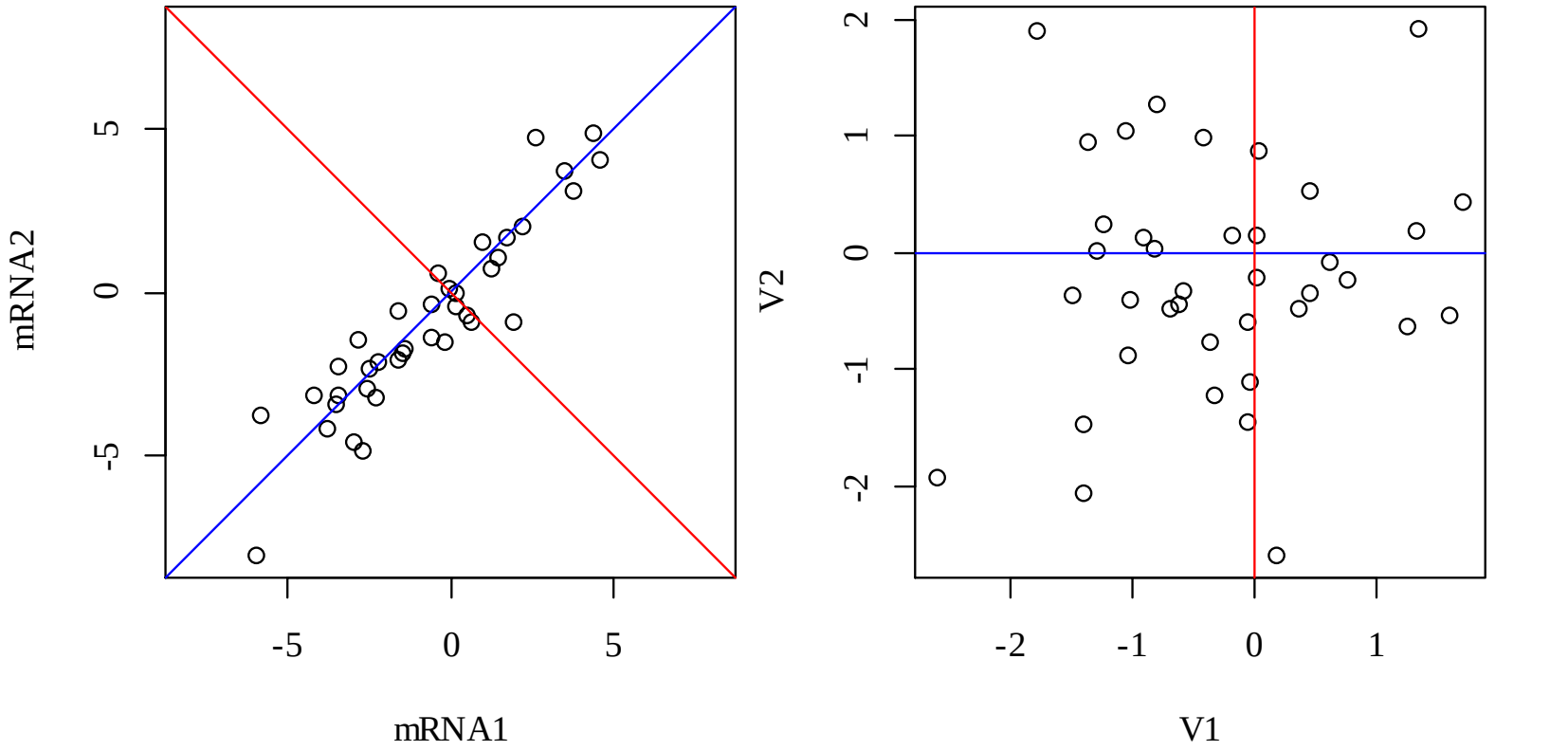




from gene expression to super-genes

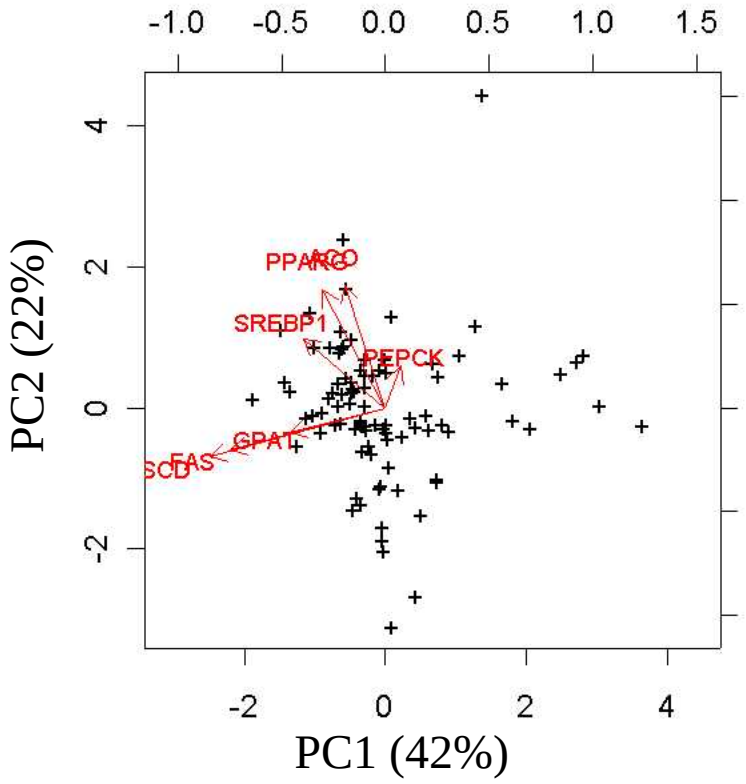
- PC or SVD decomposition of multiple traits
  - Y = t traits × n individuals
  - decompose as $Y = UDW^{T}$
    - U,W = ortho-normal transforms (eigen-vectors)
    - D = diagonal matrix with singular values
- transform problem to principal components
  - W_1 and W_2 uncorrelated "super-traits"
- interval map each PC separately
  - $W_1 = G^*_{1Q} + e^*_1$
- may only need to map a few PCs

