Mining for Low Abundance Transcripts in Microarray Data

Yi Lin[1], Samuel T. Nadler[2], Hong Lan[2], Alan D. Attie[2], Brian S. Yandell[1,3]

[1]Statistics, [2]Biochemistry, [3]Horticulture, University of Wisconsin-Madison

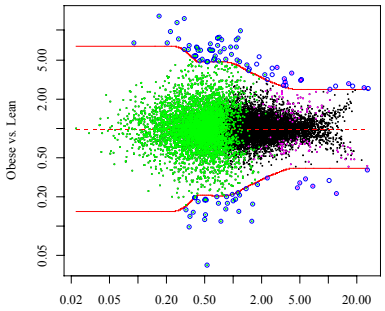
Key Issues

- differential gene expression using mRNA chips
  - diabetes and obesity study (biochemistry)
    - lean vs. obese mice: how do they differ?
    - what is the role of genetic background?
  - detecting genes at low expression levels
- inference issues
  - formal evaluation of each gene with(out) replication
    - smoothly combine information across genes
  - significance level and multiple comparisons
  - general pattern recognition: tradeoffs of false +/–
- modelling differential expression
  - gene-specific vs. dependence on abundance
  - R software module

Diabetes & Obesity Study

- 13,000+ mRNA fragments (11,000+ genes)
  - oligonuleotides, Affymetrix gene chips
  - mean(PM) - mean(NM) adjusted expression levels
- six conditions in 2x3 factorial
  - lean vs. obese
  - B6, F1, BTBR mouse genotype
- adipose tissue
  - influence whole-body fuel partitioning
  - might be aberrant in obese and/or diabetic subjects
- Nadler et al. (2000) PNAS

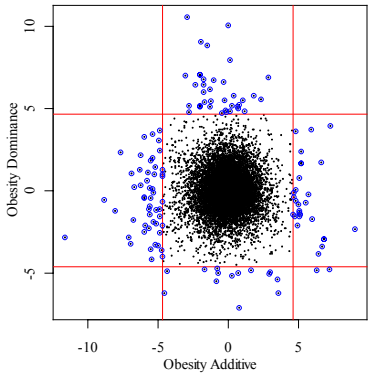
Low Abundance Genes for Obesity



Low Abundance Obesity Genes

- low mean expression on at least 1 of 6 conditions
  - negative adjusted values
  - ignored by clustering routines
- transcription factors
  - I-κB modulates transcription - inflammatory processes
  - RXR nuclear hormone receptor - forms heterodimers with several nuclear hormone receptors
- regulation proteins
  - protein kinase A
  - glycogen synthase kinase-3
- roughly 100 genes
  - 90 new since Nadler (2000) PNAS

Obesity Genotype Main Effects



Low Abundance on Microarrays

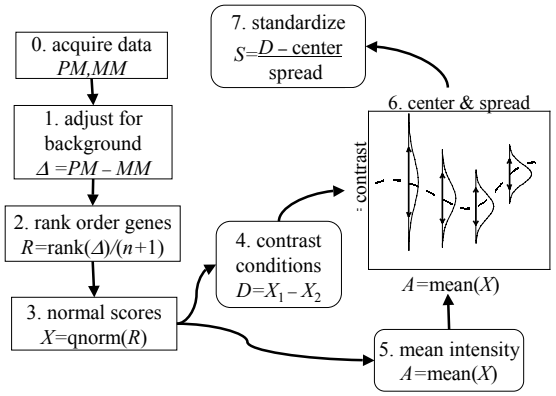
- background adjustment
  - remove local "geography"
  - comparing within and between chips
- negative values after adjustment
  - low abundance genes
    - virtually absent in one condition
    - could be important: transcription factors, receptors
  - large measurement variability
    - early technology (bleeding edge)
- prevalence across genes on a chip
  - up to 25% per chip (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

Why not use log transform?

- log is natural choice
  - tremendous scale range (100-1000 fold common)
  - intuitive appeal, e.g. concentrations of chemicals (pH)
  - looks pretty good in practice (roughly normal)
  - easy to test if no difference across conditions
  - but adjusted values $\Delta = PM - MM$ may be negative
- approximate transform to normal
  - very close to log if that is appropriate
  - handles negative background-adjusted values
  - approximate $\Phi^{-1}(F(\Delta))$ by $\Phi^{-1}(F_n(\Delta))$

Normal Scores Procedure

adjusted expression	$\Delta = PM - MM$
rank order	$R = \text{rank}(\Delta) / (n+1)$
normal scores	$X = \text{qnorm}(R)$
	$X = \Phi^{-1}(F_n(\Delta))$
average intensity	$A = (X_1+X_2)/2$
difference	$D = X_1 - X_2$
variance	$\text{Var}(D \mid A) \approx \sigma^2(A)$
standardization	$S = [D - \mu(A)]/\sigma(A)$



Robust Center & Spread

- center and spread vary with mean expression X
- partitioned into many (about 400) slices
  - genes sorted based on X
  - containing roughly the same number of genes
- slices summarized by median and MAD
  - median = center of data
  - MAD = median absolute deviation
  - robust to outliers (e.g. changing genes)
- smooth median & MAD over slices

Robust Spread Details

- MAD ~ same distribution across A up to scale
  - $\text{MAD}_i = \sigma_i S_i,\ S_i \sim S,\ i = 1,\ldots,400$
  - $\log(\text{MAD}_i) = \log(\sigma_i) + \log(S_i),\ I = 1,\ldots,400$
- regress $\log(\text{MAD}_i)$ on A_i with smoothing splines
  - smoothing parameter tuned automatically
    - generalized cross validation (Wahba 1990)
- globally rescale anti-log of smooth curve
  - $\text{Var}(D|A) \approx \sigma^2(A)$
- can force $\sigma^2(A)$ to be decreasing

