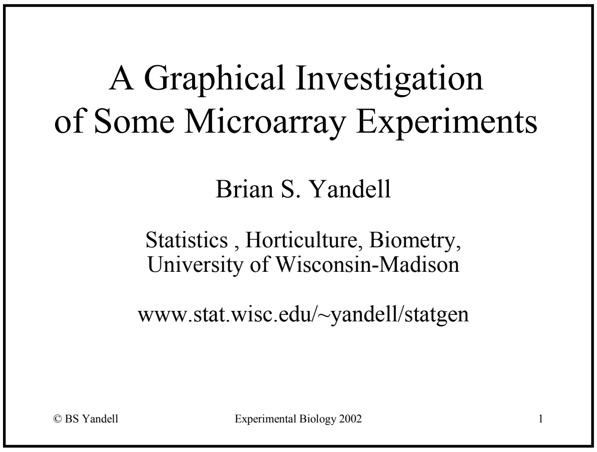
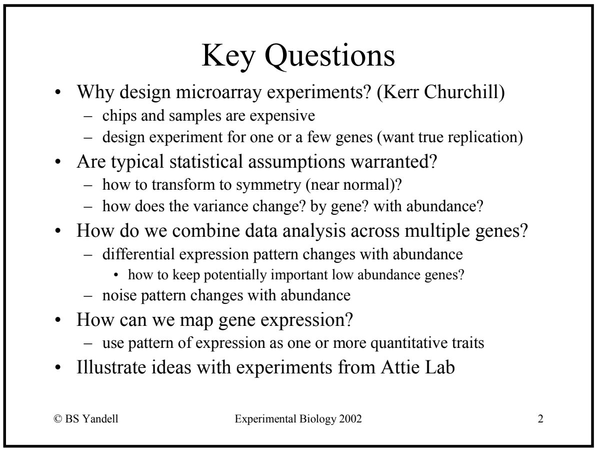


A Graphical Investigation of Some Microarray Experiments

Brian S. Yandell

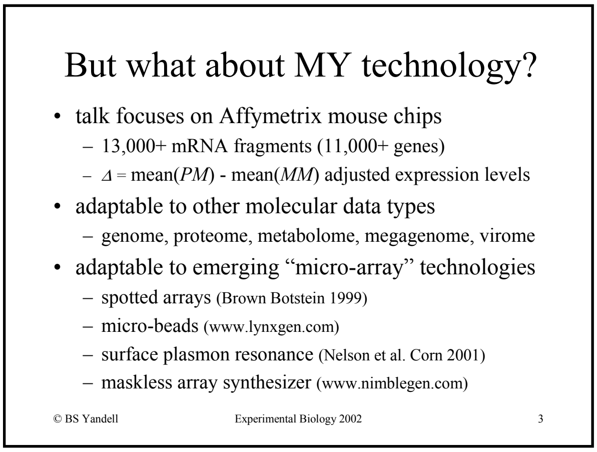
Statistics , Horticulture, Biometry,
University of Wisconsin-Madison

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



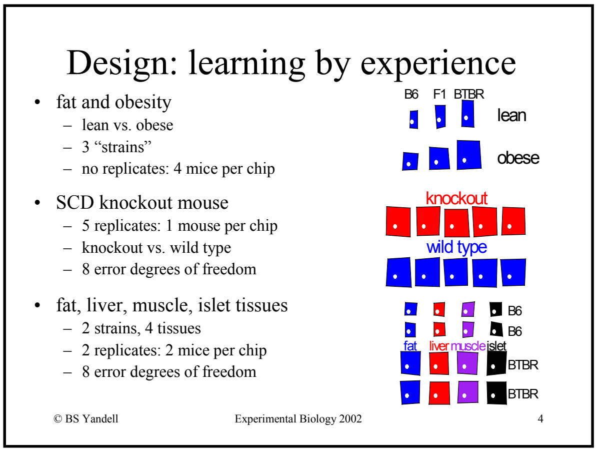
Key Questions

- Why design microarray experiments? (Kerr Churchill)
  - chips and samples are expensive
  - design experiment for one or a few genes (want true replication)
- Are typical statistical assumptions warranted?
  - how to transform to symmetry (near normal)?
  - how does the variance change? by gene? with abundance?
- How do we combine data analysis across multiple genes?
  - differential expression pattern changes with abundance
    - how to keep potentially important low abundance genes?
  - noise pattern changes with abundance
- How can we map gene expression?
  - use pattern of expression as one or more quantitative traits
- Illustrate ideas with experiments from Attie Lab



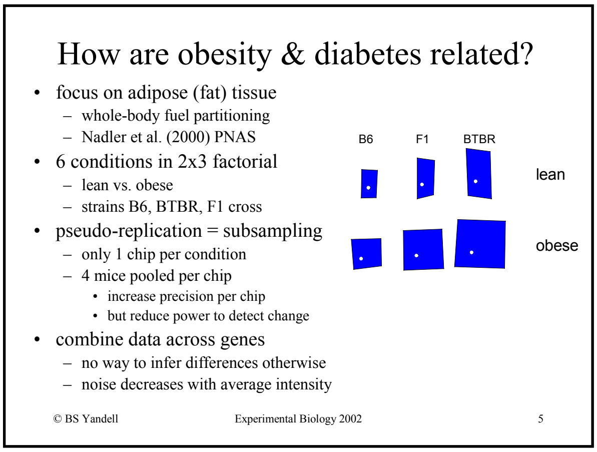
But what about MY technology?