PC simply rotates & rescales to find major axes of variation

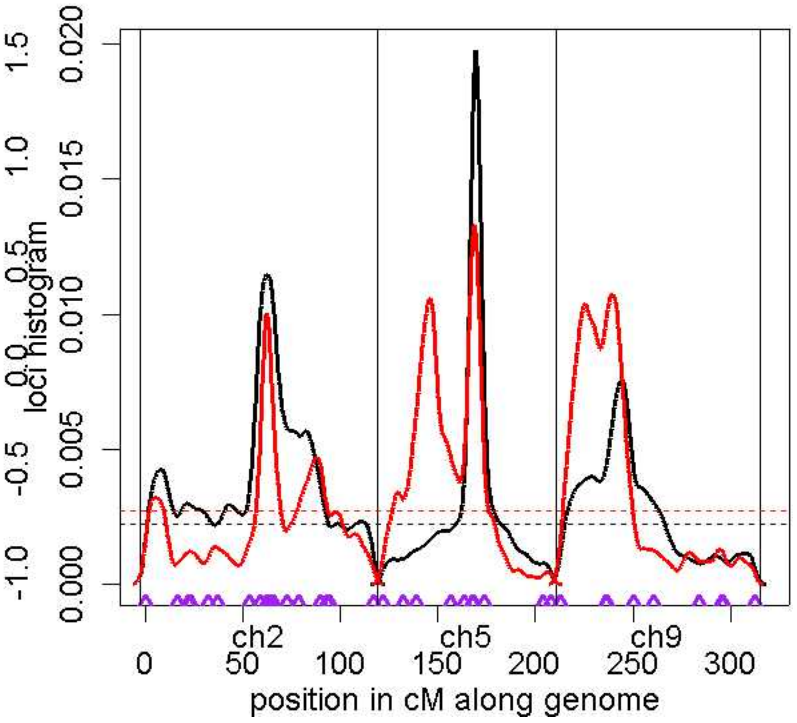


multivariate screen for gene expressing mapping

principal components



PC1(red) and SCD(black)

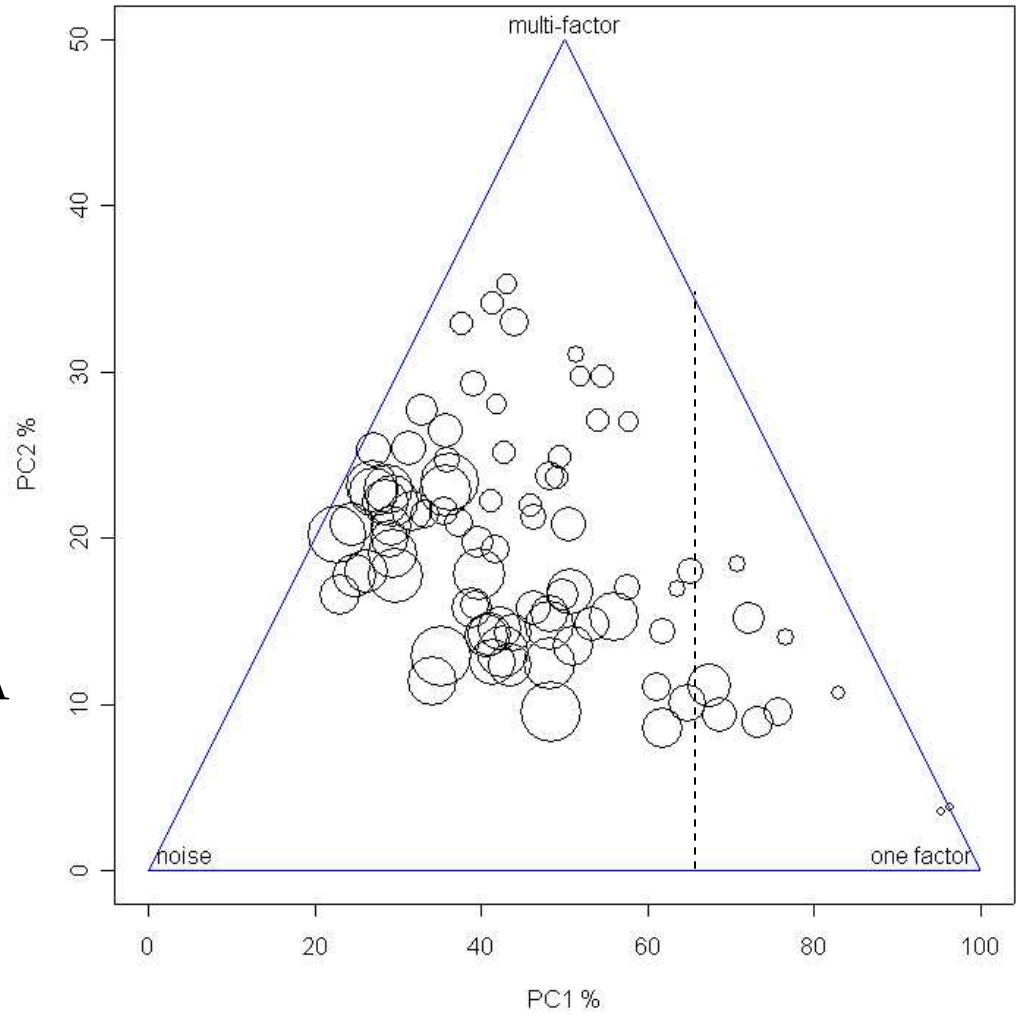


PC across microarray functional groups

1500+ mRNA of 30,000
85 functional groups
60 mice
2-35 mRNA / group
which are interesting?

