QTL Studies with Microarray Data

Long Han[1], Yi Lin[2],
Fei Zou[2], Patrick J. Gaffney[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
November 2000

Basic Idea

- study QTLs across segregating population
  - simultaneous search for multiple QTLs
- phenotype is pattern of microarray expression
  - examine many facets of biological process
  - multiple traits using principle components
- account for low abundance and signal variability
  - detect transcription factors and receptors
  - robustly adapt to variability given mean expression

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

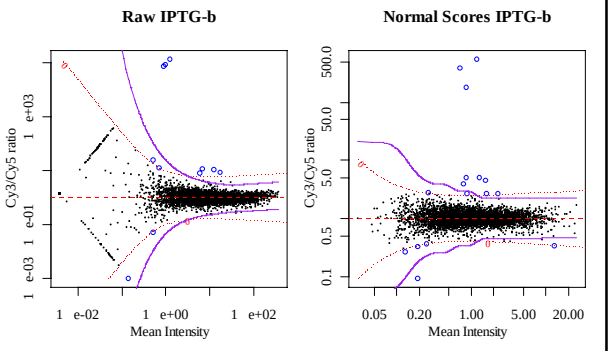
Why not use log transform?

- 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

Comparison with *E. coli* Data

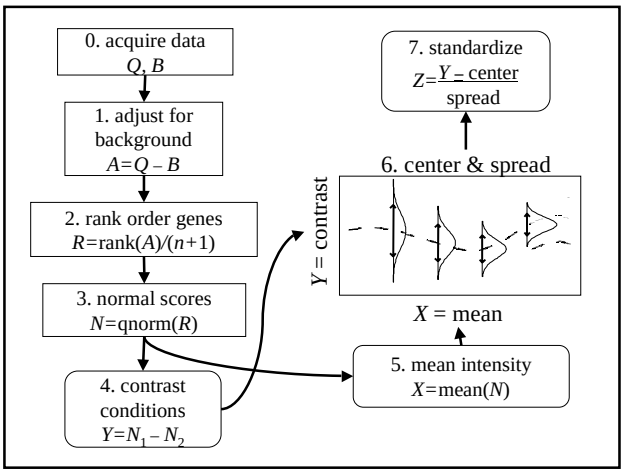
- 4,000+ genes (whole genome)
- Newton et al. (2000) J Comp Biol
  - Bayesian odds of differential expression
- IPTG-b known to affect only a few genes
  - ~150 genes at low abundance
  - including key genes

E. coli with IPTG-b



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	$S = [Y - \mu(X)]/\sigma(X)$



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

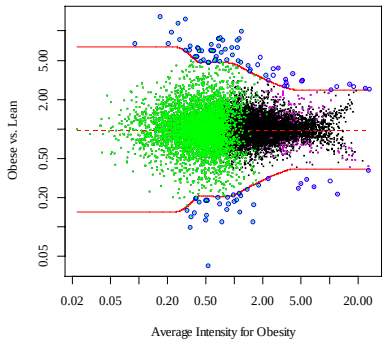
Robust Spread Details

- 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

Diabetes & Obesity Study

- 13,000+ gene 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

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

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

Microarray ANOVAs

- Kerr et al. (2000)
- gene by condition interaction
  - N_{ijk} = gene$_i$ + condition$_j$ + gene*condition$_{ij}$ + rep error$_{ijk}$
- conditions organized in factorial design
  - experimental units may be whole or part of array
- genes are random effects
  - focus on outliers (BLUPs), not variance components
  - gene*condition$_{ij}$ = differential expression
  - allow variance to depend on gene$_i$ main effect
- replication to improve precision, catch gross errors

Microarray Random Effects

- variance component for non-changing genes
  - robust estimate of MS(G*C) using smoothed MAD
  - rescale normal score response N by spread $\sigma(X)$
  - look for differential expression
    - or use clustering methods
- variance component for replication
  - robust estimate of MSE using smoothed MAD
  - look for outliers = gross errors

Principle Components

- Alter et al. (2000) for microarrays
  - see also Hilsenbeck et al. (1999)
- $N_{ij} = N_{ijk}$ = gene i for condition j (for rep k)
  - principle components (singular value decomposition)
  - $N = UDV^T$
    - D has eigen-values down diagonal
    - U has eigen-conditions as columns, genes as rows
    - V has eigen-genes as rows, conditions as column
- model for eigen-gene i
  - V_{ijk} = gene$_i$ + condition$_j$ + gene*condition$_{ij}$ + rep error$_{ijk}$

PCA Pros and Cons

- advantages
  - eigen-genes V_1, V_2, ... are orthogonal
  - may only need a few
    - how fast do eigen-values D drop?
- disadvantages
  - UDV^T may be expensive to compute
  - less efficient if many large eigen-values
  - may be difficult to interpret some eigen-genes
    - depends on choice of conditions
  - decomposition does not reflect experimental design
    - could improve via linear discriminant analysis (Fisher 1936)

Microarray QTLs using PCAs

- condition = genotype, array = individual
- V_{ijk} = gene$_i$ + QTL$_j$ + gene*QTL$_{ij}$ + individual$_{ijk}$
- QTL genotype depends on flanking markers
  - mixture model across possible QTL genotypes
  - single vs. multiple QTL
- single QTL may influence numerous genes
  - epistasis = inter-genic interaction
  - modification of biochemical pathway(s)

Multiple QTLs

- Zeng, Kao, et al. (1999, 2000)
  - multiple interval mapping (MIM)
- Satagopan, Yandell (1996); Stevens, Fisch (1998); Silanpää, Arjas (1998, 1999)
  - Bayesian interval mapping using RJ-MCMC
- True et al. (1997); Zeng et al. (2000)
  - first principle components as trait
  - MIM with interactions

LDAs for QTLs

- PCAs computed once
  - individuals are random sample from segregating population
  - expect major genotype effects to follow PCs
- LDAs could adjust to genotypes
  - start with PCs, hopefully close to LDs
  - LDA depends on unknown QTLs: decompose BW^{-1}
    - B is between genotype variation (QTL effects)
    - W is within genotype variation (error)
  - expensive computation: any shortcuts?

Obesity Genotype Main Effects