Finding Patterns of Differential Expression

identify modest number of genes to pursue
limit chance of false positives
single genes vs. joint behavior of groups
inference vs. reorganization
adjustments for multiple comparsions
diagnostic plots

classical inference (test for outliers)
Bayesian inference (probability on mixture)
dimension reduction (SVD = PCA)
unsupervised learning (clustering)
supervised learning (classification)

- formal inference
  paired comparisons (R/G, 2 conditions)
  test for outliers
  Chen et al. (1997); Dudoit et al. (2000a)

- improvements on the game
  allow variation to change with mean intensity
  experimental design & anova
  Roberts et al. (2000); Lin et al. (2001); Kerr et al. (2000abc)

- posterior probability of differential expression
  mixture model: changed & unchanged genes
  Newton et al. (2001); Sapir Churchill (2000) Efron et al. (2000); Lee et al. (2000)

- principal components: super-genes
  singular value decomposition: eigen-genes
  linear combinations of genes
  major direction of variation across conditions
  Alter et al. (2000); West et al. (2000)

- unsupervised learning
  clustering: organize genes into groups
  fancier flavors: plaid models, etc.
  prior knowledge vs. data-driven methods
  prescreening of genes?
  Eisen et al. (1998); Tamayo et al. (1999); Lazzeroni Owen (2000)

- supervised learning
  classification: assign genes to preset groups
  reduce many conditions & genes to a few
  correlation of genes with conditions
  Golub et al. (1999); Dudoit et al. (2000b)

Ratio-based Comparison of Two Conditions

Chen et al. (1997); de Risi et al. (1996)
cDNA spotted arrays: Red & Green samples
check two conditions on one microarray
large dynamic range in expression levels (3-5 orders of magnitude)

plot G vs. R on ordinary or log-log scales
note how common values of T lie along a line
– line through zero for ordinary scales
– parallel lines with slope 1 for log-log scale

$T = R/G \approx 1$ for unchanged genes
distribution of ratio is tricky
want to use most genes to derive test
combine variance estimates across genes

model for mRNA abundance

mRNA abundance kinetics

synthesis	polymerase	selected by condition
degradation	nuclease	generic cell machinery

abundance largely driven by selection factors

assumptions suspect at very high/low abundance
synthesis/degradation may be at limits

hypothesis: $H_0 : \mu_{R_k} = \mu_{G_k} = \mu_k$
hence: $\sigma_{R_k} = \sigma_{G_k} = c\mu_k$

assumptions across genes $k = 1, \cdots, n$

- some mean abundance $E(R_k) = \mu_{R_k}$
- spread proportional to abundance
  (constant coefficient of variation c)

$$SE(R_k) = \sigma_{R_k} = c\mu_{R_k}$$

$$V(R_k) = \sigma^2_{R_k} = c^2\mu^2_{R_k}$$

- distribution of abundance measurment is normal

$$R_k \sim N(\mu_{R_k}, c^2\mu^2_{R_k})$$

- same CV c for G_k

$$G_k \sim N(\mu_{G_k}, c^2\mu^2_{G_k})$$

distribution of ratio $T_k = R_k/G_k$

assume: R_k, G_k independent by experiment design
– separate measurements
– samples prepared separately
– negligible interaction at spot on array
however: strong correlation across genes
$k = 1, \cdots, n$

focus on any gene k and drop subscript

$$T = R/G = X/Y$$

equations (4)–(7) develop distribution

$f_{X,Y}(x, y) = f_X(x)f_Y(y)$ due to independence
$f_{R,G}(r, g) = f_R(r)f_G(g)$ in particular

hypothesis: $H_0 : \mu_{R_k} = \mu_{G_k} = \mu_k$
approximation: intensity measurements $Y > 0$
claim: introduces neglible error (equation 6)

$F_{T_k}(t)$ **does not depend on gene** k
follows from constant CV c and normality
note how μ and σ drop out

$$\frac{(g-\mu_k)^2}{2\sigma_k^2} = \frac{(g-\mu_k)^2}{2c^2\mu_k^2} = \frac{((g/\mu_k)-1)^2}{2c^2} = \frac{(u-1)^2}{2c^2}$$

$$\frac{(tg-\mu_k)^2}{2\sigma_k^2} = \frac{(tu-1)^2}{2c^2}$$

change of variables (calculus)
$u = g/\mu_k$ and $dg = \mu_k du$

shape of ratio distribution
approximation: normal tail is negligible (7)
– negligible mass beyond 3 SD for normal
– extend integral from $[0, \infty)$ to $(-\infty, , \infty)$
$f_T(t)$ slightly assymetric (Figure 5)

inference for ratio

confidence interval depends on c (Table 1-2, Fig 6)
"verified" by Monte Carlo simulation (details?)

maximum likelihood estimate of c equations (9) and (10) when calibrated
iteration to MLE when uncalibrated
$H_0 : \mu_{R_k} = m\mu_{G_k}$, m unknown
no closed form for MLE of m and c together

proposed using mode for m in text
actually use mean in algorithm
why might they be different?

R routine computes critical value (Newton)
$>$ `chen.poly(cv,err=.01)`
$>$ `library(microarray)`

dataset (ratios only):
`www.nhgri.nih.gov/DIR/LCG/ARRAY/expn.html`

Replicated cDNA Experiments

Dudoit et al. (2000)
diagnostic plot of raw data (Fig 3)
– R vs. G or $M = R/G$ vs. $A = \sqrt{RG}$
– log scale (base 2 or base 10)

normalization (Fig 4, 9)
$\log_2(R/G) - c_t(A)$
$c_t(A)$ = smooth fit to center of $M|A$
t = print tip for cDNA array
(variability?: later)

test statistic with replication

j = gene; 1,2 = condition
n_i arrays per condition

$$t_j = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{\sqrt{\frac{s^2_{1j}}{n_1} + \frac{s^2_{2j}}{n_2}}}$$

large $|t_j|$ indicate differential expression
how large?

replication needed to assess variation
and reduce chance of gross error
evaluate t_j using permutation test
allow for multiple comparisons

adjusting p-values with multiple comparisons

p_j = unadjusted, $\tilde{p}_j$ = unadjusted

Bonferroni method: $\tilde{p}_j = \min(np_j, 1)$
conservative, controls family-wise error (VWE) rate

Zidak method: $\tilde{p}_j = 1 - (1 - p_j)^n$
conservative, controls VWE rate, assumes independence

Holms step-down method:
rank-order the test statistics
compare to those left untested
$\tilde{p}_{(1)} = \min(np_{(1)}, 1)$
$\tilde{p}_{(j+1)} = \min(\tilde{p}_{(j)}, (n-j)p_{(j+1)}, 1)$

Westfall Young step-down method:
uses dependence structure of data
$\tilde{p}_{(1)} = \text{prob}\{\min_{\ell \in \{1:n\}} P_{(\ell)} \le p_{(1)} | H_0\}$
$\tilde{p}_{(j)} = \min(\tilde{p}_{(j-1)}, \text{prob}\{\min_{\ell \in \{j:n\}} P_{(\ell)} \le p_{(j)} | H_0\})$