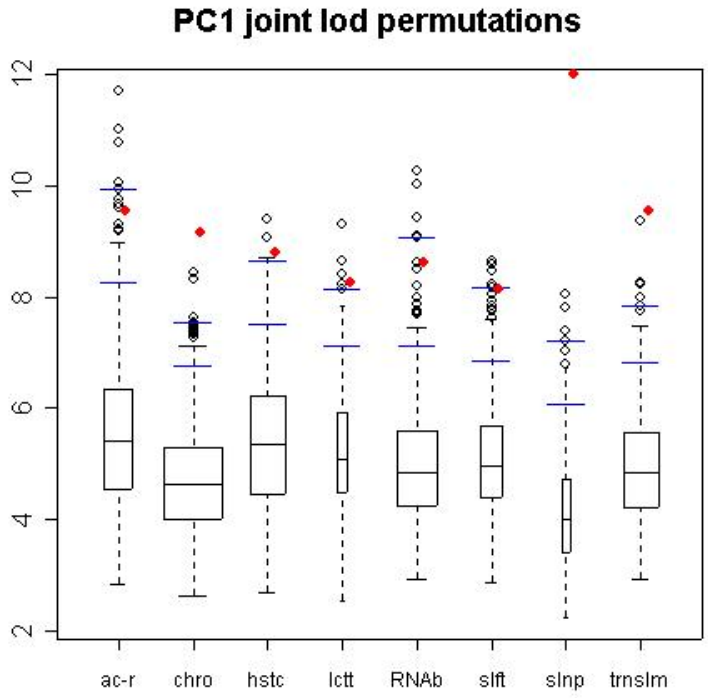
examine PC1, PC2

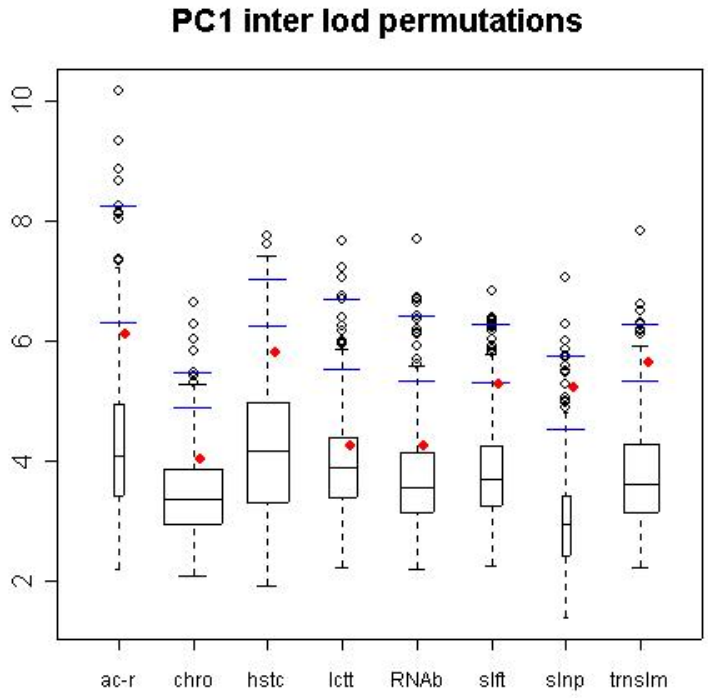
circle size = # unique mRNA



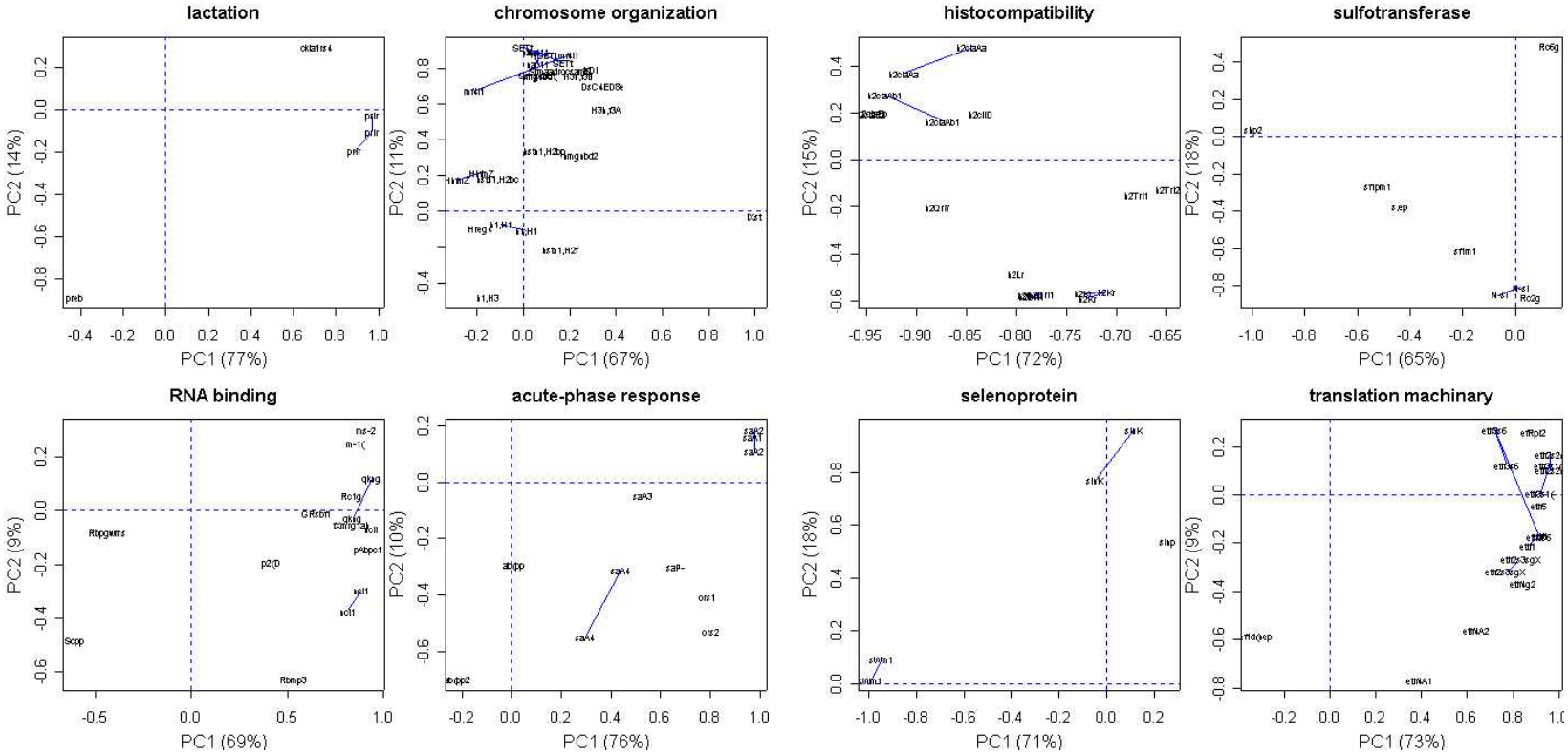
how well does PC1 do?