- talk focuses on Affymetrix mouse chips
  - 13,000+ mRNA fragments (11,000+ genes)
  - Δ = mean(*PM*) - mean(*MM*) adjusted expression levels
- adaptable to other molecular data types
  - genome, proteome, metabolome, megagenome, virome
- adaptable to emerging “micro-array” technologies
  - spotted arrays (Brown Botstein 1999)
  - micro-beads (www.lynxgen.com)
  - surface plasmon resonance (Nelson et al. Corn 2001)
  - maskless array synthesizer (www.nimblegen.com)



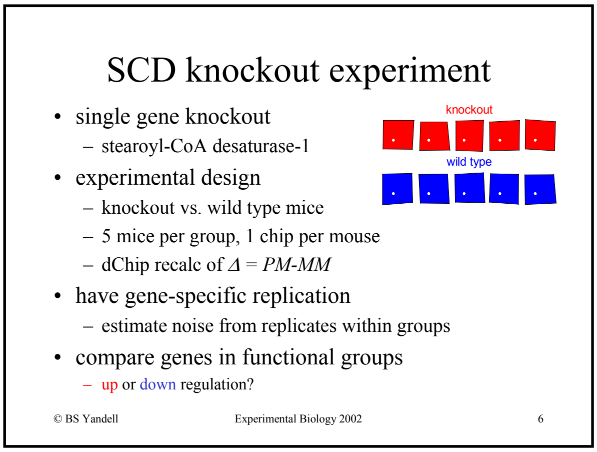
Design: learning by experience

- fat and obesity
  - lean vs. obese
  - 3 “strains”
  - no replicates: 4 mice per chip
- SCD knockout mouse
  - 5 replicates: 1 mouse per chip
  - knockout vs. wild type
  - 8 error degrees of freedom
- fat, liver, muscle, islet tissues
  - 2 strains, 4 tissues
  - 2 replicates: 2 mice per chip
  - 8 error degrees of freedom



How are obesity & diabetes related?

- focus on adipose (fat) tissue
  - whole-body fuel partitioning
  - Nadler et al. (2000) PNAS
- 6 conditions in 2x3 factorial
  - lean vs. obese
  - strains B6, BTBR, F1 cross
- pseudo-replication = subsampling
  - only 1 chip per condition
  - 4 mice pooled per chip
    - increase precision per chip
    - but reduce power to detect change
- combine data across genes
  - no way to infer differences otherwise
  - noise decreases with average intensity

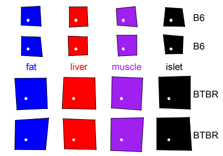


SCD knockout experiment

- single gene knockout
  - stearoyl-CoA desaturase-1
- experimental design
  - knockout vs. wild type mice
  - 5 mice per group, 1 chip per mouse
  - dChip recalc of $\Delta = PM\text{-}MM$
- have gene-specific replication
  - estimate noise from replicates within groups
- compare genes in functional groups
  - up or down regulation?

Diabetes action in whole body

- tissues important for diabetes
  - fat, muscle, liver, islets
  - focus on fat & liver here
- two obese strains
  - BTBR diabetic
  - B6 non-diabetic
- experimental design
  - only 16 Affymetrix chips
  - 2 replicates each tissue*genotype condition
  - 4 mice per condition in pools of 2 per chip
  - some benefits of pooling & independent replication



How do we infer strain differences?

strain differences?

same mean

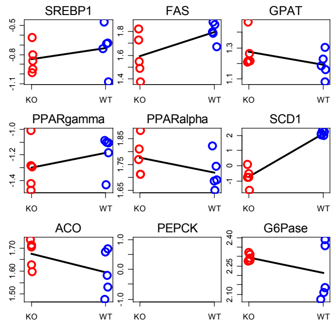
flat line

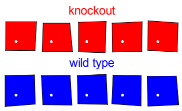
noise negligible?

high noise

signal unreliable

test inconclusive





Why is noise so important?

