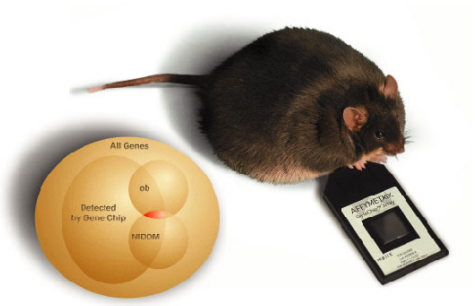


Graphical Diagnostics for Multiple QTL Investigation

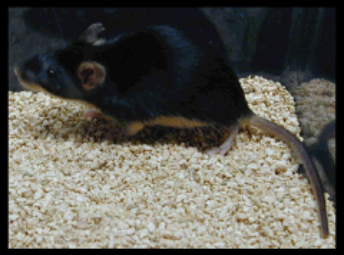
Brian S. Yandell
University of Wisconsin-Madison

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

- studying diabetes with microarrays
- taking a multiple QTL approach
- handling high throughput phenotypes
- designing for expensive phenotypes

Insulin Resistant Mice

Bill Dove



BTBR strain

glucose insulin

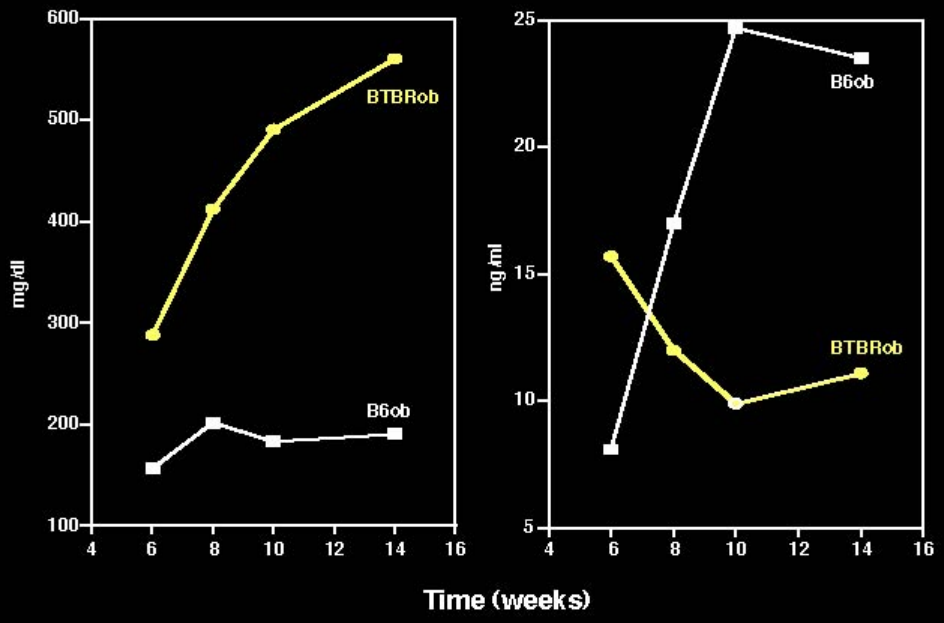
insulin resistance alleles

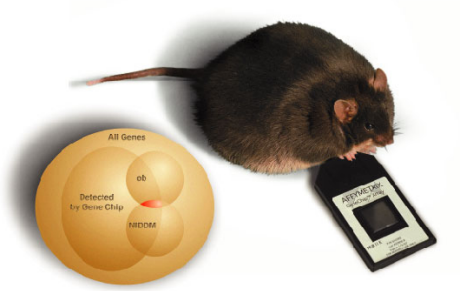
\+

??? → diabetes

obesity

(courtesy AD Attie)



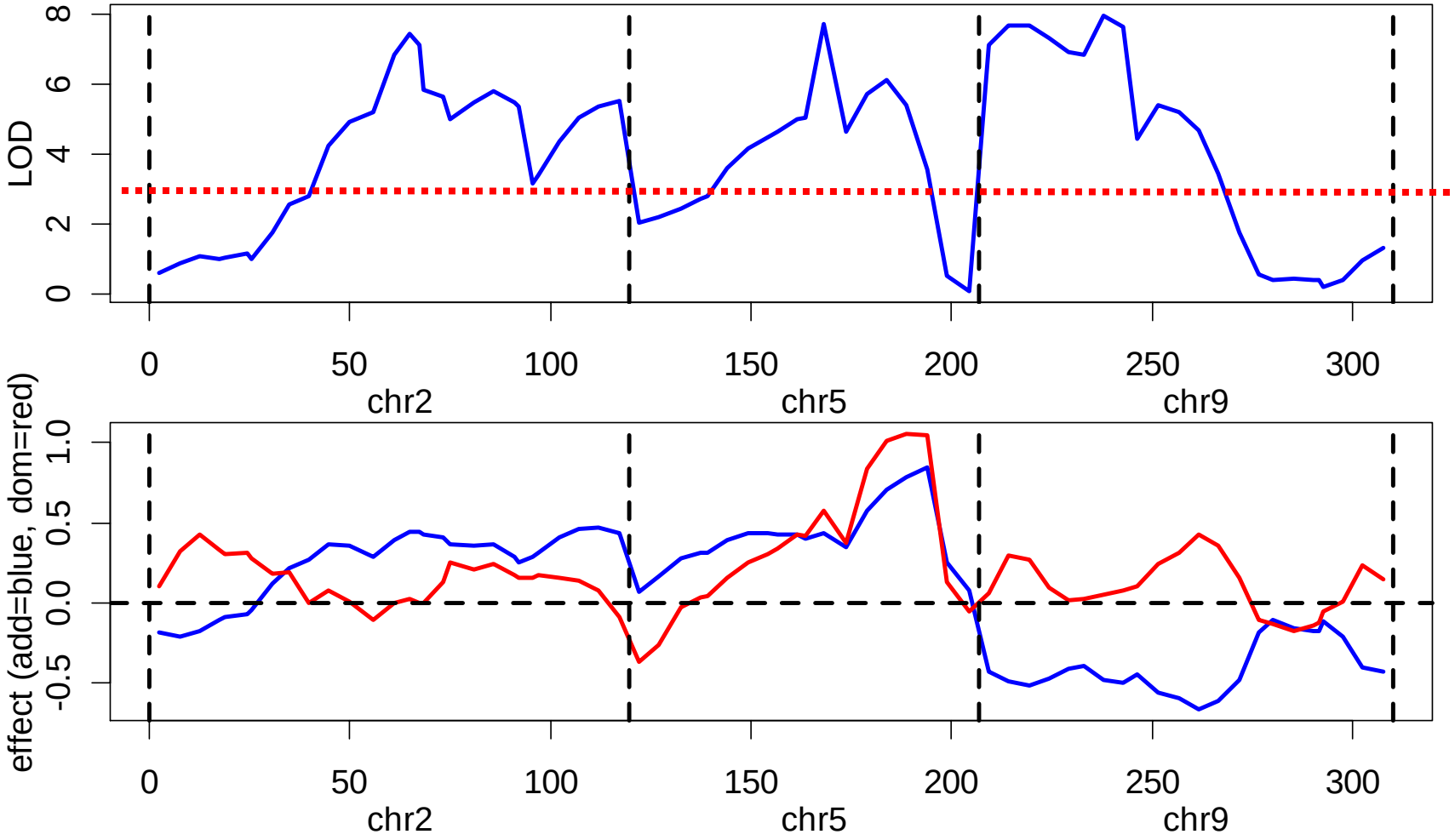


1. studying diabetes in an F2

- 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.)

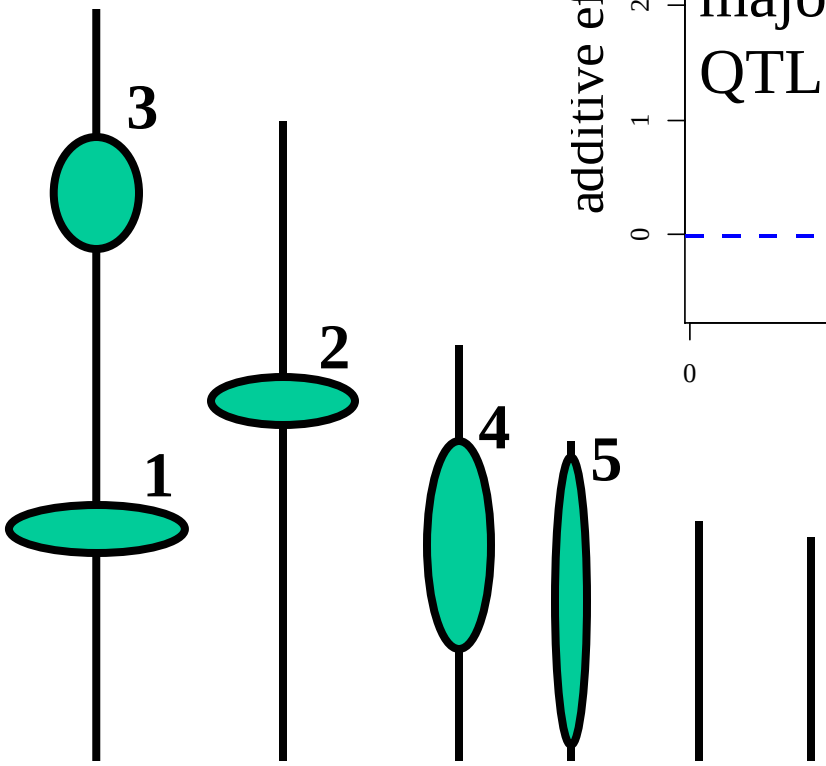
mRNA expression as phenotype:

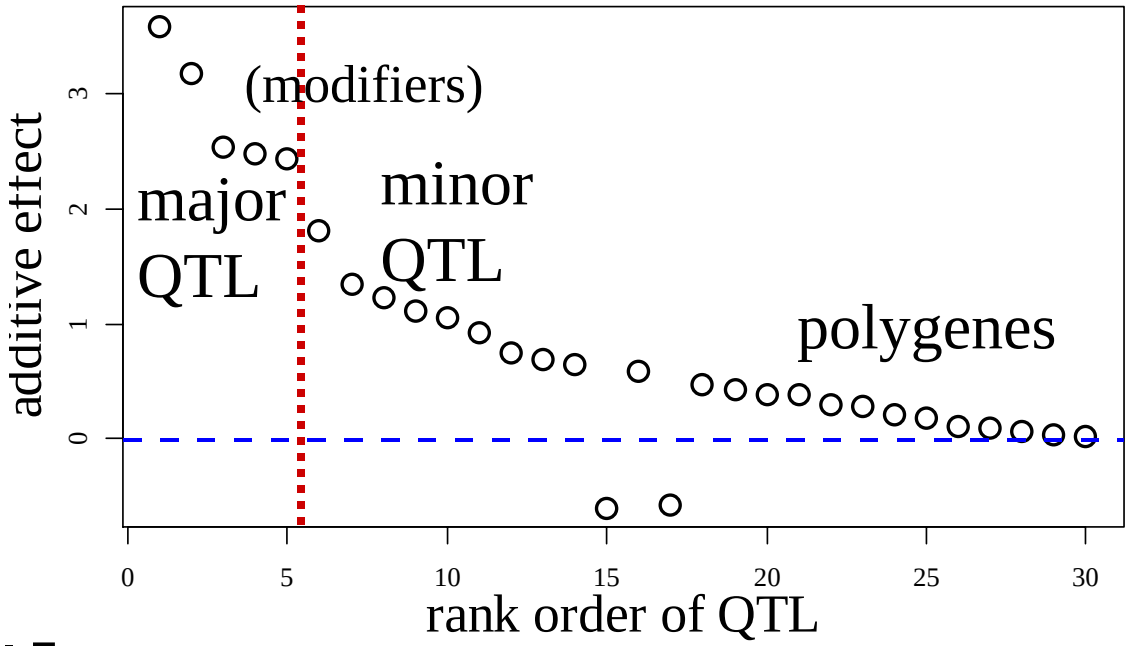
interval mapping for SCD1 is complicated