lod peaks for 2 QTL at best pair of chr vs. 500 permutations

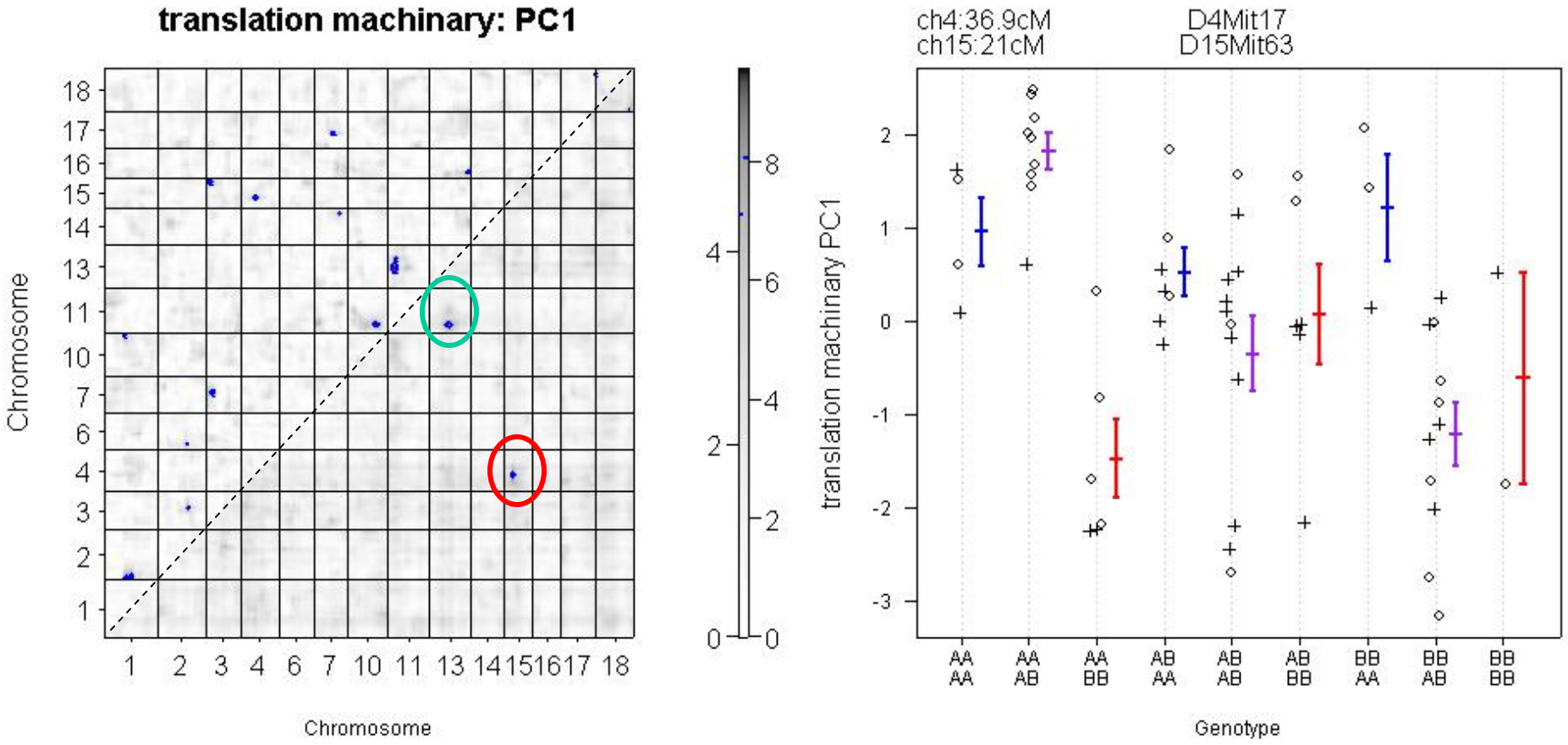


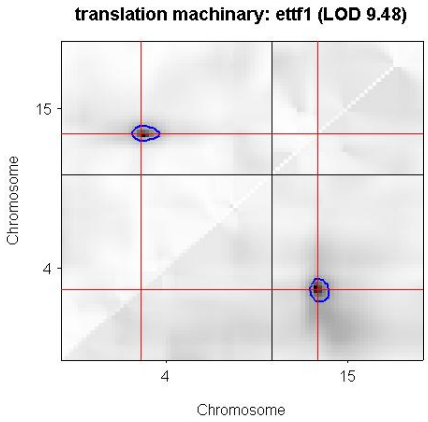


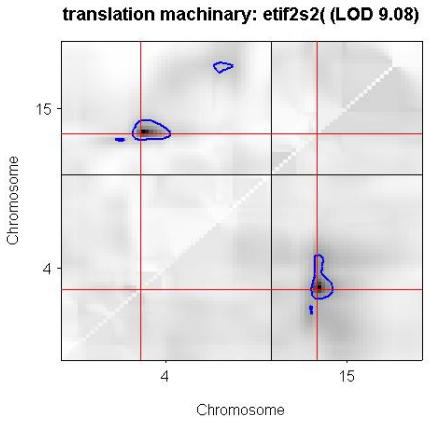
factor loadings for PC1&2

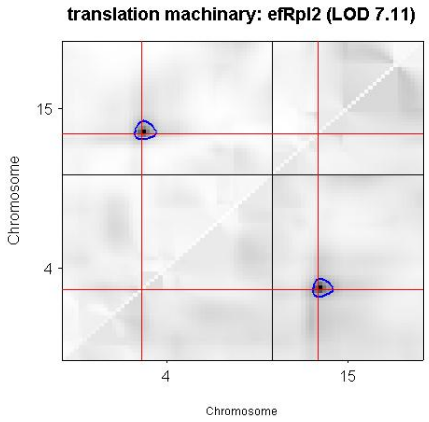


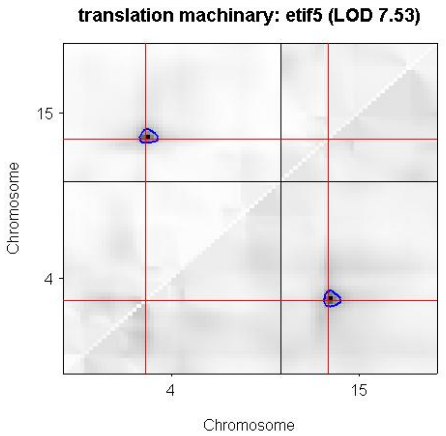
focus on translation machinery (EIF)