- is differential expression “signal” large relative to “noise”?
  - signal is difference across conditions of interest
    - lean vs. obese, knockout vs. wild, B6 vs. BTBR
  - noise assessed by “true” replicates, not pseudo-replicates
- sources of noise
  - conditions: mouse, strain, tissue
    - can vary with mRNA abundance, gene-specific features
  - materials: chip, mRNA sample preparation
    - hybridization and reading mechanics
  - watch out for pseudo-replication
    - pooled mRNA from multiple mice on one chip
    - multiple chips from same mRNA source
- experimental unit is tissue from mouse (or set of pooled mice)
  - increase power with more mice on distinct chips
  - think of experiment for a single mRNA at a time

Are typical statistical assumptions warranted for microarrays?

- independence: address at design phase
  - want chips independent, but gene spots on chip?
  - often expect genes to correlate--coordinated expression
- equal variance
  - log (almost) takes care of this--or does it?
  - what affects variance? abundance? gene function?
- normality (symmetry, bell-shaped histogram)
  - log (almost) transforms to symmetry?

To log or not to log?

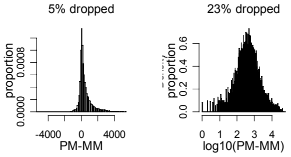
- log is natural choice
  - tremendous dynamic range (100-1000 fold common)
    - intuitive appeal, e.g. concentrations of chemicals (pH)
    - fold changes becomes additive
  - nice statistical properties ideally
    - noise variance roughly constant(?)
    - histogram roughly symmetric/normal
  - but adjusted values $\Delta = PM - MM$ may be negative
- approximate log transform: normal scores
  - there is an exact transform to normality
    - close to log(Δ) but exact form unknown: $\Phi^{-1}(F(\Delta))$
    - handles negative background-adjusted values
  - close approximation easy to compute: $X = \Phi^{-1}(F_n(\Delta))$
  - plot using anti-log to approximate fold changes

Approximate log transform: normal scores

whole chip

- $\Delta = PM - MM$
- log10(Δ)

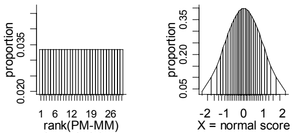
note dropped data



sample of 30

- rank(Δ)
- $X = \Phi^{-1}(F_n(\Delta))$

squish blocks into bell shaped curve



How do we analyze multiple genes?

- assume transformed expression is roughly normal
  - at least roughly symmetric
  - or use methods that account for data shape
- find common patterns of differential expression
  - compare genes across conditions
  - how can we combine gene patterns?
- use common patterns in noise
  - is variation in noise constant? probably not
    - mRNA abundance, gene function, gene-to-gene variability
  - how to model changing variation easily?
- let design drive analysis
  - linear model based on experimental design
  - incorporate sources of variation

Gene-specific model for data analysis

- fit linear model with conditions, genes, replicates
  - $X_{cgr} = \mu + C_c + G_g + D_{cg} + N_{cgr}$
    - c = condition; g = gene; r = replicate
    - $C_c = 0$ if arrays normalized separately
    - D_{cg} = differential expression for condition i, gene j
    - Kerr Churchill (2001)
- mean abundance of gene g: $A_g = X_{\bullet g \bullet}$
- differential expression: $D_g = D_{1g} - D_{2g}$
  - contrast among conditions = “signal”
  - lean vs. obese, B6 vs. BTBR, …

How to assess differential expression?

- differential expression: $D_g = \Sigma_c\, w_c X_{cg\bullet}$
  - $D_g = \Sigma\, w_i D_{cg} + \Sigma\, w_c N_{cg\bullet}$
  - $\text{Var}(D_g) = \delta_g^2 + \sigma_g^2/R$ = signal + noise
    - standardized contrasts: $\Sigma\, w_c = 0$, $\Sigma\, w_c^2 = 1$
- gene-specific variance of difference
  - $\text{Var}(D_g) = \sigma_g^2/R$ no differential expression
  - $\text{Var}(D_g) = \delta_g^2 + \sigma_g^2/R$ differential expression
- infer gene-specific differential expression
  - is signal δ_g “large” relative to noise σ_g?
  - how to estimate $\text{SD}_g = \sigma_g$?

Two ways to measure noise SD

- SD decreases with abundance
  - mechanics of hybridization, reading
  - $\text{SD}_g^2 = \sigma(A_g)^2$
    - can estimate without replication
    - combine information across genes
  - Newton et al. (2001), Roberts et al. (2000), Lin et al. (2001)
- SD varies from gene to gene
  - biochemistry of specific mRNA
  - SD_g^2 = gene-specific σ_g^2
    - need “substantial” replication (say 5?)
    - analyze genes separately
  - Efron et al. (2002), Lönnstedt Speed (2001)

Are SDs constant across genes?