Pareto diagram of QTL effects

major QTL on linkage map





2. taking a multiple QTL approach

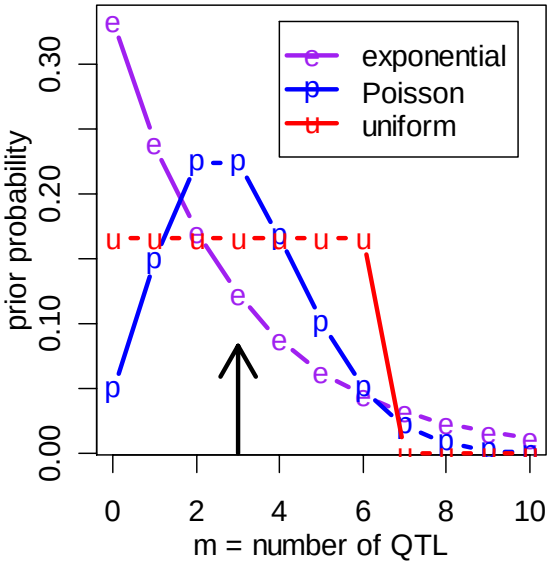
- improve statistical power, precision
  - increase number of QTL detected
  - better estimates of loci: less bias, smaller intervals
- improve inference of complex genetic architecture
  - patterns and individual elements of epistasis
  - appropriate estimates of means, variances, covariances
    - 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

comparing QTL models

- balance model fit with model "complexity“
  - want best fit (maximum likelihood or posterior)
  - without too complicated a model
- information criteria or Bayes factor quantifies the balance
  - Bayes information criteria (BIC) for classical approach
  - Bayes factors (BF) for Bayesian approach
- find “better” models
  - avoid selection bias (see Broman 2001)
    - QTL of modest effect only detected sometimes
    - genotypic effects biased upwards when detected
  - stochastic QTL detection
    - avoid sharp in/out dichotomy
    - average over better models

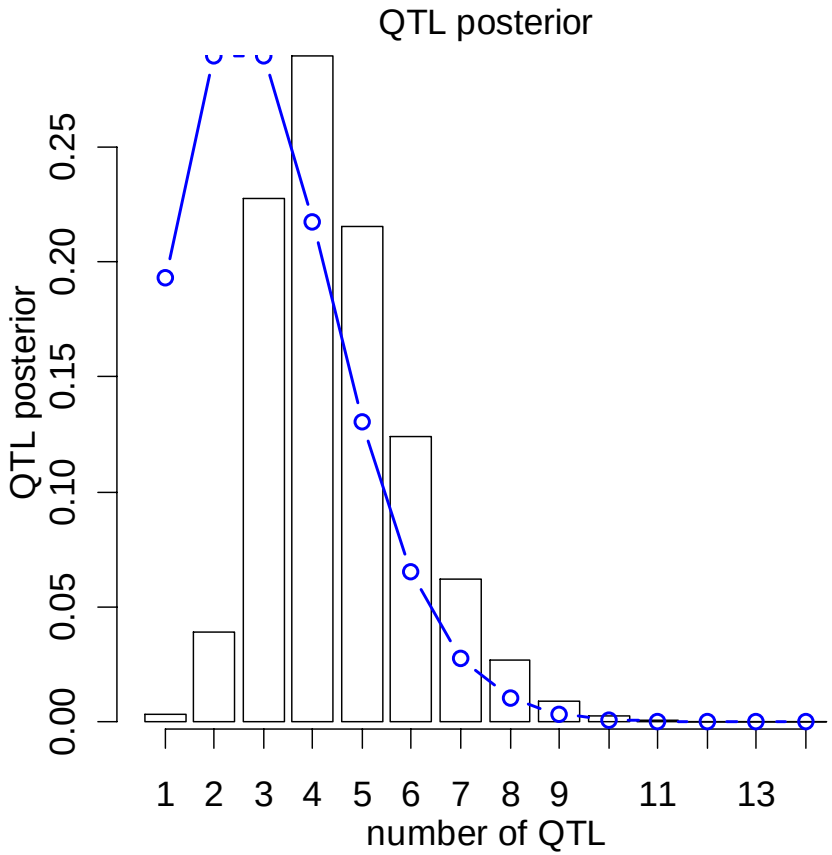
QTL Bayes factors

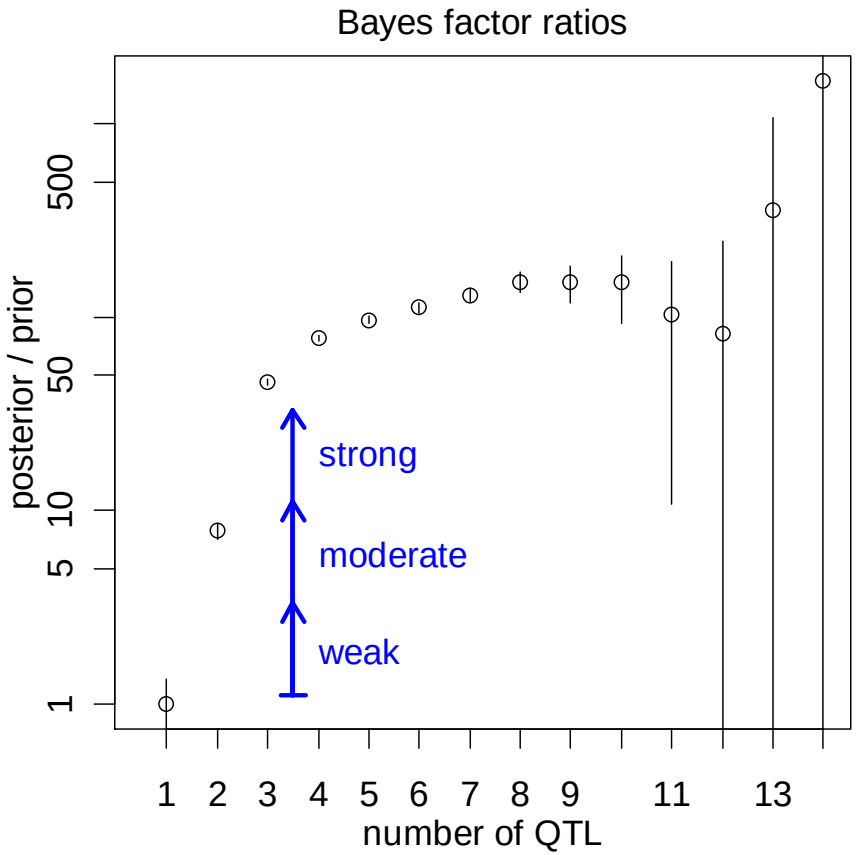
- BF = posterior odds / prior odds
- BF equivalent to BIC
  - simple comparison: 1 vs 2 QTL
    - same as LOD test
  - general comparison of models
  - want Bayes factor >> 1
- *m* = number of QTL
  - indexes model complexity
  - genetic architecture also important



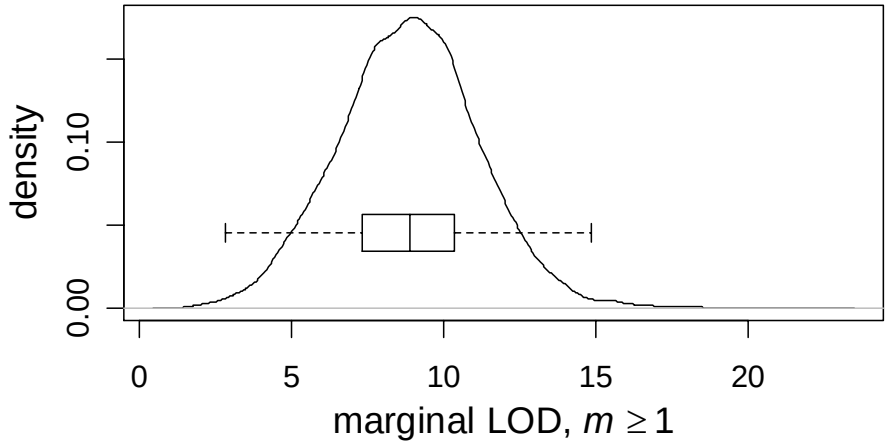
$$BF_{m,m+1} = \frac{\mathrm{pr}(m|\mathrm{data})/\mathrm{pr}(m)}{\mathrm{pr}(m+1|\mathrm{data})/\mathrm{pr}(m+1)}$$

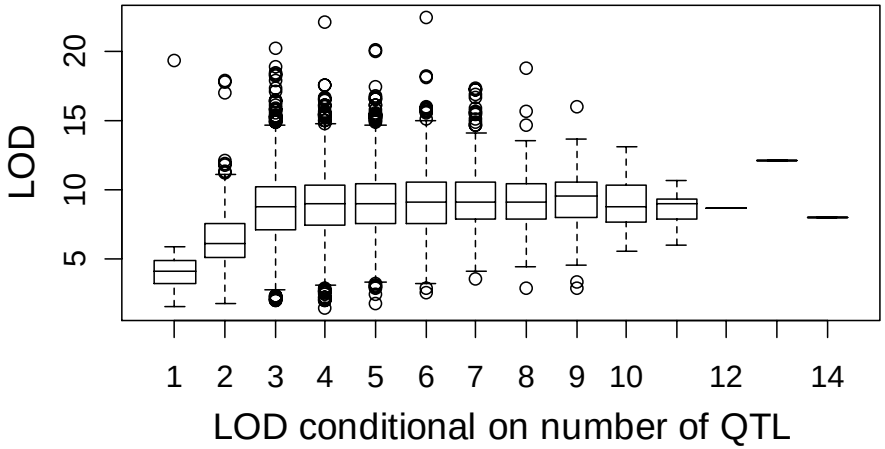
Bayesian model assessment: number of QTL for SCD1

