Mining for Low Abundance Transcripts in Microarray Data

Yi Lin¹, Samuel T. Nadler², Hong Lan², Alan D. Attie², Brian S. Yandell¹,³

¹Statistics, ²Biochemistry, ³Horticulture, 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)
  - oligonucleotides, 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 mean expression on at least 1 of 6 conditions
  - negative adjusted values
  - ignored by clustering routines
- transcription factors
  - I-kB 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
  - 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

1. adjust for background
   $\Delta = PM - MM$
2. rank order genes
   $R = rank(\Delta) / (n+1)$
3. normal scores
   $X = qnorm(R)$
   $X = \Phi^{-1}(F_n(\Delta))$
4. contrast conditions
   $D = X_1 - X_2$
5. mean intensity
   $A = mean(X)$
6. center & spread
    $D = center(A)$
7. standardize
    $S = D - center(A)$

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
  - MAD$_i = \sigma S_i$, $S_i = S$, $i = 1, ..., 400$
  - log(MAD$_i$) = log($\sigma$) + log($S_i$), $I = 1, ..., 400$
- regress log(MAD$_i$) on $A_i$ with smoothing splines
  - smoothing parameter tuned automatically
    - generalized cross validation (Wahba 1990)
  - globally rescale anti-log of smooth curve
    - $\text{Var}(D_i|A) = \sigma^2(\cdot)$
    - can force $\sigma^2(\cdot)$ to be decreasing
Anova Model
• transform to normal: $X = \Phi^{-1}(F_n(D))$
• $X_{ijk} = \mu + C_i + G_j + (CG)_{ij} + E_{jjk}$
  - $i=1,...,I$ conditions; $j=1,...,J$ genes; $k=1,...,K$ replicates
  - $C_j = 0$ if arrays normalized separately
• $Z_i = 1(0)$ if (no) differential expression
• Variance $(A_j = \sum_{jk} X_{ijk}/IK)$
  - $\text{Var}(X_{ijk} | A_j) = \gamma(A_j)^2 + \delta(A_j)^2 + \sigma(A_j)^2$ if $Z_i = 1$
  - $\text{Var}(X_{ijk} | A_j) = \gamma(A_j)^2 + \sigma(A_j)^2$ if $Z_i = 0$

Differential Expression
• $D_k = \sum w_j X_{ijk}$ with $\sum w_j = 0$, $\sum w_j^2 = 1$
• $D_k = \sum w_j (CG)_{jk} + \sum w_j E_{jk}$
• Variance depending on abundance
  - $\text{Var}(D_k | A_j) = \delta(A_j)^2 + \sigma(A_j)^2$ if $Z_i = 1$
  - $\text{Var}(D_k | A_j) = \sigma(A_j)^2$ if $Z_i = 0$
• Variance depending on gene $j$
  - $\text{Var}(D_k | j, A_j) = \sigma(A_j)^2 V_j$, with $V_j \sim \Gamma^{-1}(\alpha, \nu)$
  - gene-specific variance
  - gene function-specific variance

Bonferroni-corrected $p$-values
• standardized differences
  - $S_j = \left( D_j - \mu(A_j) \right) / \sigma(A_j) \sim \text{Normal}(0,1)$?
  - genes with differential expression more dispersed
• Zidak version of Bonferroni correction
  - $p = 1 - (1 - p)^n$
  - 13,000 genes with an overall level $p = 0.05$
    - each gene should be tested at level $1.95 \times 10^{-6}$
    - 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)

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

Gene-specific variance?

Comparison of multiple comparisons
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 $\text{Normal}(0, 1)$
  - 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: $\text{var} = c(\text{mean})^2$
    - 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 $D_j - \Phi$, $\text{var}(D_j) - 1^\ast_i(X_j - \text{Bind})$
    - gamma
      - Bayes (Newton et al. 2001, Tsodikov et al. 2000)
        - gene-specific $X_j - T$, $Z_j - \text{Bind}$
  - anova (Kerr et al. 2000, Dudoit et al. 2000)
    - log normal: $\text{var} = c(\text{mean})^2$
    - 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

```r
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 )
```