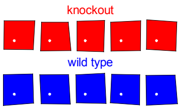
gene-specific SD

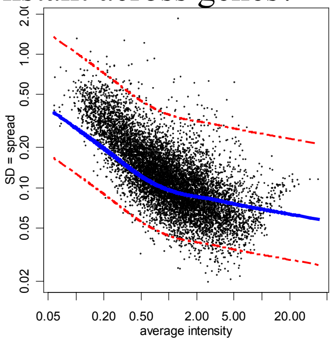
using replicates

abundance-based SD

using mean contrasts

95% χ_8^2 limits

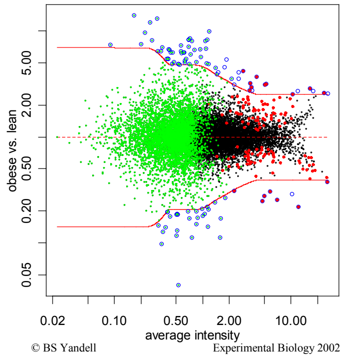




How to Estimate Spread of Noise?

- focus on genes with no differential expression
  - assume SD changes with abundance A_g
  - use robust estimate $\text{SD}_g = \sigma(A_g)$ across genes
  - screens out changing genes as “outliers”
- focus on replication
  - measure expression noise by deviations from mean
    - $\text{SD}_g^2 = \Sigma_{cr}(X_{cgr} - X_{cg\bullet})^2 / \nu_1$
- Combine ideas into gene-specific hybrid
  - Gene-specific SDs vary around $\sigma(A_g)$
    - “prior” $\sigma_g^2 \sim \text{inv-}\chi^2(\nu_0, \sigma(A_g)^2)$
  - combines two “statistically independent” estimates

Adipose: no replication



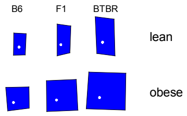
low abundance

(some $\Delta < 0$)

Clustering

(Nadler et al. 2000)

significant change

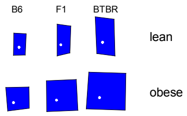


Why Worry about Low Abundance Genes?

- expression may be at or below background level
  - background adjustment: $\Delta = PM - MM$
    - removes local “geography”
    - allows comparison within and between chips
    - can be negative--problem with log transform
  - large measurement variability
    - early technology (bleeding edge)
    - do next generation chips really fix this?
  - low abundance genes
    - mRNA virtually absent in one condition
    - could be important: transcription factors, receptors, regulators
- high prevalence across genes on a chip
  - up to 25% per early Affy chips (reduced to 3-5% with www.dChip.org)
  - 10-50% across multiple conditions
- low abundance signal may be very noisy
  - 50% false positive rate even after adjusting for variance
  - may still be worth pursuing: high risk, high research return

Adipose: What was Found?

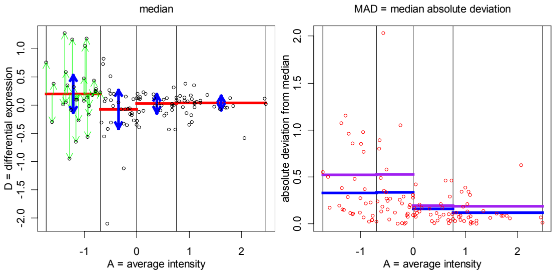
- transcription factors
  - I-κB modulates transcription - inflammatory processes
  - RXR nuclear hormone receptor - forms heterodimers with several nuclear hormone receptors
- regulatory proteins
  - protein kinase A
  - glycogen synthase kinase-3
- roughly 100 genes
  - 90 new since Nadler (2000) PNAS
  - but 50% false positives!



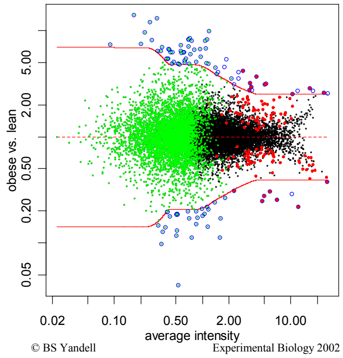
Robust SD varying with abundance

- median & median absolute deviation (MAD)
  - robust to outliers (e.g. changing genes)
  - easy to compute
  - adapt to patterns in data rather than idealized model
- partition genes into slices based on abundance A_g
  - use many slices to assess how SD varies
  - ~30 genes per slice for Affy mouse chips (400 slices)
- smooth median & MAD over slices
  - automated smoothing splines (Wahba 1990)
  - smoothes out slice-to-slice chatter

Median & MAD by abundance slices (1% sample)



Adipose: center and spread



low abundance

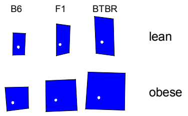
(some $\Delta < 0$)

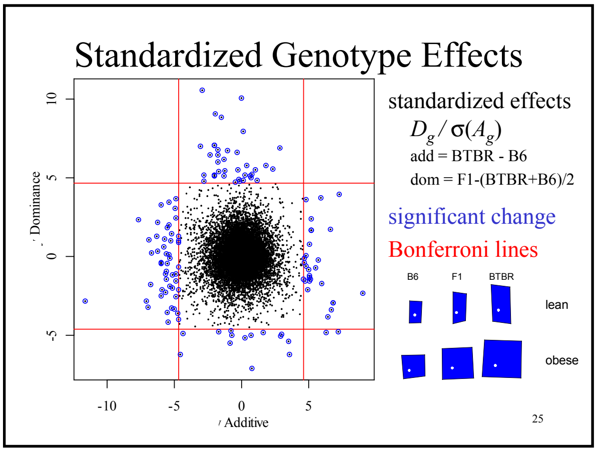
Clustering

(Nadler et al. 2000)

significant change

(Bonferroni limits)