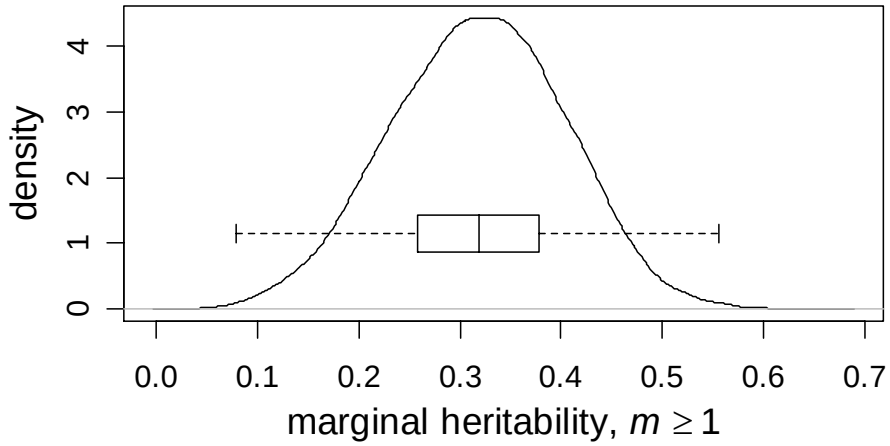


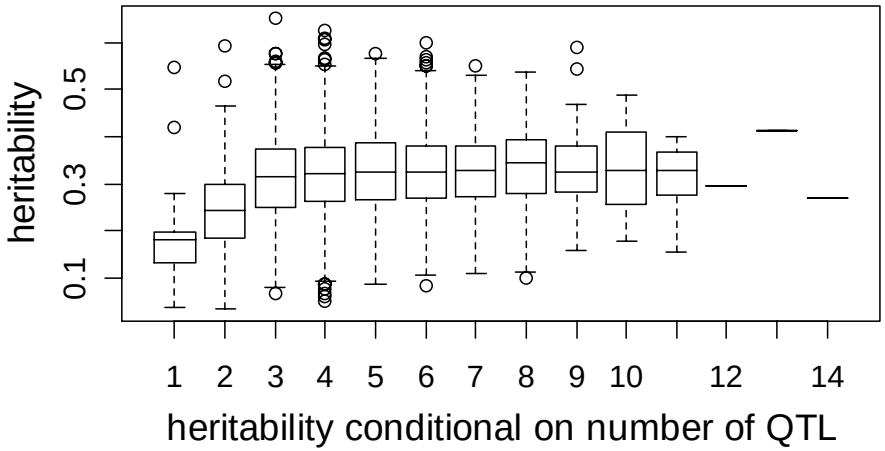


Bayesian LOD and h^2 for SCD1
(summaries from R/bim)



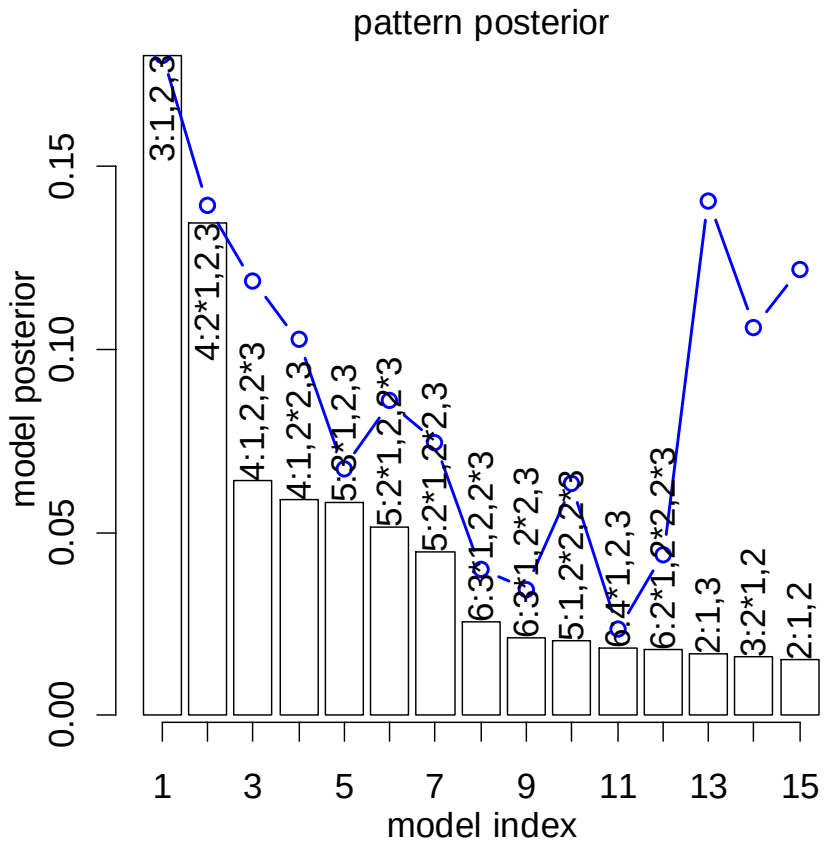


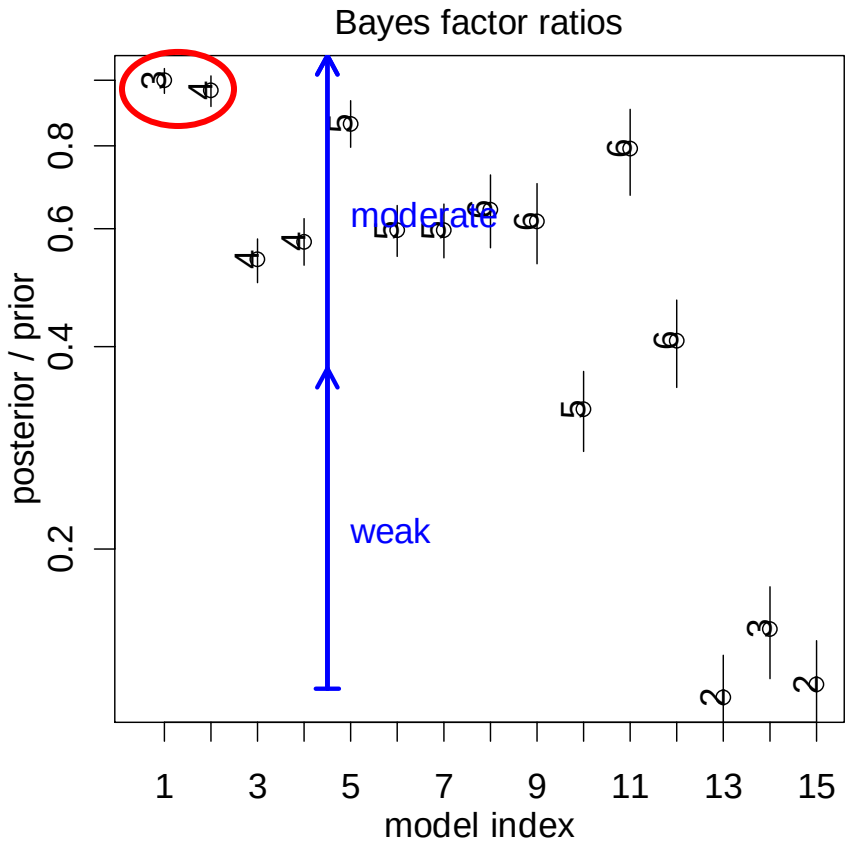




Bayesian model assessment

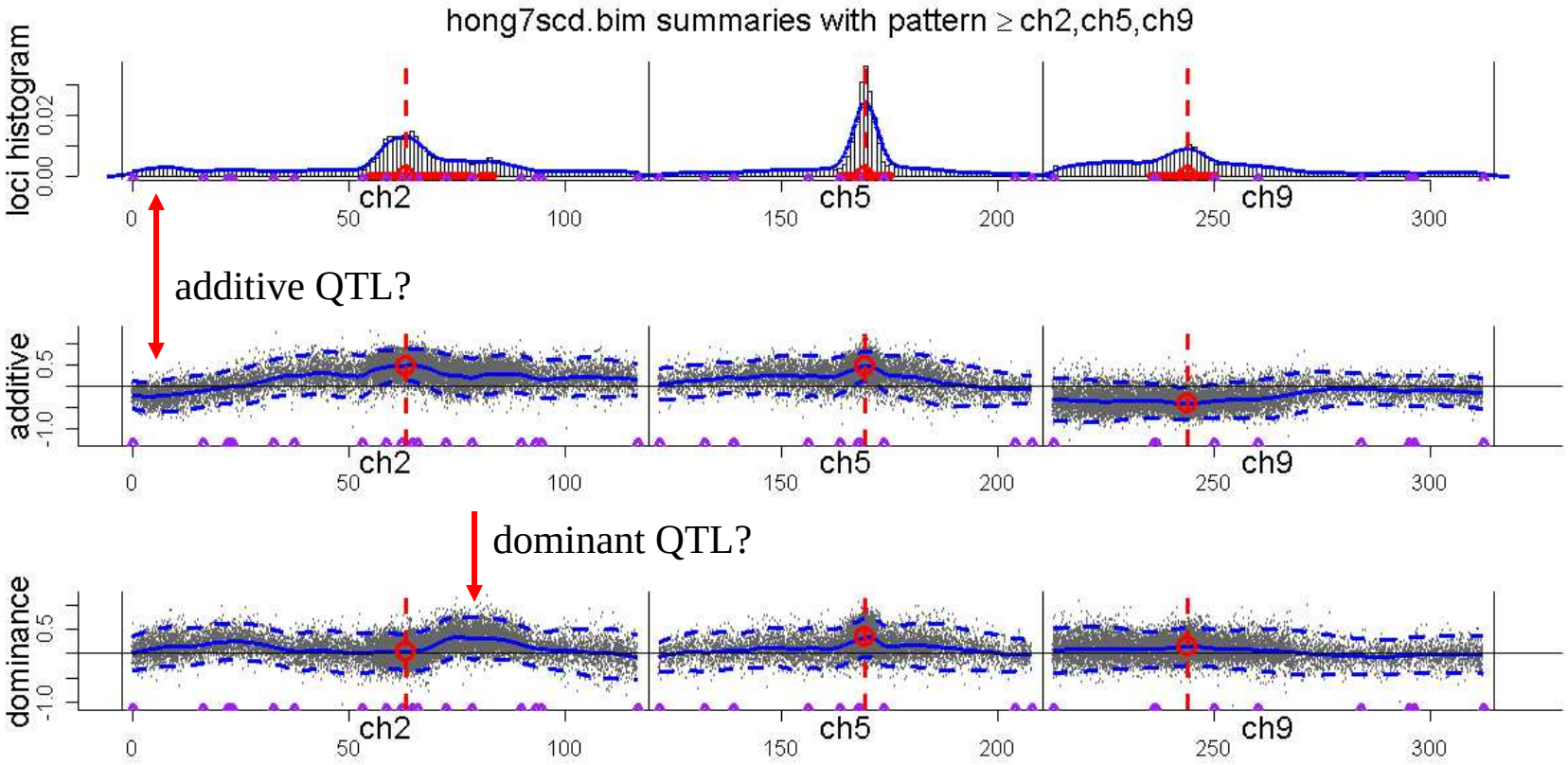
genetic architecture: chromosome pattern