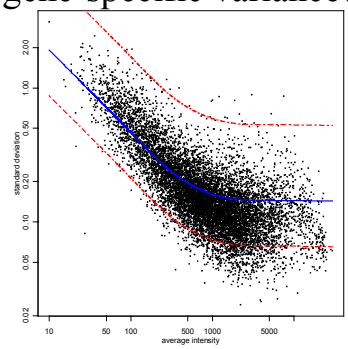
Anova Model

- transform to normal: $X = \Phi^{-1}(F_n(\Delta))$
- $X_{ijk} = \mu + C_i + G_j + (CG)_{ij} + E_{ijk}$
  - i=1,…,I conditions; j=1,…,J genes; k=1,…,K replicates
  - $C_i = 0$ if arrays normalized separately
- Z_i = 1(0) if (no) differential expression
- Variance ($A_j = \Sigma_{jk}\ X_{ijk}/IK$)
  - $\mathrm{Var}(X_{ijk} \mid A_j) = \gamma(A_j)^2 + \delta(A_j)^2 + \sigma(A_j)^2$ if $Z_i = 1$
  - $\mathrm{Var}(X_{ijk} \mid A_j) = \gamma(A_j)^2 + \sigma(A_j)^2$ if $Z_i = 0$

Differential Expression

- $D_{jk} = \Sigma\ w_i X_{ijk}$ with $\Sigma\ w_i = 0$, $\Sigma\ w_i^2 = 1$
  - $D_{jk} = \Sigma\ w_i (CG)_{ij} + \Sigma\ w_i E_{ijk}$
- Variance depending on abundance
  - $\mathrm{Var}(D_{jk} \mid A_j) = \delta(A_j)^2 + \sigma(A_j)^2$ if $Z_i = 1$
  - $\mathrm{Var}(D_{jk} \mid A_j) = \sigma(A_j)^2$ if $Z_i = 0$
- Variance depending on gene j ?
  - $\mathrm{Var}(D_{jk} \mid j, A_j) = \sigma(A_j)^2 V_j$, with $V_j \sim \Gamma^{-1}(\alpha,\nu)$
  - gene-specific variance
  - gene function-specific variance

gene-specific variance?

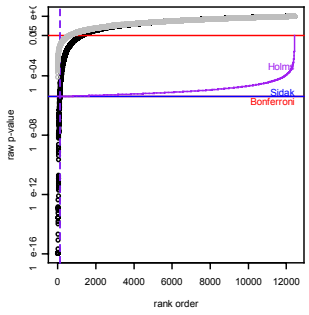


Bonferroni-corrected p-values

- standardized differences
  - $S_j = [D_j - \mu(A_j)]/\sigma(A_j) \sim$ Normal(0,1) ?
  - genes with differential expression more dispersed
- Zidak version of Bonferroni correction
  - $p = 1 - (1 - p_1)^n$
  - 13,000 genes with an overall level $p = 0.05$
    - each gene should be tested at level $1.95*10^{-6}$
    - differential expression if $S > 4.62$
  - differential expression if $|D_j - \mu(A_j)| > 4.62\sigma(A_j)$
- too conservative? weight by A_j?
  - Dudoit et al. (2000)

comparison of multiple comparisons

uniform $j/(1+n)$	grey
p-value	black
nominal .05	red
Holms	purple
Sidak	blue
Bonferroni	

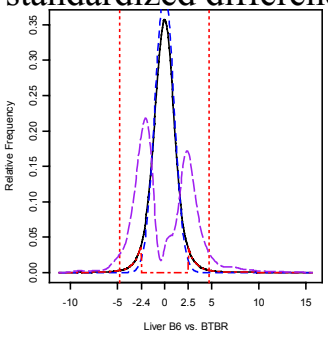


Patterns of Differential Expresssion

- (no) differential expression: $Z = (0)1$
  - $S_j | Z_j \sim$ density f_Z
    - f_0 = standard normal
    - f_1 = wider spread, possibly bimodal
  - $S_j \sim$ density $f = (1 - \pi_1) f_0 + (1 - \pi_1) f_1$
- chance of differential expression: π_1
  - $\mathrm{prob}(Z_j = 1) = \pi_1$
  - $\mathrm{prob}(Z_j = 1 \mid S_j) = \pi_1 f_1(Z_j) / f(Z_j)$

density of standardized differences

- $S = [D - \mu(A)]/\sigma(A)$
  - f = black line
- standard normal
  - f_0 = blue dash
- differential expression
  - f_1 = purple dash
- Bonferroni cutoff
  - vertical red dot



Looking for Expression Patterns

- differential expression: $D = X_1 - X_2$
  - $S = [D - \text{center}]/\text{spread} \sim \text{Normal}(0,1)$?
  - classify genes in one of two groups:
    - no differential expression (most genes)
    - differential expression more dispersed than N(0,1)
  - formal test of outlier?
    - multiple comparisons issues
  - posterior probability in differential group?
    - Bayesian or classical approach
- general pattern recognition
  - clustering / discrimination
  - linear discriminants (Fisher) vs. fancier methods

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 $D_j \sim \Phi$, var(D_j) ~Γ^{-1}, $Z_j \sim$ Bin(p)
  - gamma
    - Bayes (Newton et al. 2001, Tsodikov et al. 2000)
      - gene-specific $X_j \sim \Gamma$, $Z_j \sim$ 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)

R Software Implementation

- quality of scientific collaboration
  - hands on experience of researcher
  - save time of stats consultant
  - raise level of discussion
  - focus on graphical information content
- needs of implementation
  - quick and visual
  - easy to use (GUI=Graphical User Interface)
  - defensible to other scientists
  - public domain or affordable?
- www.r-project.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 )
```