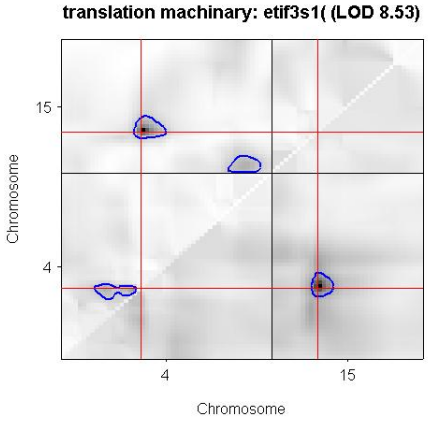


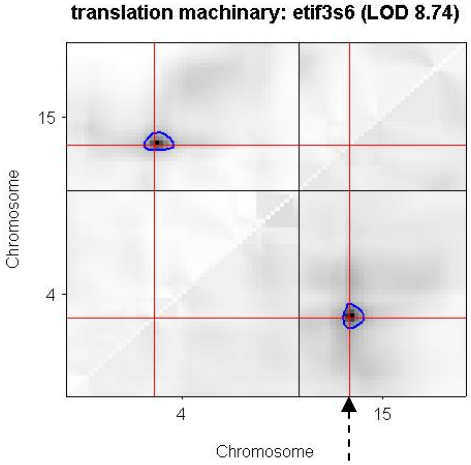

translation machinary: ettf1 (LOD 9.48)
Chromosome
15
4
4
15
Chromosome

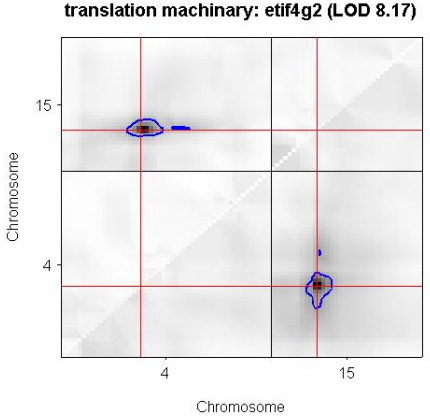

translation machinary: etif2s2((LOD 9.08)
Chromosome
15
4
4
15
Chromosome

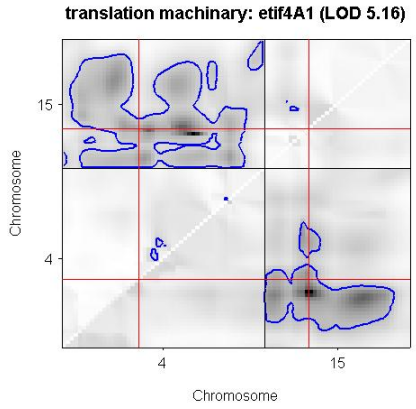

translation machinary: efRpl2 (LOD 7.11)
Chromosome
15
4
4
15
Chromosome

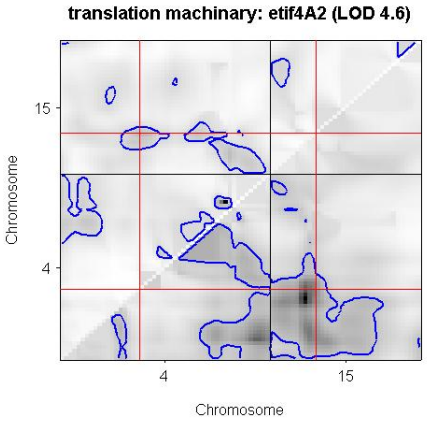

translation machinary: etif5 (LOD 7.53)
Chromosome
15
4
4
15
Chromosome

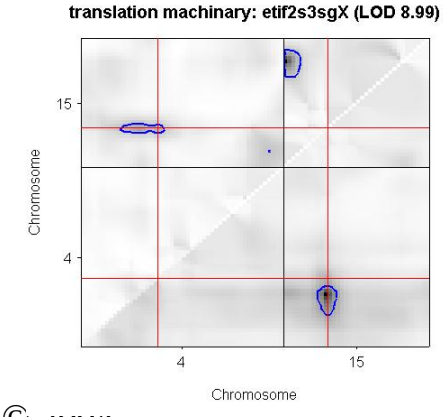

translation machinary: etif3s1((LOD 8.53)