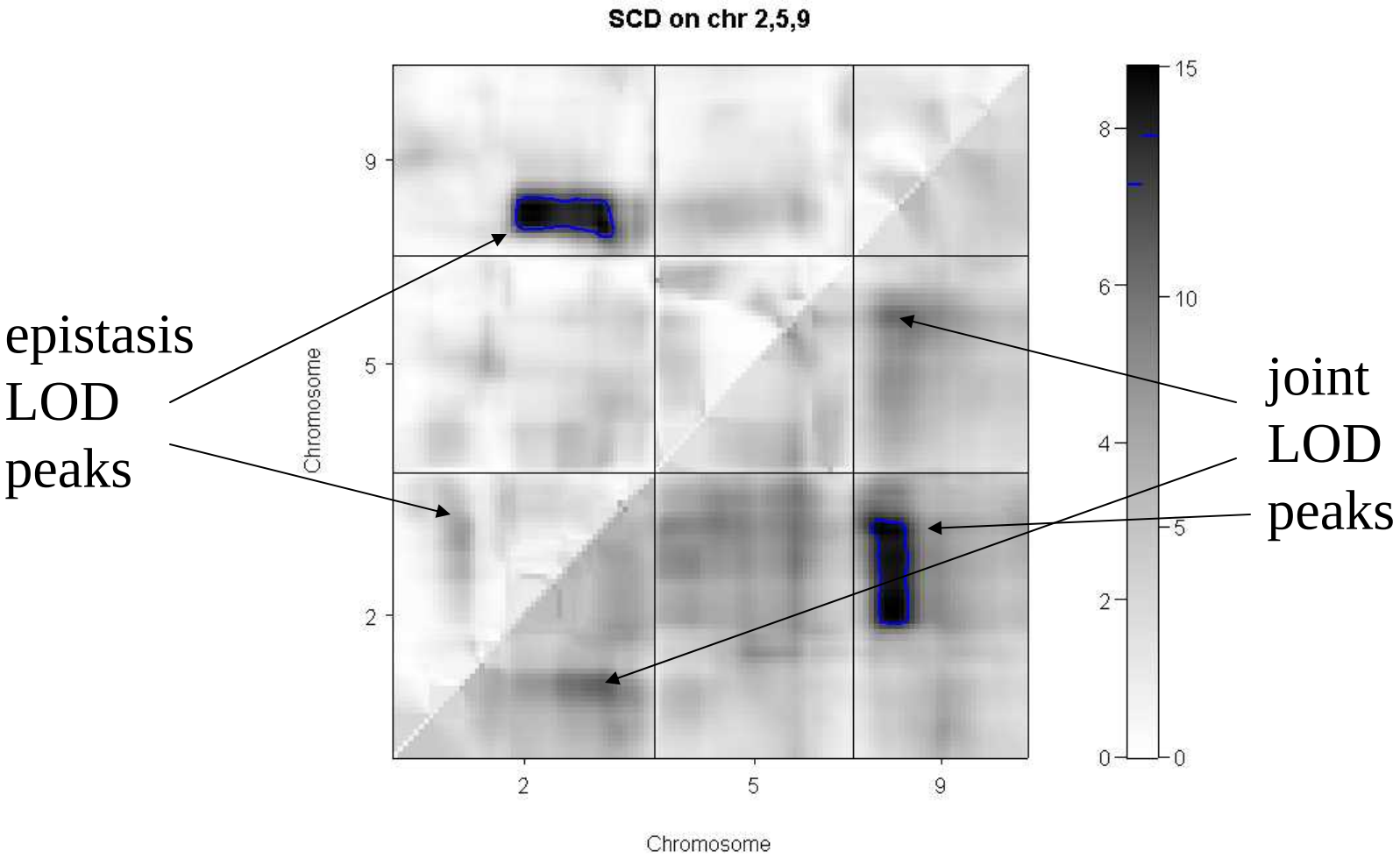


trans-acting QTL for SCD1

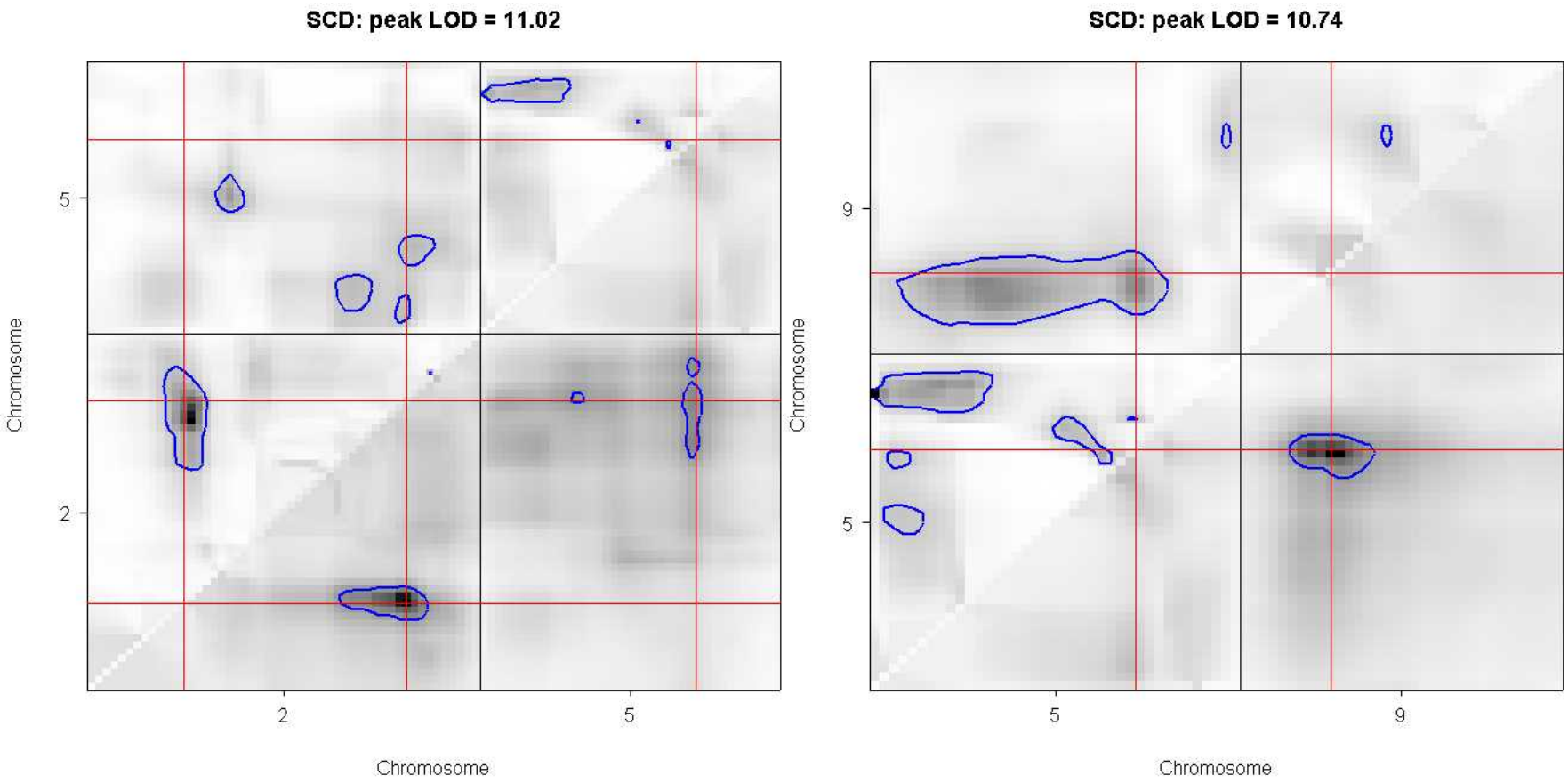
multiple QTL Bayesian model averaging



2-D scan: assumes only 2 QTL
(scantwo with HK method, from R/qtl)



sub-peaks can be easily overlooked



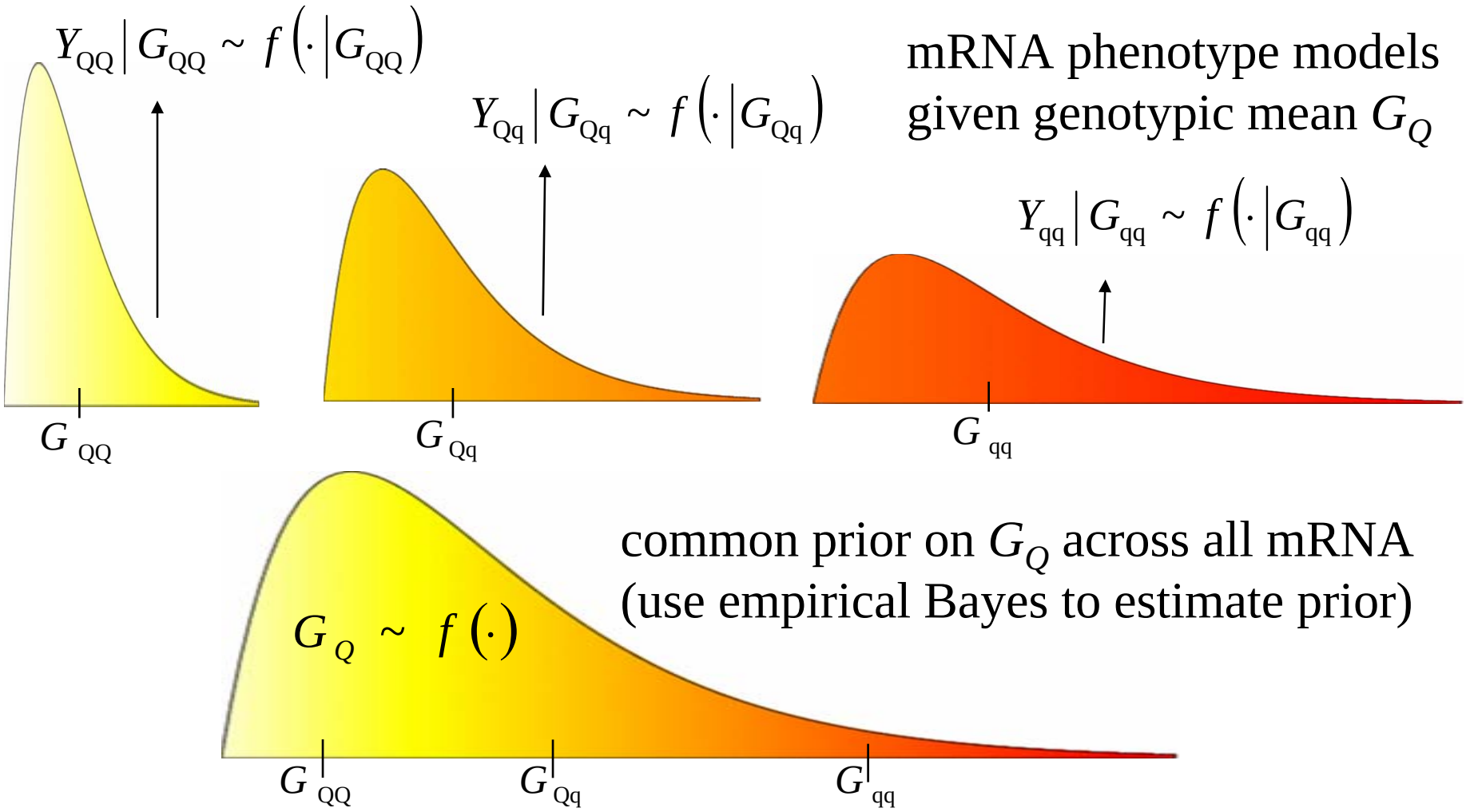

GeneChip
AFFYMETRIX

3. handling high throughput dilemma

- want to focus on gene expression network
  - ideally capture functional group in a few dimensions
  - allow for complicated genetic architecture
- may have multiple loci influencing expression
  - quick interval mapping assessment may be misleading
  - many genes with epistasis affect coordinated fashion?
- focus gene mapping using dimension reduction
  - initial screening using EB arrays to 2500+ mRNA
  - identify 85 functional groups from 1500+ mRNA
  - model selection for groups with stronger PC signals

hierarchical model for expression phenotypes

(EB arrays: Christina Kendziorski)



mRNA phenotype models given genotypic mean G_Q

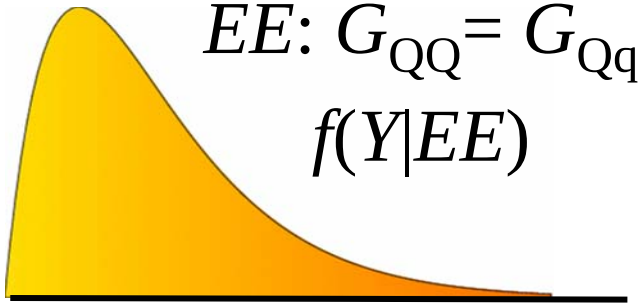
common prior on G_Q across all mRNA (use empirical Bayes to estimate prior)