Standardized Genotype Effects

standardized effects

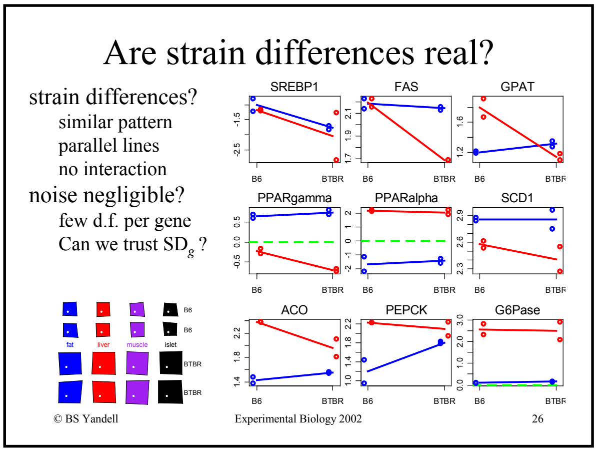
$D_g / \sigma(A_g)$

add = BTBR - B6

dom = F1-(BTBR+B6)/2

significant change

Bonferroni lines



Are strain differences real?

strain differences?

similar pattern

parallel lines

no interaction

noise negligible?

few d.f. per gene

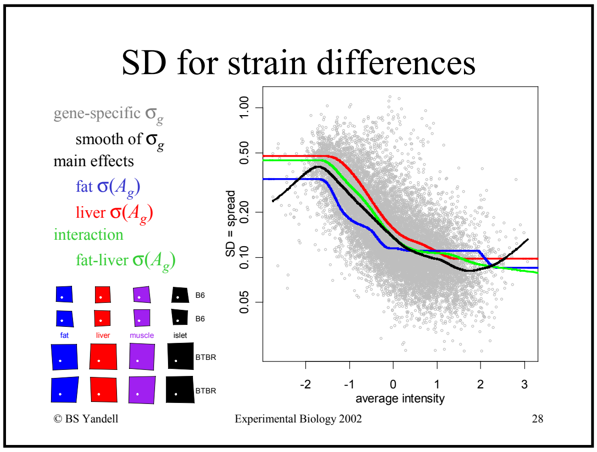
Can we trust SD_g ?

Improving on gene-specific SD

- gene-specific SD from replication
  - $SD_g^2 = \Sigma_{cr}(X_{cgr} - X_{cg\bullet})^2 / \nu_1$
- robust abundanced-based estimate
  - $\sigma(A_g)$ = smoothed MAD
- Combine ideas into gene-specific hybrid
  - “prior” $\sigma_g^2 \sim \text{inv-}\chi^2(\nu_0, \sigma(A_g)^2)$
  - “posterior” shrinkage estimate