Chromosome
15
4
4
15
Chromosome

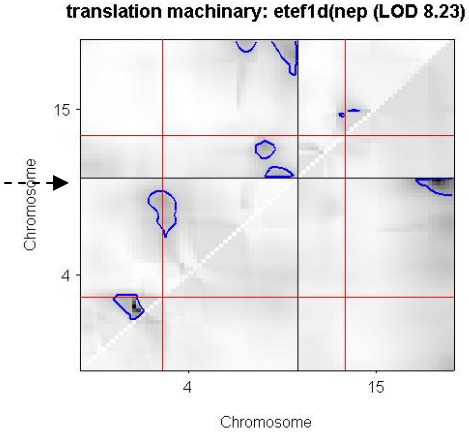

translation machinary: etif3s6 (LOD 8.74)
Chromosome
15
4
4
15
Chromosome


translation machinary: etif4g2 (LOD 8.17)
Chromosome
15
4
4
15
Chromosome

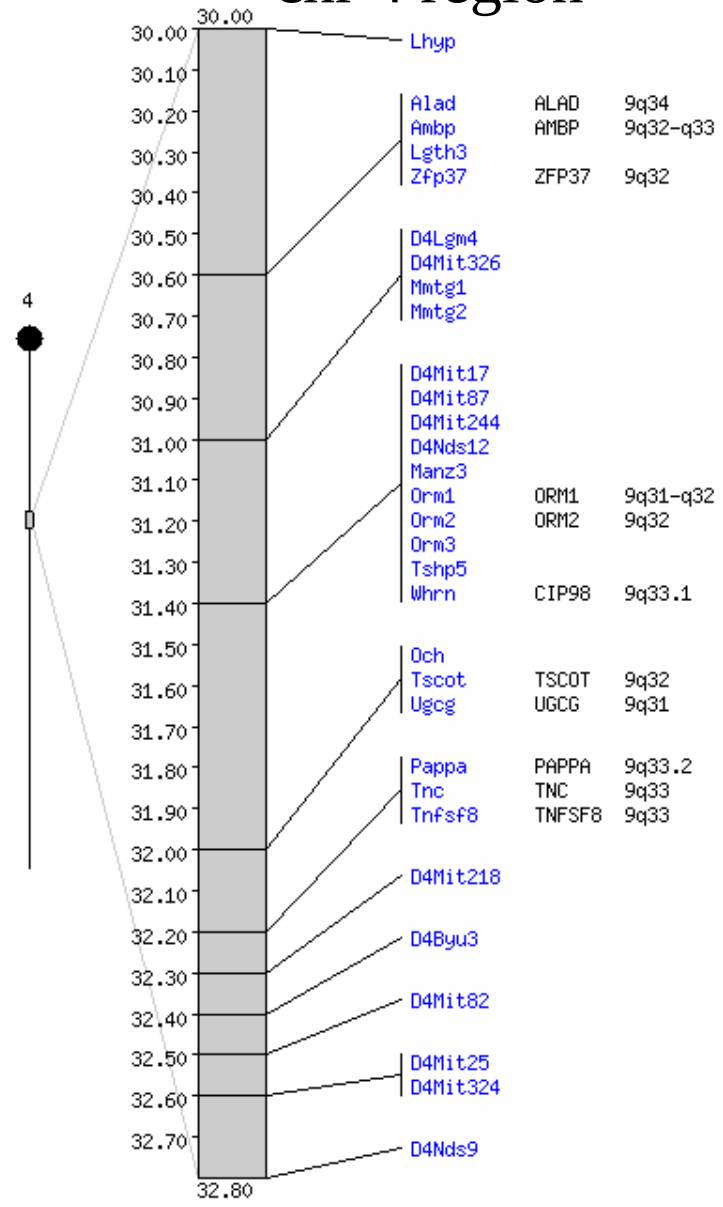

translation machinary: etif4A1 (LOD 5.16)
Chromosome
15
4
4
15
Chromosome


translation machinary: etif4A2 (LOD 4.6)
Chromosome
15
4
4
15
Chromosome

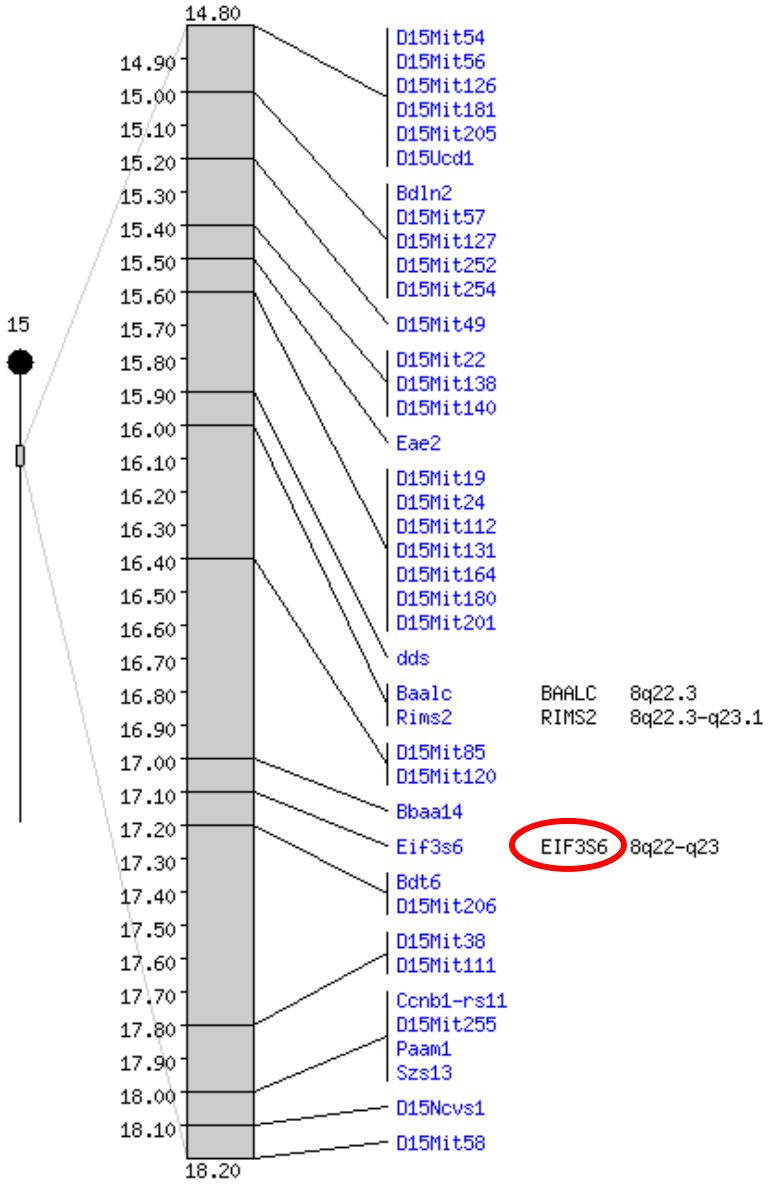

translation machinary: etif2s3sgX (LOD 8.99)
Chromosome
15
4
4
15
Chromosome


translation machinary: etef1d(nep (LOD 8.23)