For every mRNA transcript, two possible patterns (DE, EE)

no QTL present

EE: $G_{QQ} = G_{Qq}$

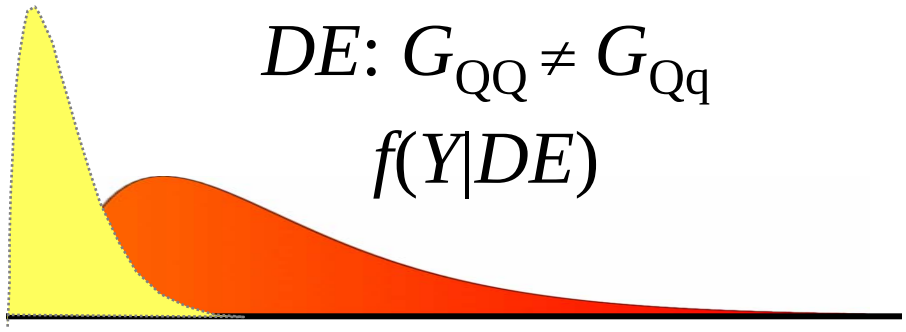
$f(Y|EE)$



QTL present

DE: $G_{QQ} \neq G_{Qq}$

$f(Y|DE)$



$$\text{odds} = \frac{P(DE|Y)}{P(EE|Y)} \frac{f(Y|DE)\,P(DE)}{f(Y|EE)\,P(EE)}$$

Empirical Bayes methods (EB arrays) make use all of the data to make mRNA-specific inferences.

hierarchical model across expression phenotypes
(Christina Kendziorski)

- vector of mRNA phenotypes organized by QTL genotype

  $Y = (Y_1,\ldots,Y_n) = (Y_{QQ}, Y_{Qq}, Y_{qq})$

  $Y \sim f(\, Y \mid \mu\,)$ if no QTL present

  $Y \sim f(Y_{QQ}|G_{QQ})\, f(Y_{Qq}|G_{Qq})\, f(Y_{qq}|G_{qq})$ if QTL present

- marginal for phenotype across possible genotypic means

  $Y \sim f_0(Y) = \int f(\mu)\, f(Y|\mu)\, \mathrm{d}\mu$ if no QTL present

  $Y \sim f_1(Y) = f_0(Y_{QQ})\, f_0(Y_{Qq})\, f_0(Y_{qq})$ if QTL present

- mixture across possible patterns of expression

  $Y \sim p_0 f_0(Y) + p_1 f_1(Y)$

  p_1 = prior probability of QTL present

  (could allow more possibilities—gene action, multiple QTL)

PC across microarray functional groups

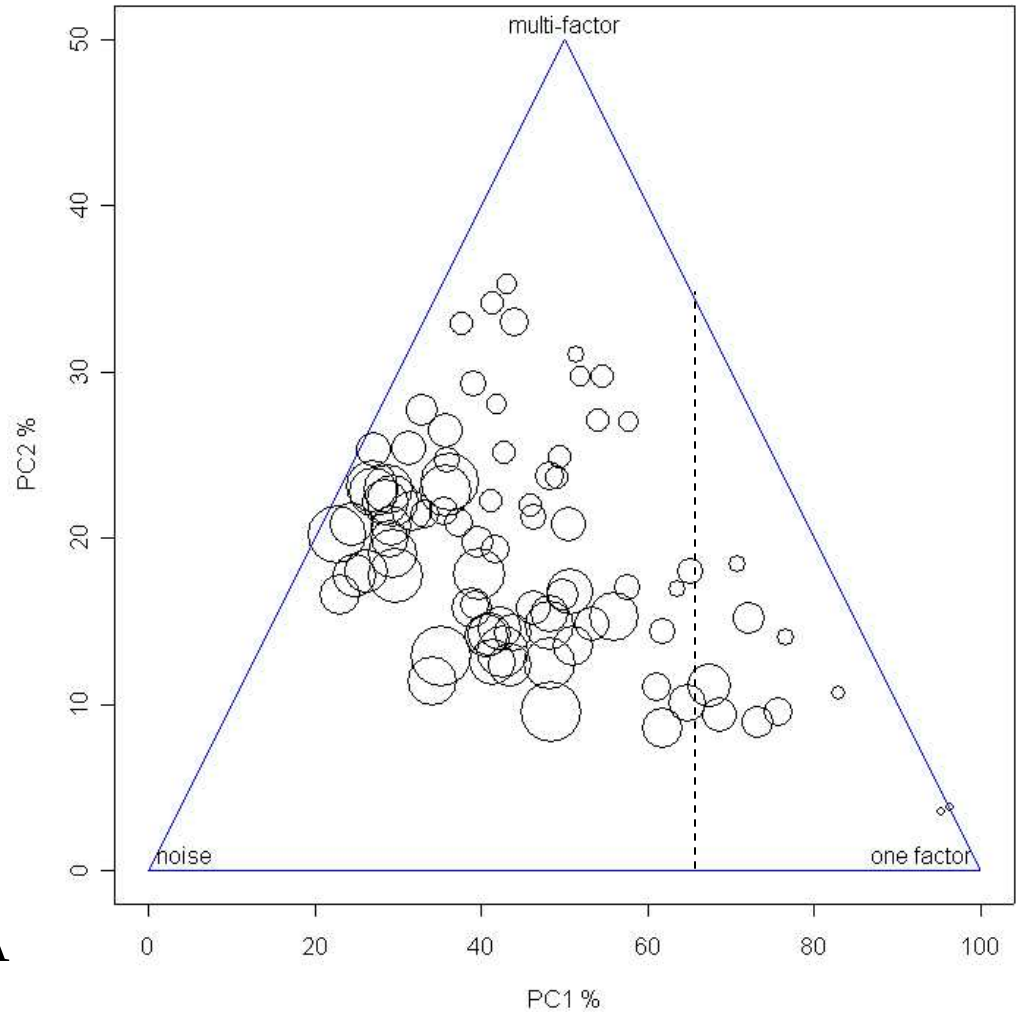
Affy chips on 60 mice
~40,000 mRNA

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

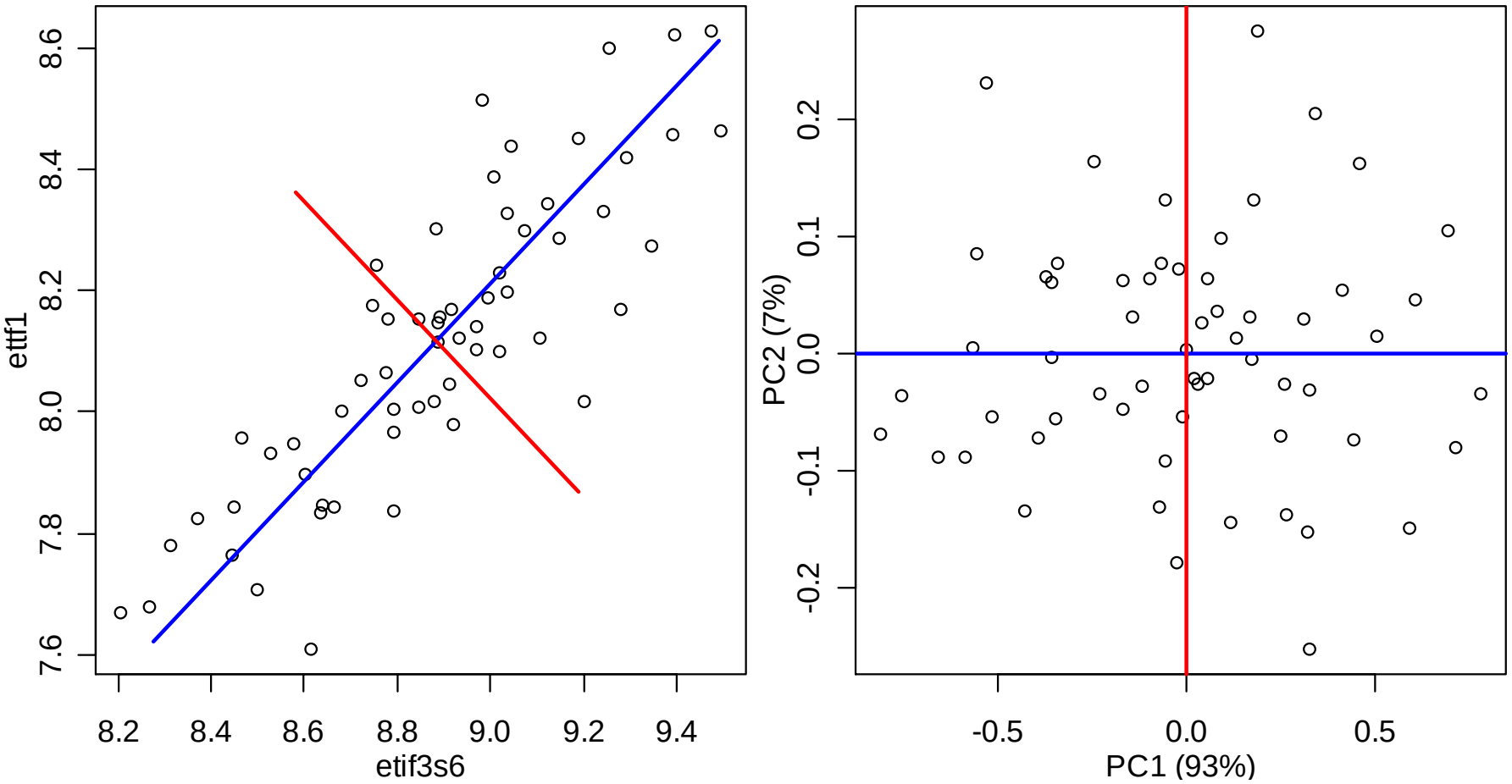
1500+ organized in
85 functional groups
2-35 mRNA / group

