Smooth Collaboration in Statistical Genomics

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

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

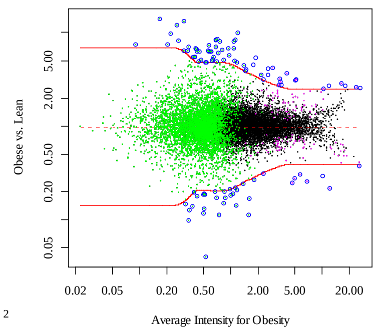
Key Issues

- what are we doing?
  - lean vs. obese mice: how do they differ?
    - gene expression using mRNA chips
  - formal evaluation of each gene without replication
    - smoothly combine information across genes
- to test or not to test?
  - significance level and multiple comparisons
  - general pattern recognition: tradeoffs of false +/–
- show me how to do it myself!
  - concepts: smooth center and spread
  - training: R software implementation

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

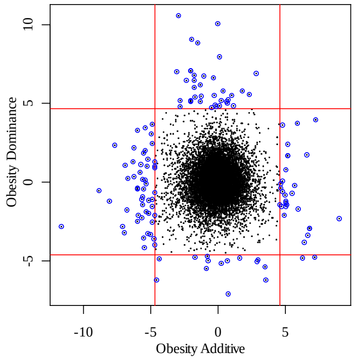
Low Abundance Genes for Obesity



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

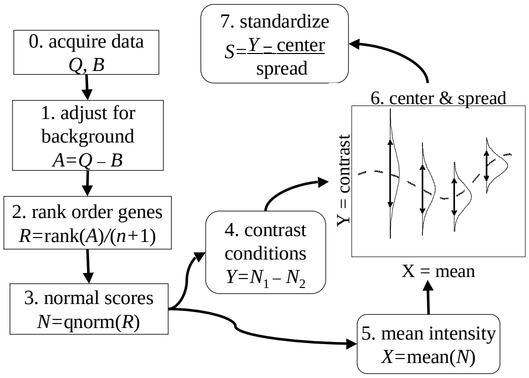
- background adjustment
  - remove local “geography”
  - comparing within and between chips
- negative values after adjustment
  - low abundance genes
    - virtually absent in one condition
    - could be important: transcription factors, receptors
  - large measurement variability
    - early technology (bleeding edge)
- prevalence across genes on a chip
  - 0-20% per chip
  - 10-50% across multiple conditions

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
- approximate transform to normal
  - normal scores of ranks (Li et al. 2000)
  - very close to log if that is appropriate
  - handles negative background-adjusted values

Normal Scores Procedure

adjusted expression	$A = Q - B$
rank order	$R = \text{rank}(A) / (n+1)$
normal scores	$N = \text{qnorm}(R)$
average intensity	$X = (N_1+N_2)/2$
difference	$Y = N_1 - N_2$
variance	$\text{Var}(Y \mid X) \approx \sigma^2(X)$
standardization	$S = [Y - \mu(X)]/\sigma(X)$



0. acquire data Q, B

1. adjust for background $A=Q-B$

2. rank order genes $R=\text{rank}(A)/(n+1)$

3. normal scores $N=\text{qnorm}(R)$

4. contrast conditions $Y=N_1-N_2$

5. mean intensity $X=\text{mean}(N)$

6. center & spread

7. standardize $S=\dfrac{Y-\text{center}}{\text{spread}}$

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 X up to scale
  - $\text{MAD}_i = \sigma_i Z_i$, $Z_i \sim Z$, $i = 1,\ldots,400$
  - $\log(\text{MAD}_i) = \log(\sigma_i) + \log(Z_i)$, $I = 1,\ldots,400$
- regress $\log(\text{MAD}_i)$ on X_i with smoothing splines
  - smoothing parameter tuned automatically
    - generalized cross validation (Wahba 1990)
- globally rescale anti-log of smooth curve
  - $\text{Var}(Y|X) \approx \sigma^2(X)$
- can force $\sigma^2(X)$ to be decreasing

Bonferroni-corrected *p*-values

- standardized normal scores
  - $S = [Y - \mu(X)]/\sigma(X) \sim \text{Normal}(0,1)$?
  - 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 $|Y - \mu(X)|\ > 4.62\sigma(X)$
- too conservative? weight by *X*?
  - Dudoit (2000)

Looking for Expression Patterns

- differential expression: $Y = N_1 - N_2$
  - *Score* = [*Y* – center]/spread ~ 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

Comparing Conditions

- comparing two conditions
  - ratio-based decisions (Chen et al. 1997)
    - constant variance of ratio on log scale, use normality
  - Bayesian inference (Newton et al. 2000, Tsodikov et al. 2000)
    - Gamma-Gamma model
    - variance proportional to squared intensity
  - error model (Roberts et al. 2000, Hughes et al. 2000)
    - variance proportional to squared intensity
    - transform to log scale, use normality
- anova (Kerr et al. 2000, Dudoit et al. 2000)
  - handles multiple conditions in anova model
  - constant variance on log scale, use normality

Publish or Perish

- academic vs. industry
- what is our audience?
  - biologists wanting to use proper methods
  - statististicians wanting to develop new methods
- who writes what? who understands what?
  - all authors responsible for content
  - mutual comprehension for the long term
- one paper or an ongoing collaboration?

Software Implementation is Key

- 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?

R Statistical System

- public domain, graphics-friendly system
  - developed maintained by top-flight statisticians
  - has standard and modern statistical methods
  - easy to install, easy-to-use graphics
  - command-line use: no GUI menus (yet)
  - extensible, scalable
- much activity with R and microarrays
  - Harvard group: Li Wong, Gentleman et al.
  - Berkeley group: Speed et al.
  - Jackson Labs: Churchill, Kerr et al.
  - Madison group: library(microarray)
    - implements Li et al. (2001); Newton et al. (2001)