Chromosome
15
4
4
15
Chromosome

chr 4 region

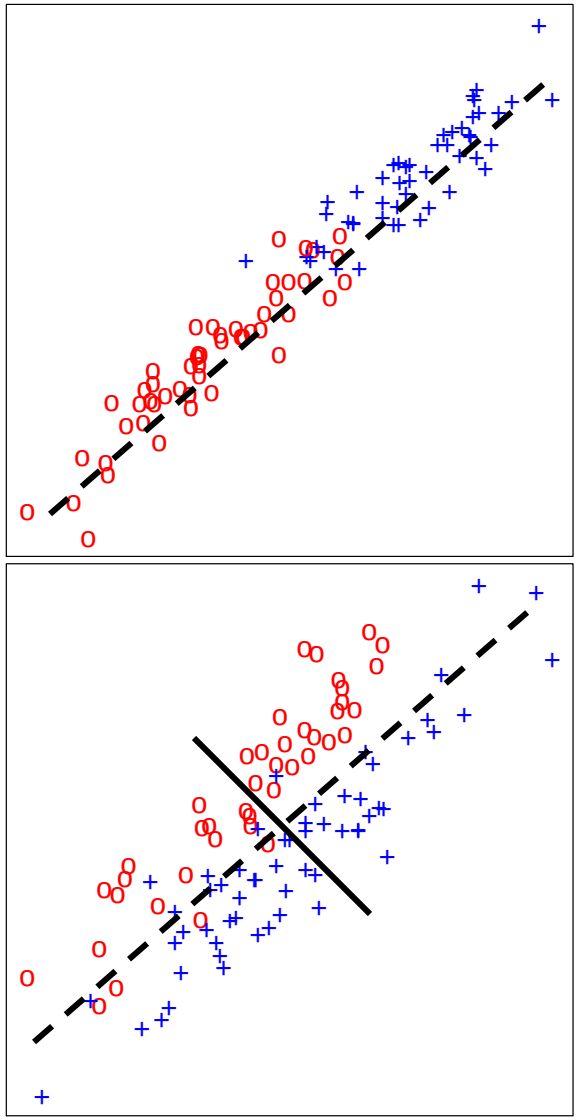


chr 15 region

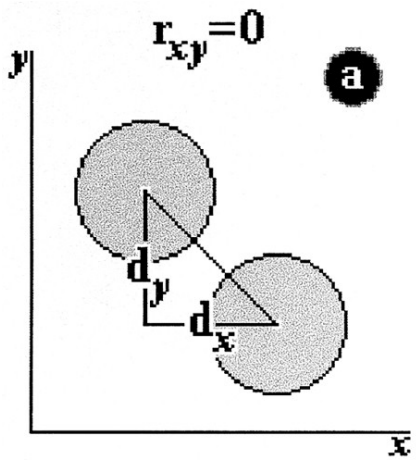


improvements on PC?

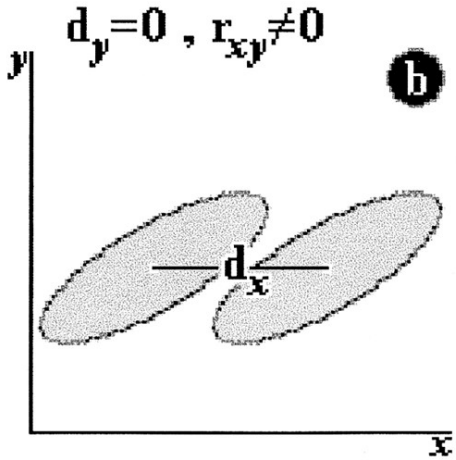
- what is our goal?
  - reduce dimensionality
  - focus on QTL
- PC reduces dimensionality
  - but may not relate to genetics
- discriminant analysis (DA)
  - rotate to improve discrimination
  - redo at each putative QTL
  - Gilbert and le Roy (2003,2004)



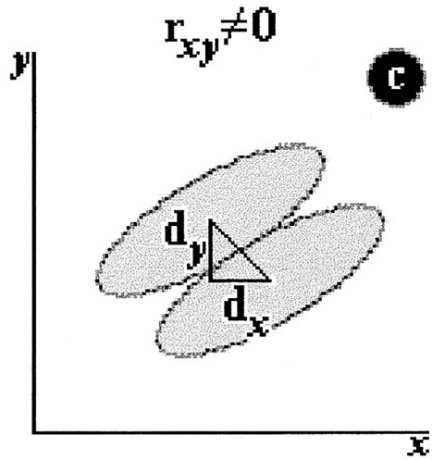
genetic & environmental correlation with multiple traits