which are interesting?
examine PC1, PC2

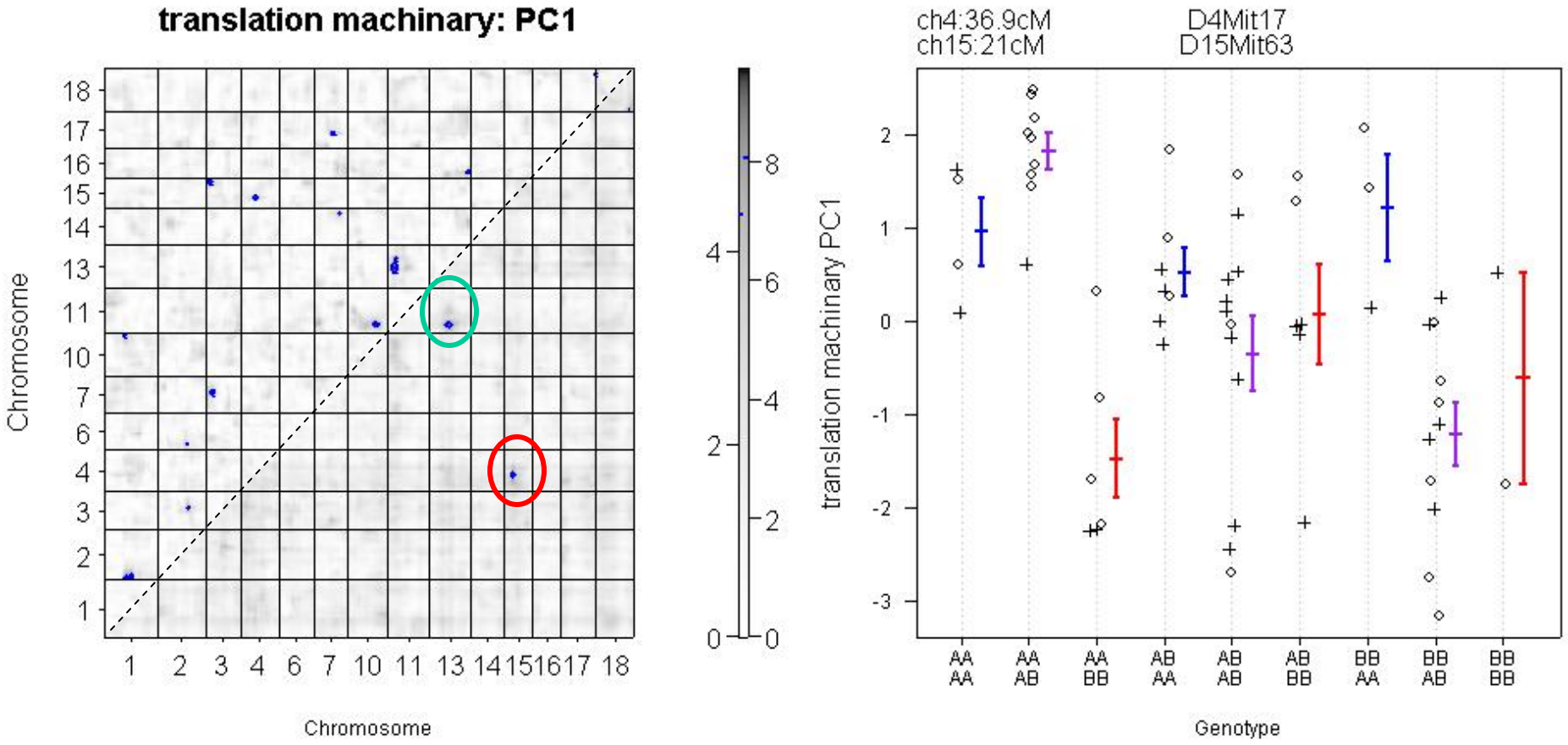
circle size = # unique mRNA

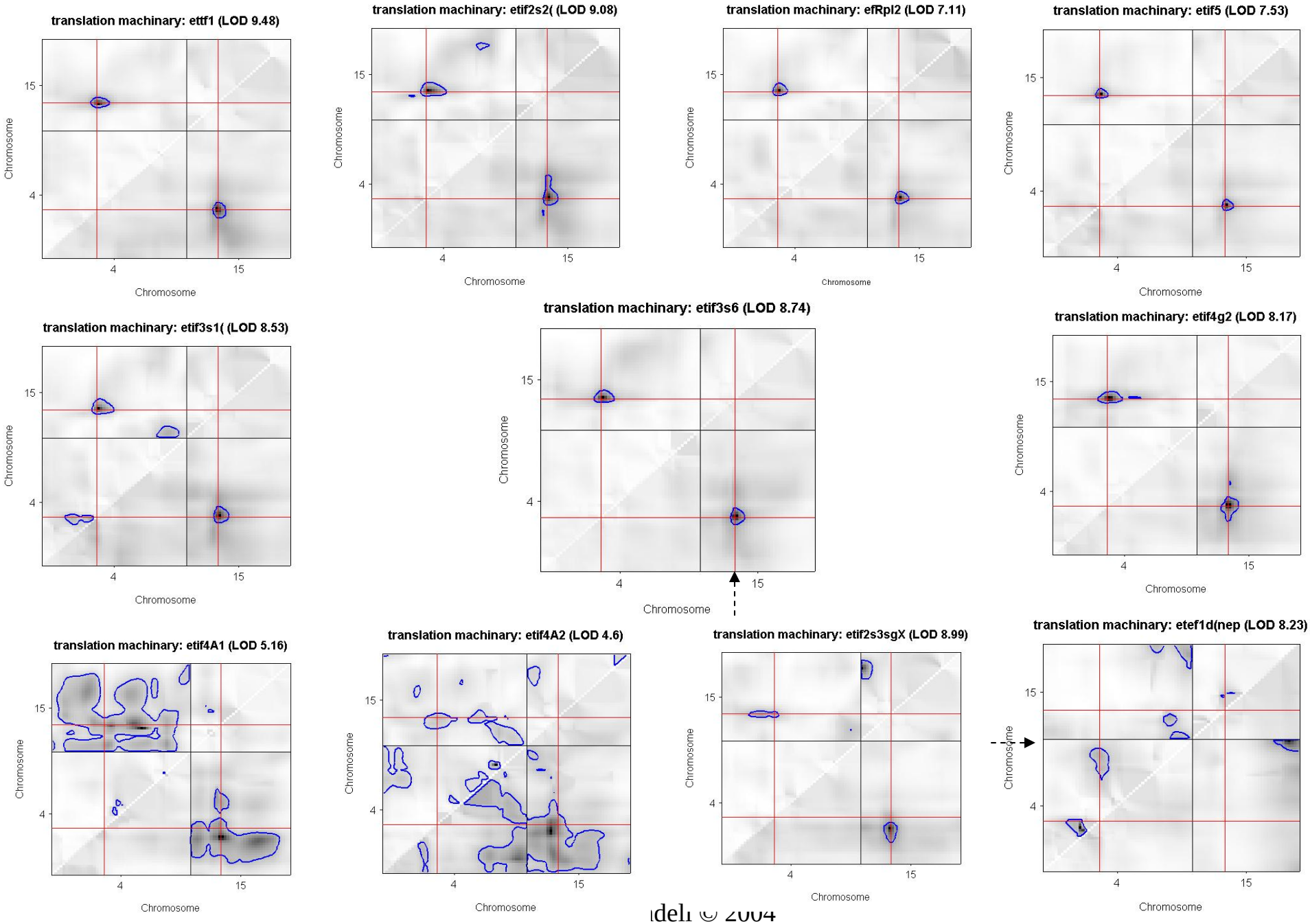


PC for two correlated mRNA



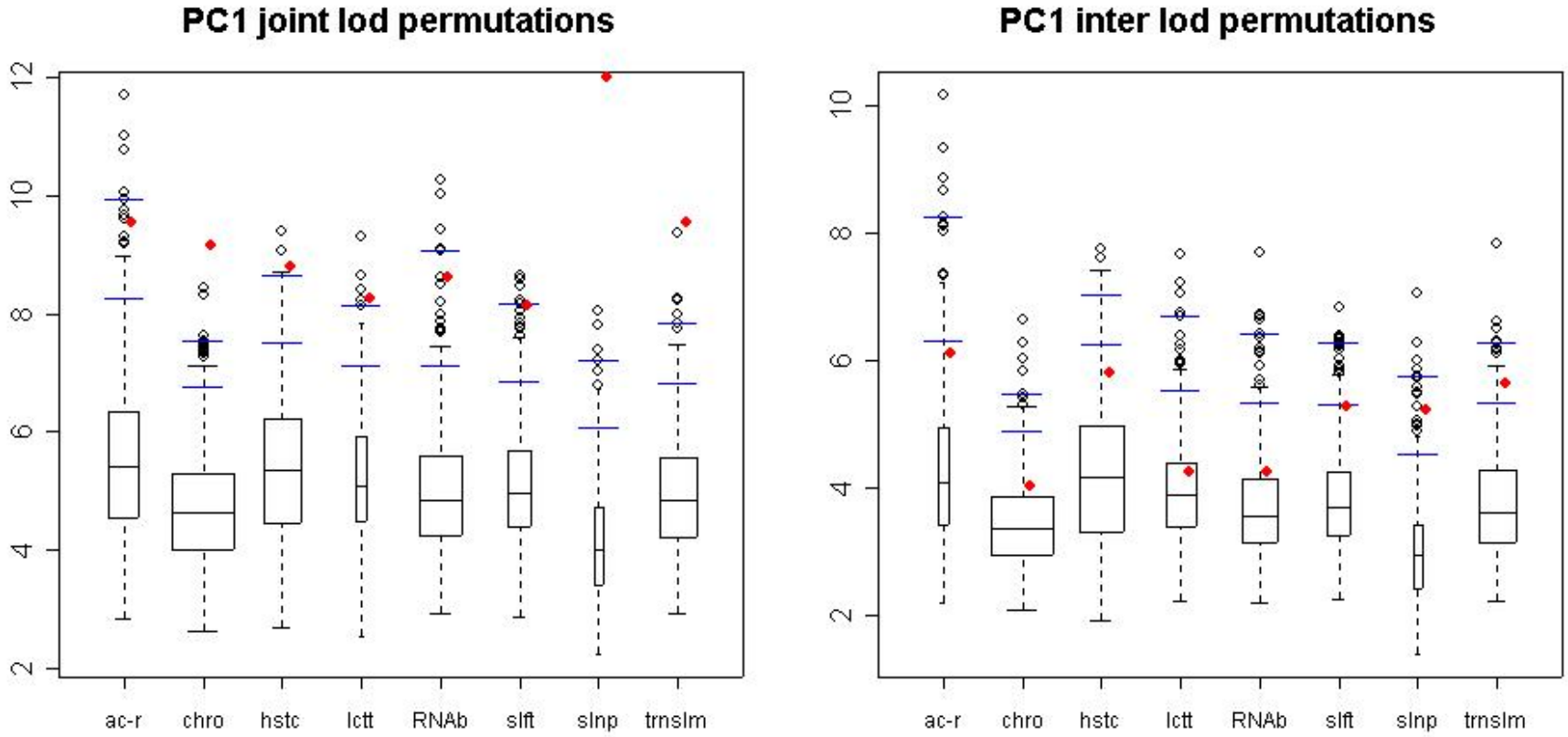
focus on translation machinery (EIF)




translation machinary: ettf1 (LOD 9.48)
translation machinary: etif2s2((LOD 9.08)
translation machinary: efRpl2 (LOD 7.11)
translation machinary: etif5 (LOD 7.53)
translation machinary: etif3s1((LOD 8.53)
translation machinary: etif3s6 (LOD 8.74)
translation machinary: etif4g2 (LOD 8.17)
translation machinary: etif4A1 (LOD 5.16)
translation machinary: etif4A2 (LOD 4.6)
translation machinary: etif2s3sgX (LOD 8.99)
translation machinary: etef1d(nep (LOD 8.23)
Chromosome
15
4

how well does PC1 do?