$$\frac{\nu_1 SD_g^2 + \nu_0 \sigma(A_g)^2}{\nu_1 + \nu_0}$$

  - combines two “statistically independent” estimates



SD for strain differences

gene-specific σ_g

smooth of σ_g

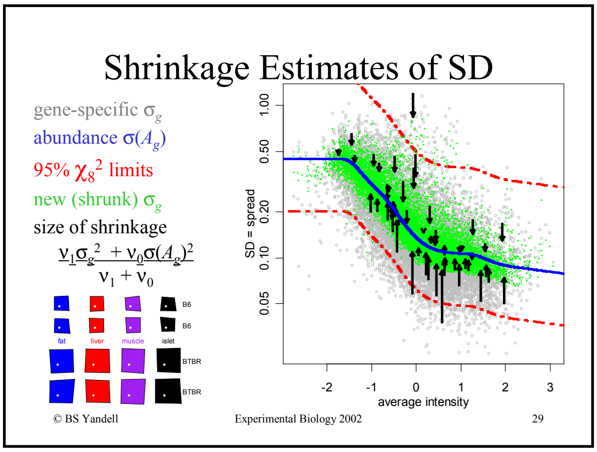
main effects

fat $\sigma(A_g)$

liver $\sigma(A_g)$

interaction

fat-liver $\sigma(A_g)$



Shrinkage Estimates of SD

gene-specific σ_g

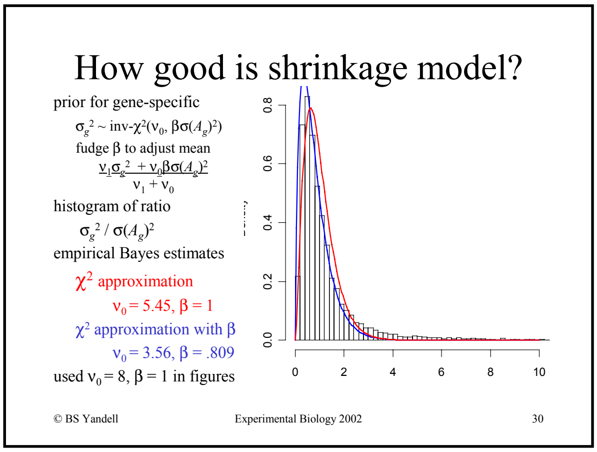
abundance $\sigma(A_g)$

95% χ_8^2 limits

new (shrunk) σ_g

size of shrinkage

$$\frac{\nu_1 \sigma_g^2 + \nu_0 \sigma(A_g)^2}{\nu_1 + \nu_0}$$



How good is shrinkage model?