genetic only



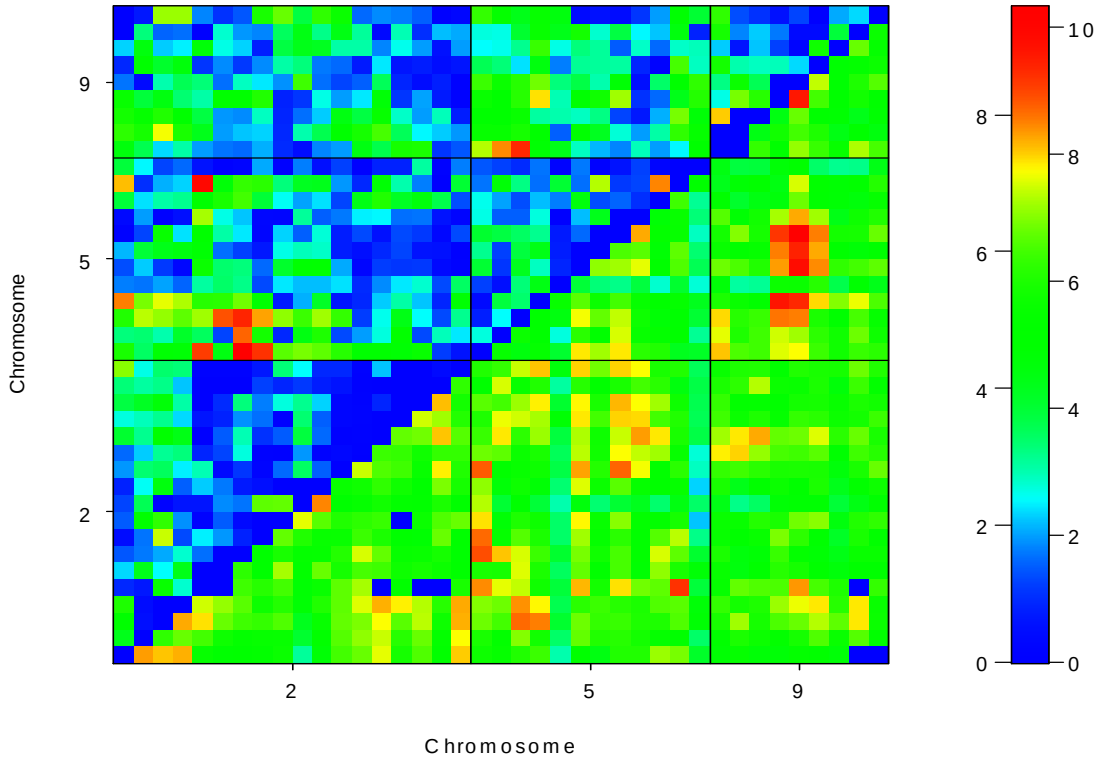
environmental only



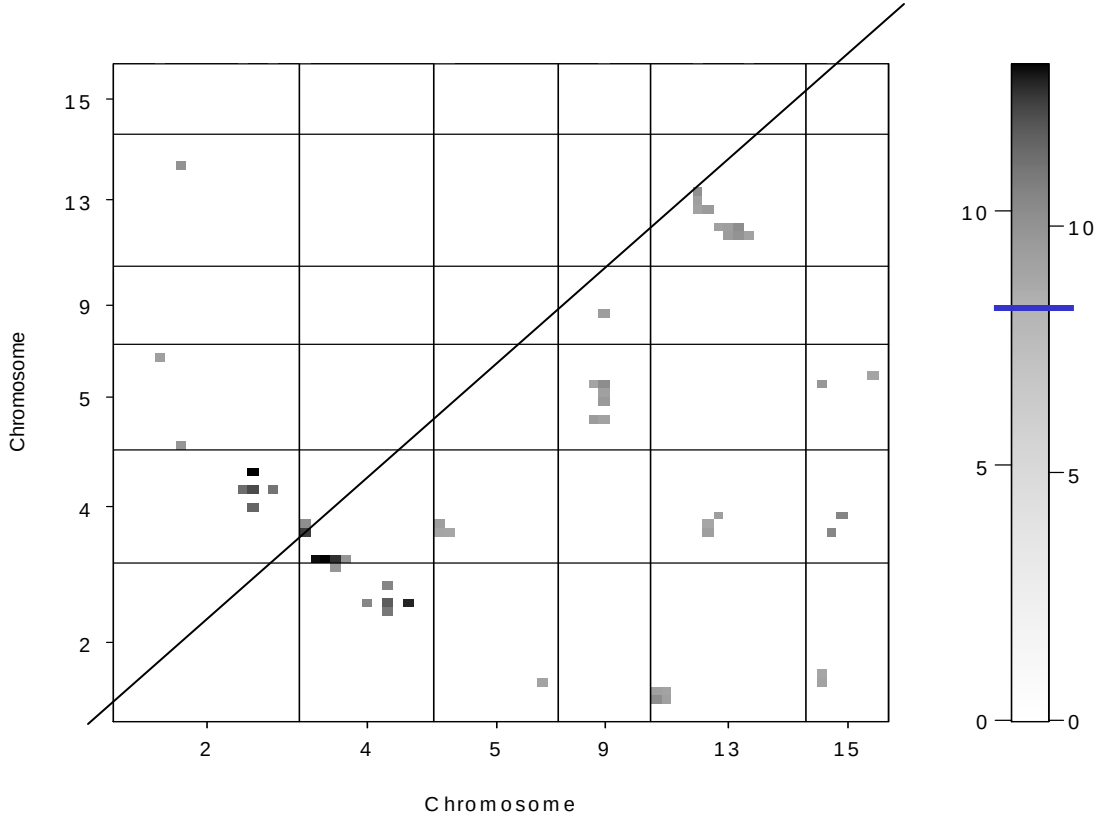
both

Korol et al. (2001)

discriminant analysis by marker pairs for SCD1-influencing chromosomes



DA for more chromosomes (mask values below 8)



what is the biological goal?

- understand biology of diabetes & obesity
- find genes influencing mRNA expression
  - localize genomic regions of high influence
    - coordinated regulation of many mRNA?
  - search databases for candidate genes there
- find mRNA expression with strong signals
  - prioritize subset of 30,000 mRNA
  - find genomic regions that influence them
- conduct followup experiments
  - new genetic crosses, more tissues
  - detailed assays of biochemical pathways