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

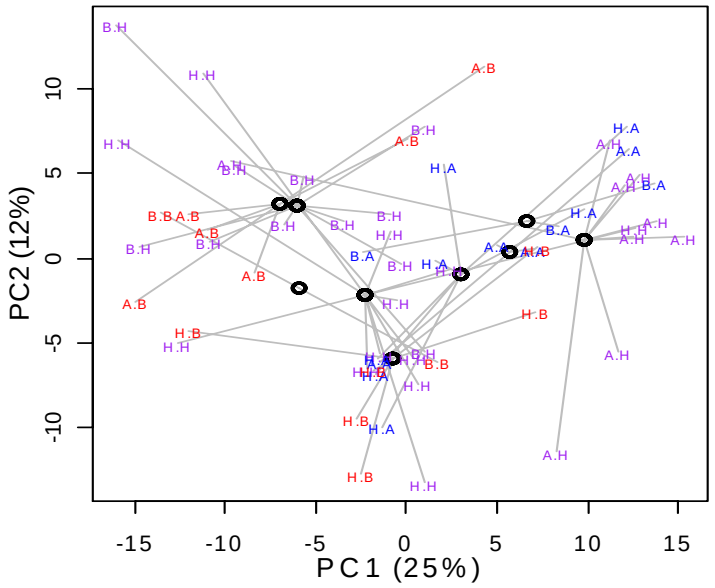


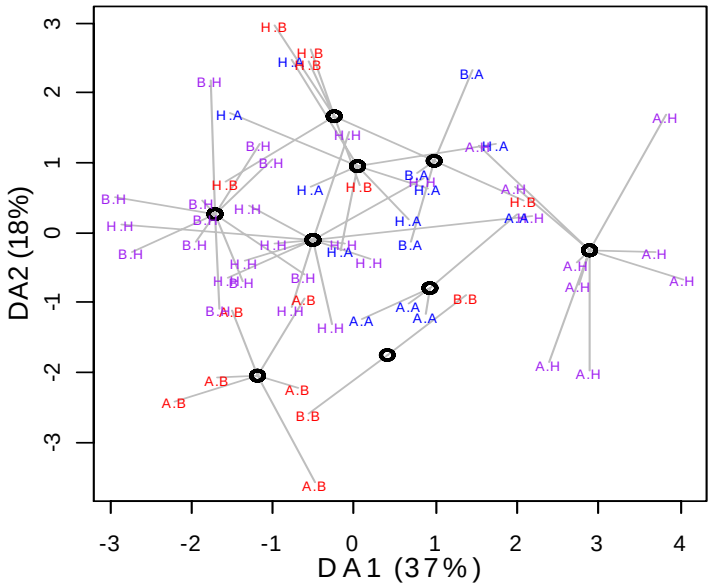
blue bars at 1%, 5%; width proportional to group size