prior for gene-specific

$\sigma_g^2 \sim \text{inv-}\chi^2(\nu_0, \beta\sigma(A_g)^2)$

fudge β to adjust mean

$$\frac{\nu_1 \sigma_g^2 + \nu_0 \beta\sigma(A_g)^2}{\nu_1 + \nu_0}$$

histogram of ratio

$\sigma_g^2 / \sigma(A_g)^2$

empirical Bayes estimates

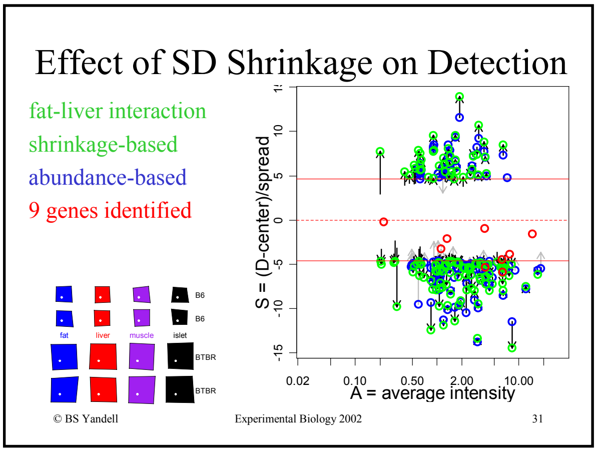
χ^2 approximation

$\nu_0 = 5.45, \beta = 1$

χ^2 approximation with β

$\nu_0 = 3.56, \beta = .809$

used $\nu_0 = 8, \beta = 1$ in figures



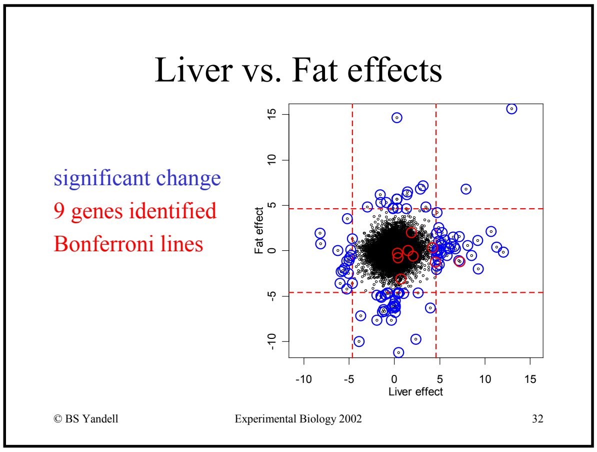
Effect of SD Shrinkage on Detection

fat-liver interaction

shrinkage-based

abundance-based

9 genes identified

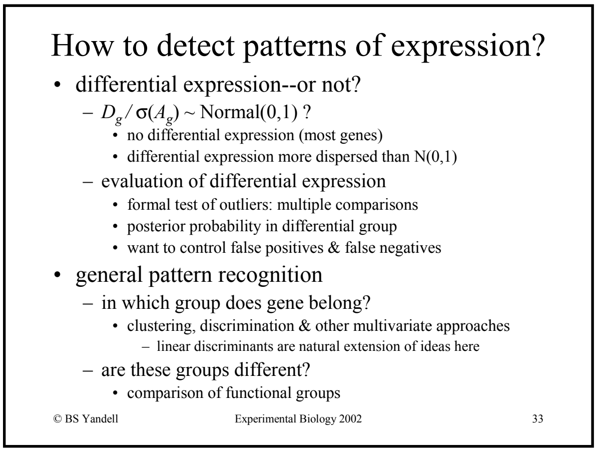


Liver vs. Fat effects

significant change

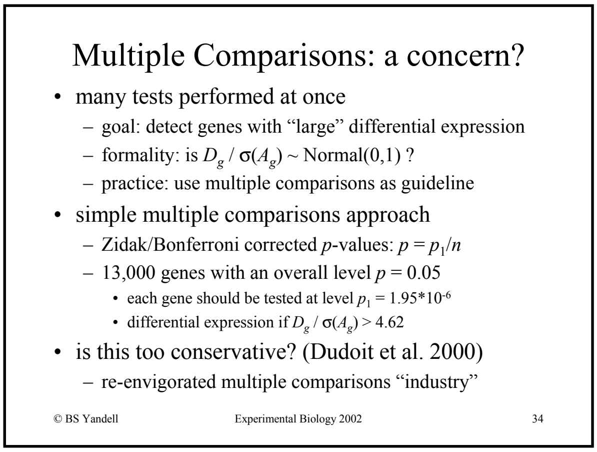
9 genes identified

Bonferroni lines



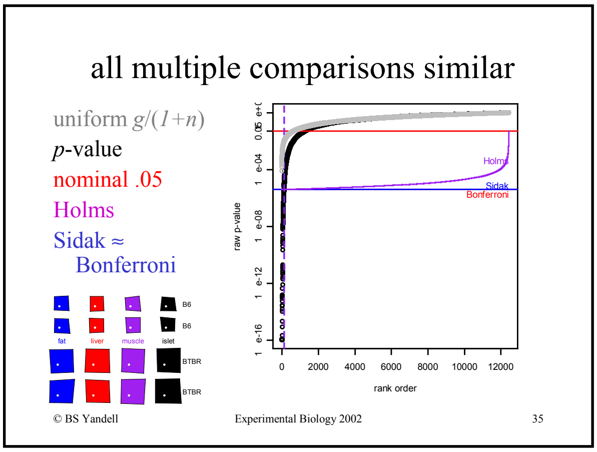
How to detect patterns of expression?

- differential expression--or not?
  - $D_g / \sigma(A_g) \sim \text{Normal}(0,1)$?
    - no differential expression (most genes)
    - differential expression more dispersed than N(0,1)
  - evaluation of differential expression
    - formal test of outliers: multiple comparisons
    - posterior probability in differential group
    - want to control false positives & false negatives
- general pattern recognition
  - in which group does gene belong?
    - clustering, discrimination & other multivariate approaches
      - linear discriminants are natural extension of ideas here
  - are these groups different?
    - comparison of functional groups



Multiple Comparisons: a concern?

- many tests performed at once
  - goal: detect genes with “large” differential expression
  - formality: is $D_g / \sigma(A_g) \sim \text{Normal}(0,1)$?
  - practice: use multiple comparisons as guideline
- simple multiple comparisons approach
  - Zidak/Bonferroni corrected p-values: $p = p_1/n$
  - 13,000 genes with an overall level $p = 0.05$
    - each gene should be tested at level $p_1 = 1.95*10^{-6}$
    - differential expression if $D_g / \sigma(A_g) > 4.62$
- is this too conservative? (Dudoit et al. 2000)
  - re-envigorated multiple comparisons “industry”



all multiple comparisons similar

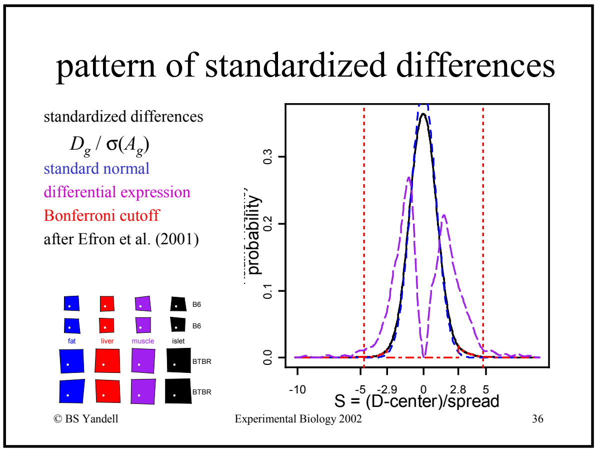
uniform $g/(1+n)$

p-value

nominal .05

Holms

Sidak ≈
Bonferroni



pattern of standardized differences

standardized differences

$D_g / \sigma(A_g)$

standard normal

differential expression

Bonferroni cutoff

after Efron et al. (2001)

Comparing gene function groups

9 functional groups

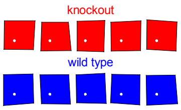
115 significant genes

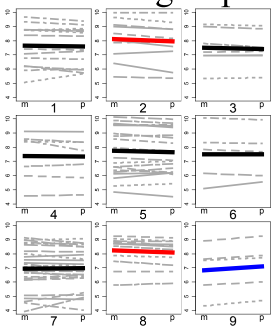
5-20 genes/group

dropped unknowns

up or down regulation?