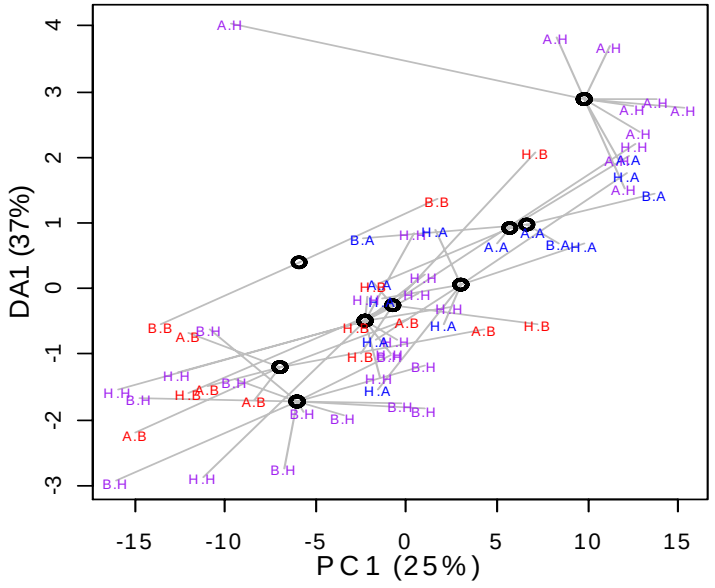
PC and DA for 1500+ mRNA traits

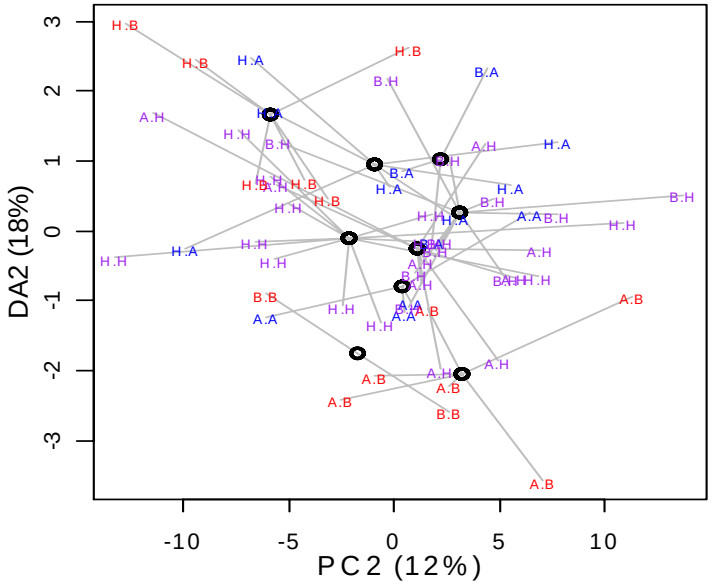
PC shows little relation to genotypes

DA based on best fit with marker pair D4Mit17 and D15Mit63

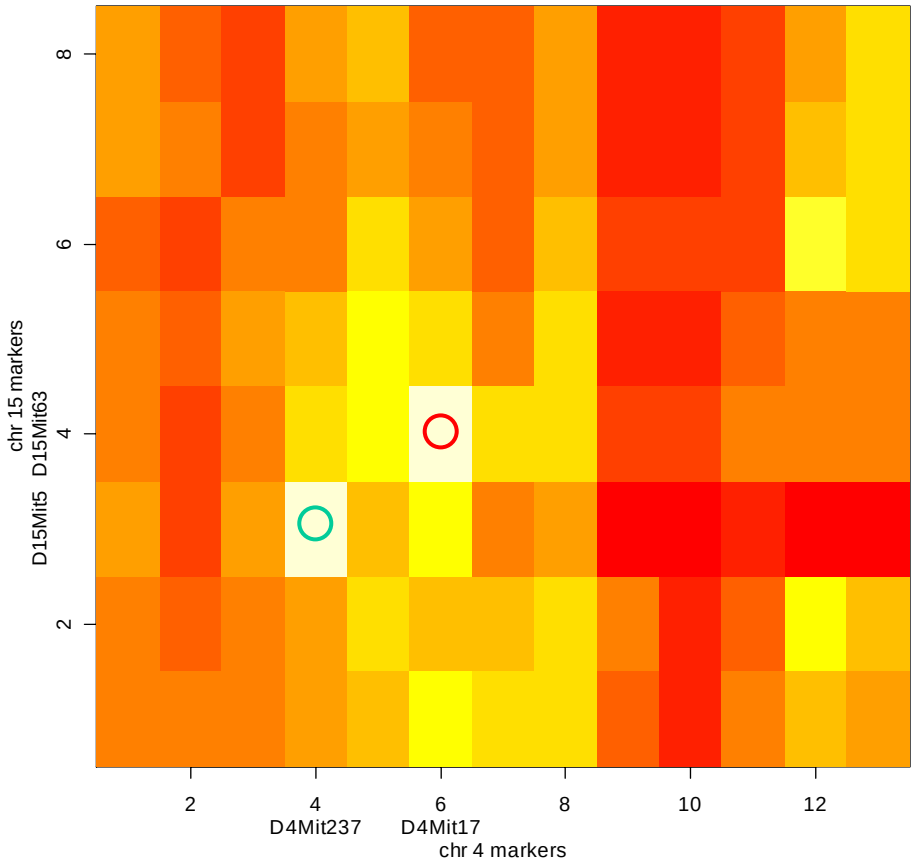


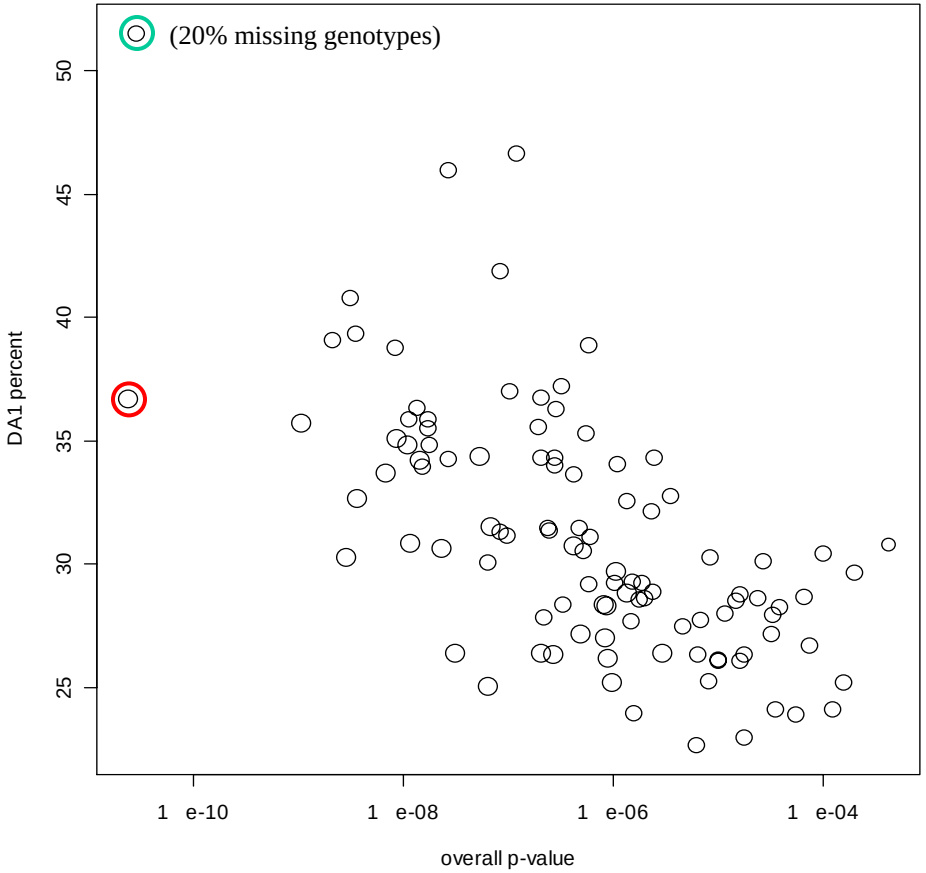




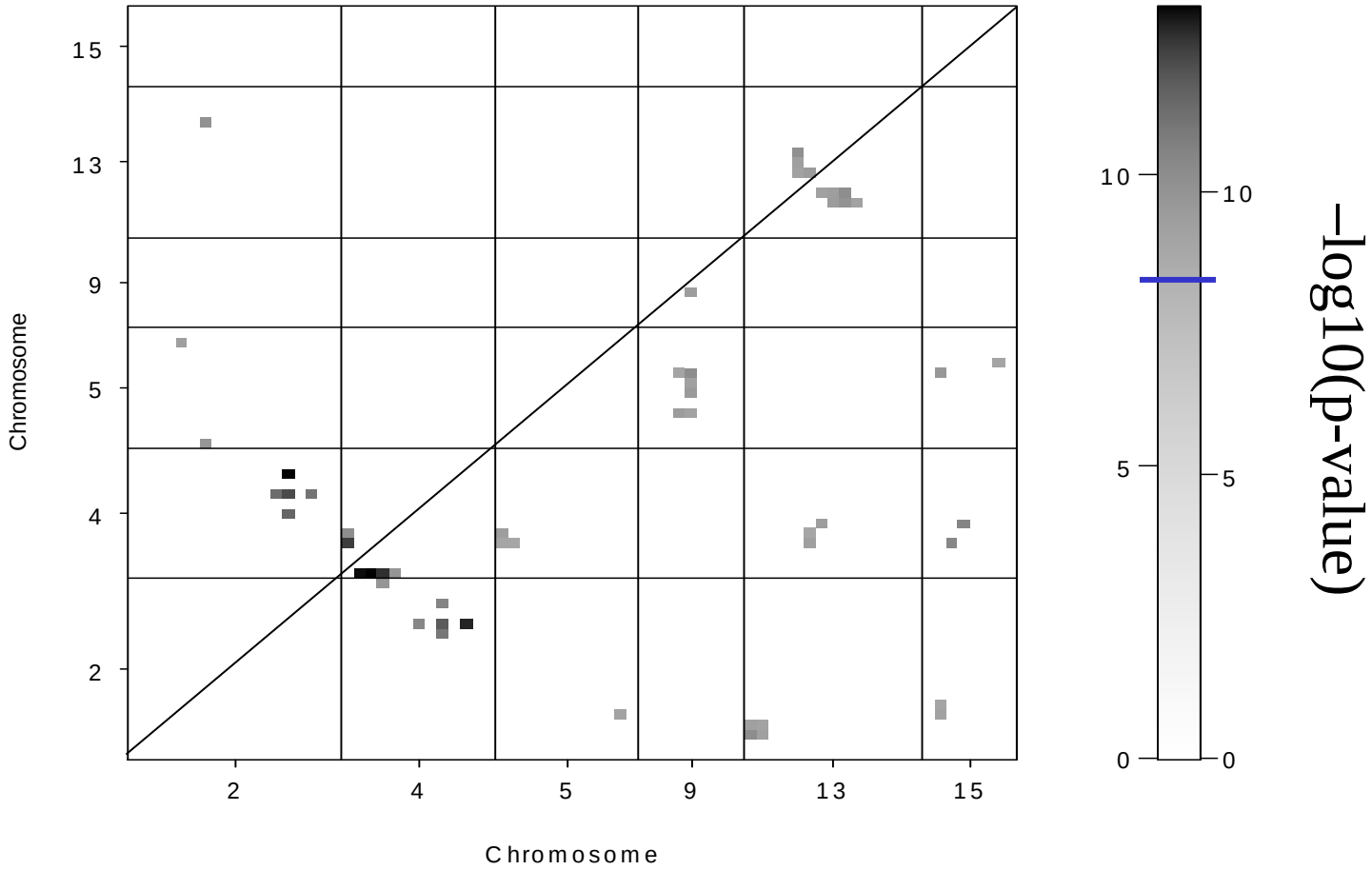


2-marker regression for DA1
on chr 4 & 15 across 1500+ mRNA traits





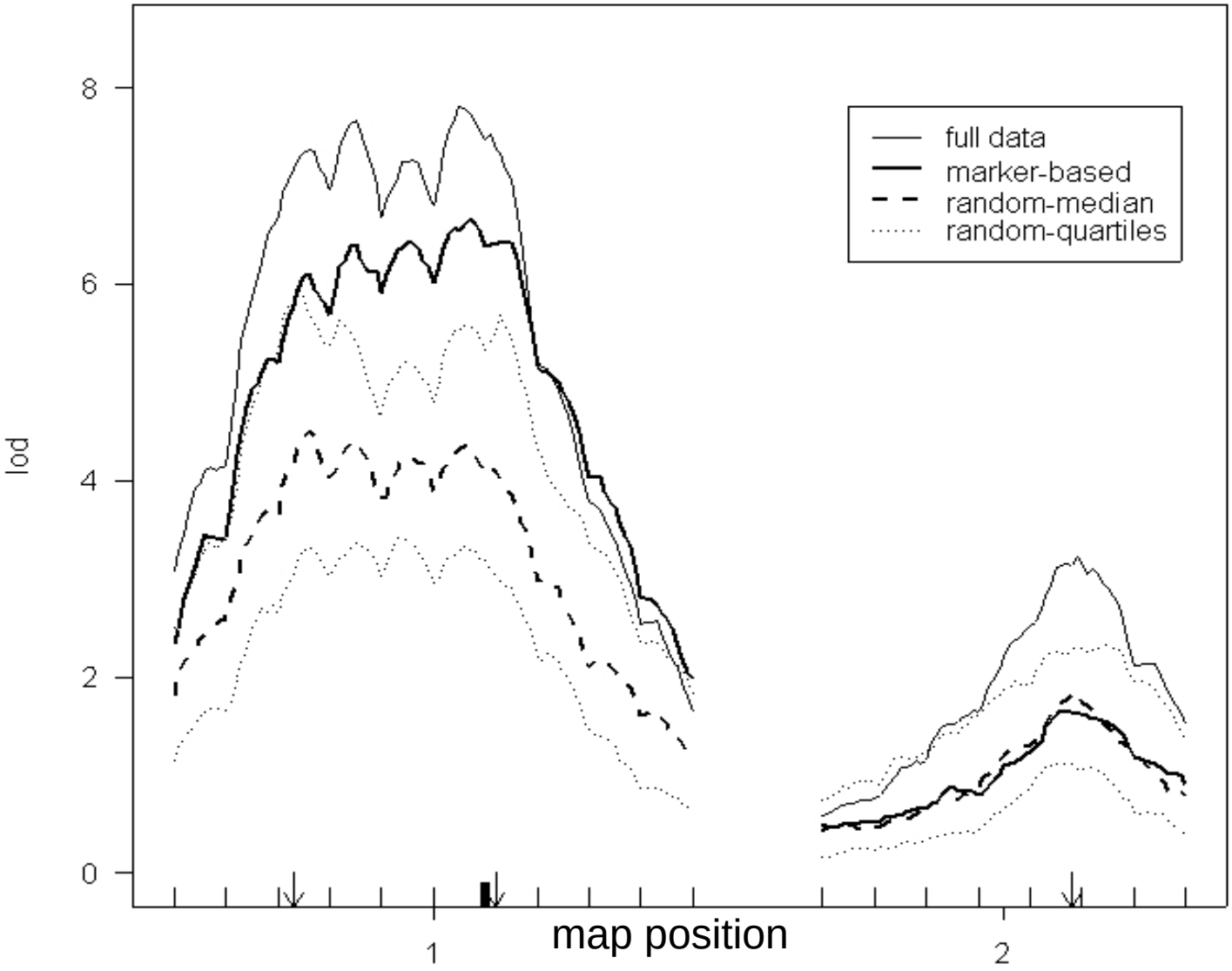
DA for selected chromosomes (mask pairs below p-value = 10^{-8})



4. designing for expensive phenotypes (Jin et al. 2004)

- microarray analysis ~ $1000 per mouse
  - could only afford to assay 60 of 108 in panel
  - wanted to preserve power to detect QTL
- selective phenotyping
  - identify set of key markers
    - framework map across subset of genome
    - or key regions identified in previous studies
      - chr 2,4,5,9,16,19 for physiological traits in diabetes/obesity study
  - genotype all individuals in panel at markers
  - select subset for phenotyping based on genotype
  - interval map with no bias

simulated LOD profiles with 3 QTL on 2 chr



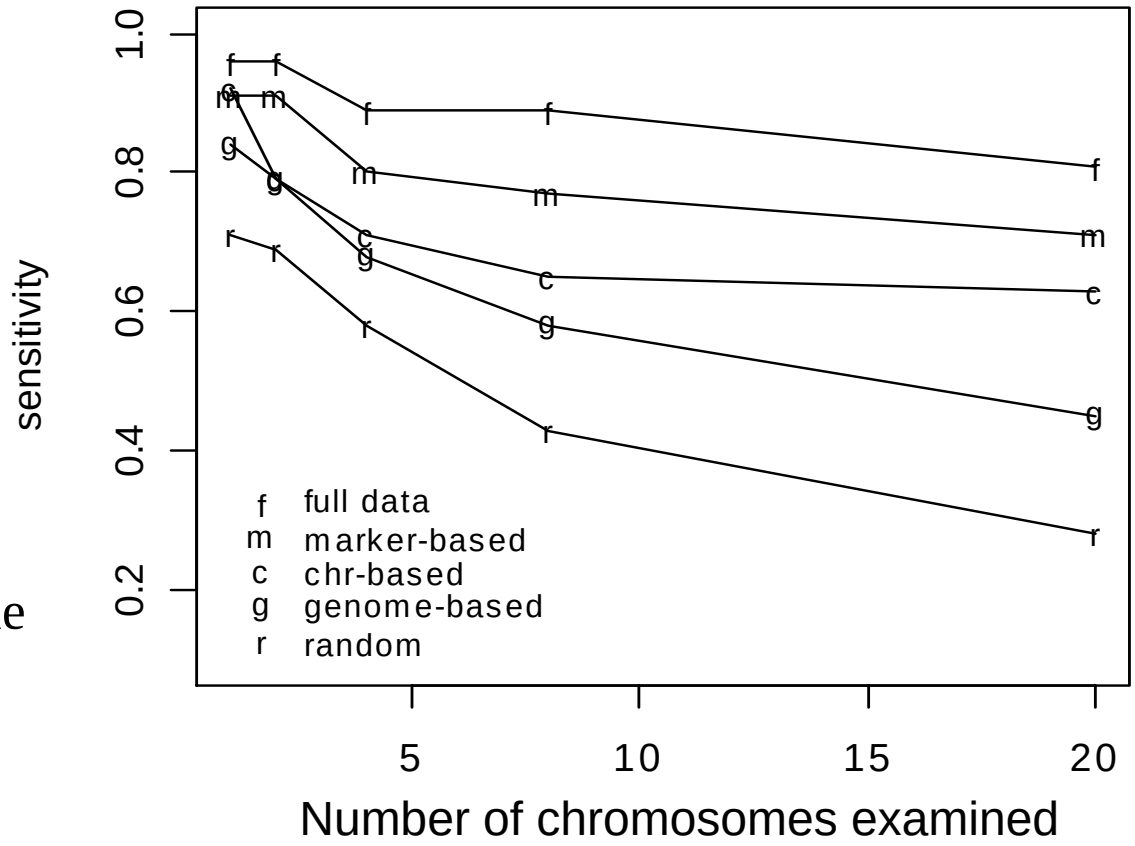
comparison of different selection methods
improved power over random sample

sensitivity = pr(detect QTL | QTL is real)

up to 80% sensitivity
of full panel

best with few markers
near QTL

genome-wide selection
better than random sample



is this relevant to large QTL studies?

- selectively phenotype 50-75% of F2 mapping panel
  - may capture most effects
    - 1:2:1 F2 allele ratio of genotypes A:H:B
    - 1:0:1 best for additive effects (50%)
    - 1:1:1 best for general effects (75%)
  - with little loss of power
  - and dramatic reduction in cost
- 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

contact information & resources

- email: byandell@wisc.edu
- web: www.stat.wisc.edu/~yandell/statgen
  - QTL & microarray resources
  - references, software, people
- R/bim freely available
  - download R from cran.r-project.org for your system (Mac, Windows, Linux)
  - Packages... Install package(s) from CRAN... qtl
  - Packages... Install package(s) from Bioconductor... bim
- thanks:
  - students: Chunfang “Amy” Jin, Fei Zou, Pat Gaffney, Jaya Satagopan, Meng Chen (UW Statistics)
  - faculty/staff: Alan Attie, Hong Lan (UW Biochemistry); Michael Newton, Christina Kendziorski, Jason Fine (UW Biostatistics); Gary Churchill, Hao Wu (Jackson Labs)
  - USDA/CSREES, NIH/NIDDK