relative to gene-to-gene variation





Related Literature

- comparing two conditions
  - log normal: var=c(mean)^2
    - ratio-based (Chen et al. 1997)
    - error model (Roberts et al. 2000; Hughes et al. 2000)
    - empirical Bayes (Efron et al. 2002; Lönnstedt Speed 2001)
      - gene-specific $D_g \sim \Phi$, var(D_g) ~Γ^{-1}, $Z_g \sim$ Bin(p)
  - gamma
    - Bayes (Newton et al. 2001, Tsodikov et al. 2000)
      - gene-specific $X_g \sim \Gamma$, $Z_g \sim$ Bin(p)
- anova (Kerr et al. 2000, Dudoit et al. 2000)
  - log normal: var=c(mean)^2
  - handles multiple conditions in anova model
  - SAS implementation (Wolfinger et al. 2001)
- See www.stat.wisc.edu/~yandell/statgen References

R Software Implementation

- quality of scientific collaboration
  - hands on experience to researcher
  - focus on graphical information content
- needs of implementation
  - quick and visual
  - easy to use (GUI=Graphical User Interface)
  - defensible to other scientists
  - open source in public domain?
- www.r-project.org
  - www.bioconductor.org

library(pickgene)

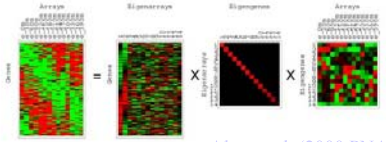
```
### R library
library(pickgene)

### create differential expression plot(s)
result <- pickgene( data, geneID = probes,
   renorm = sqrt(2), rankbased = T )

### print results for significant genes
print( result$pick[[1]] )

### density plot of standardized differences
pickedhist( result, p1 = .05, bw = NULL )
```

Mapping Gene Expression as a Quantitative Trait?

- gene expression in segregating population
  - assume one gene locus (QTL) influences expression
  - create backcross (BC) or intercross (F2)
  - map QTL using expression as quantitative trait
    - scan entire genome for possible QTL
    - MapMaker, QTL Cart or other package
  - gene expression may be controlled by other QTL
- multiple genes influenced by same QTL?
  - is QTL at a regulatory gene?
- multiple QTL affecting some regulatory gene?

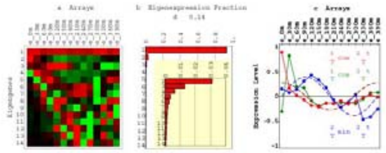
From genes to regulatory genes

- X = expression data from chips for F2 population
  - too many gene expressions to map separately
  - reduce dimension using multivariate approach
  - principle components (singular value decomposition)
    - $X = UDV^T$
    - V has eigen-genes as rows, individuals as columns
- V = combined expression of coordinated genes
  - map first few important rows of V as quantitative traits
  - suppose coordination due to gene regulation
    - elicit biochemical pathways (Henderson et al. Hoeschele 2001)
  - increase power to detect expression-modifying QTL



$X = UDV^T$

Alter et al. (2000 PNAS)
yeast cell cycle



Idea of test for QTL at one locus

(using graphs from West et al. 2000)

t-tests at QTL locus

colored correlation matrix

regulators

genes

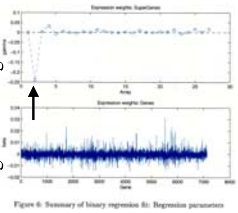
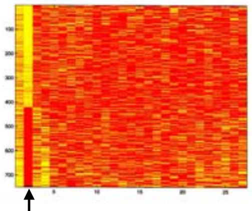


Figure 6: Summary of binary regression fit: Regression parameters

X = gene signal



V = regulator signal

Multiple QTLs

- mapping principle component as quantitative trait
  - Liu et al. (1996); Zeng et al. (2000)
  - multiple interval mapping with interactions
- research groups working on expression QTLs
  - Doerge et al. (Purdue)
  - Jansen et al. (Waginingen)
- multiple QTL literature
  - multiple interval mapping
    - Zeng, Kao, et al. (1999, 2000)
  - Bayesian interval mapping
    - Satagopan et al. (1996); Satagopan, Yandell (1996); Stevens, Fisch (1998); Silanpää, Arjas (1998, 1999)

Summary

- Why design microarray experiments? (Kerr Churchill)
  - chips and samples are expensive: use resources well
  - design experiment for one gene with true replication
- Are typical statistical assumptions warranted?
  - not automatically--plot your data!
  - find transform to symmetry (near normal)
  - examine how SD changes with abundance
- How do we combine data analysis across multiple genes?
  - keep low abundance data & allow model noise with abundance
  - use formal tests as guide to false positive rate
- How can we map gene expression?
  - use multivariate summaries to capture functional patterns
  - expression may be controlled by other (regulatory) gene
- Ongoing collaboration requires continual dialog

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

Alan D. Attie[3]	Yi Lin[1]
Hong Lan[3]	Yang Song[1]
Samuel T. Nadler[3]	Fei Zou[5]
	Christina Kendziorski[4]

[3]UW-Madison Biochemistry

[1]UW-Madison Statistics
[4]UW-Madison Biostatistics
[5]UNC Biostatistics