Identifying Differentially Expressed Genes:

Dealing with the Multiple Comparison Issue when Simultaneously Testing Thousands of Hypotheses

(an example of the Storey & Tibshirani method of controlling the FDR)

Rhonda DeCook
University of Iowa

Application of the FDR error control

- Biomedical
- Geophysical
- Internet data
- Wavelet shrinkage
- Anywhere you have massive data sets (and multiple tests)
Microarray experiments - background

- DNA sequencing

- Genes are DNA sequences that code for proteins

- Proteins are the building block of organisms

- Measuring Proteins or mRNA in an organism
  - Medical and health studies
  - Progress in crop production

- Microarray Chips...

---

Microarray experiments - background

- Measuring mRNA with an Affymetrix GeneChip®
  - Creation of chip
    Thousands of genes represented, e.g. 22,810 for Arabidopsis plant
  - mRNA extraction from organism
  - Hybridization (applying sample to the chip) and measurement
Microarray experiments - background

- Example data for experiment with two conditions:

<table>
<thead>
<tr>
<th>Probe</th>
<th>EU 1-1</th>
<th>EU 1-2</th>
<th>EU 2-1</th>
<th>EU 2-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>219_at2678_at</td>
<td>173.8</td>
<td>593.0</td>
<td>582.2</td>
<td>320.3</td>
</tr>
<tr>
<td>2619_at518_at</td>
<td>417.5</td>
<td>387.0</td>
<td>552.0</td>
<td>542.8</td>
</tr>
</tbody>
</table>

A Probe is like a ‘gene’. The response is a measure of expression.

Microarray experiments - background

- Result: Thousands of measurements taken on each experimental unit (plant, person, etc.) in one condition.

- Common hypothesis at every gene:
  \[ H_0: \mu_1 = \mu_2 \quad \text{vs.} \quad H_A: \text{not } H_0 \]

- Often, this means testing thousands of hypotheses simultaneously leading to thousands of \( p \)-values to sift through.
Multiple comparison adjustment

- With no adjustment, use comparison-wise $\alpha = 0.05$
  - What is the probably of making no type I errors?
    for 22,810 independ. Tests $\left( \frac{0.95}{22810} \right) = 0.00008$
    $\rightarrow$ very likely to make at least one mistake

- Family-wise error rate $\alpha = 0.05$
  - Bonferroni adjustment (possible very few significant findings)
    - e.g. $p = 0.000002 \times 22,810 = 0.045$
    - raw p-value

---

Multiple comparison adjustment

- Researcher often thinks there are many genes that are differentially expressed.

- Trade-off between type I and type II errors.
  - Want enough significant genes for future research along with a reasonable error rate.

- Researcher is often willing to accept a reasonable number of false positives to detect more true positives.
Answer: False Discovery Rate (FDR)

- Benjamini Hochberg (1995) *first method introduced
- Storey & Tibshirani (2001) *modification of B & H method
- FDR is the expected proportion of false positives among all positive results.
- Allows researcher to choose a cut-off associated with a list of significant genes and an estimated error rate.

Distribution of $p$-values

- Main Idea:
  1) The distribution of $p$-values coming from null genes are uniformly distributed (0,1)
  
  ![Uniform Distribution](image)

  2) The distribution of $p$-values coming from genes that are differentially expressed is stochastically smaller than the uniform (0,1)

  ![Differential Expression](image)
Distribution of $p$-values

Histogram of observed 22,810 $p$-values testing

$H_0: \mu_1 = \mu_2$

---

Distribution of $p$-values

Dashed line is what we would expect if ALL 22,810 tests were actually null.

Initial estimated FDR (from B & H) when the chosen raw $p$-value cut-off is 0.001:

99 rejections (comes from the ordering of $p$-values)

$22,810 \times 0.001 = 22.8$ expected errors if all null,

estimated $FDR = \frac{p_{(k)}^m}{k} = 22.8/99 = 23\%$
False Discovery Rate

- Most likely, this is an **OVERESTIMATE** of the error rate.

- Issue: How can we estimate the proportion ($\pi_0$) of genes with a true null hypothesis? (because we probably don’t have 22,810 true nulls)

  This value will greatly effect our estimated FDR.

---

Estimating proportion of null genes $\pi_0$

- Use region of test statistic values where you expect only null test statistics for estimating $\pi_0$, the proportion of nulls (i.e. pick the region near 1.0 if using $p$-values)

  - **NOTE:** Region choice affects bias vs. variability trade-off
    - Storey & Tibshirani (2001)

  Using the region (0.95, 1.0), if all tests were null, I'd expect 1140.5 $p$-values to be in this interval. But only 705 are actually there.

  Thus, $\hat{\pi}_0 = 705/1140.5 = 62\%$

  **New FDR estimate:**

$$P_{\text{FDR}} \leq \frac{(0.001 \times 22,810 \times 0.62)}{99} = 14\%$$
Estimating proportion of null genes

Cubic Spline fit to $\pi$
estimates as a function of ‘region width’

- Storey and Tibshirani from (PNAS, 2003).
- Need to choose the ‘window’ for estimating null proportion.
- $\hat{\pi}$ is chosen to be the limit of the spline as ‘region width’ goes to 0.

Some examples of gene expression data sets...

Animal Science: muscle undergoing hypertrophy vs. stable muscle

Plant Pathology: roots infected with soybean cyst nematodes vs. unaffected roots

Genetics: wheel-running mice vs. non-runners

Plant Pathology: interaction between multiple kinds of powdery mildew fungus and multiple genotypes of barley.
Conclusions

- Massive data sets (and large-scale multiple testing) becoming more prevalent.

- FDR is a nice way to quantify the error rate when doing thousands of tests simultaneously.

- Variety of methods for estimating \( \pi \) (the proportion of nulls), and this is useful in FDR estimates.
From OLRT p. 96

The Student Newman–Keuls (SNK) procedure for controlling the FDR for pairwise comparisons is a step-down procedure, but it uses the studentized range distribution.

The procedure is very similar to the REGWR procedure, but the critical value for significance is less for stretches of means of a-2, a-3, ...

Order the differences, and start with a stretch of \( k = a \) means, declare the \( (\bar{y}_{\text{min}} - \bar{y}_{\text{max}}) \) to be significant if

\[
|\bar{y}_{\text{min}} - \bar{y}_{\text{max}}| > q_{\alpha} (k, df) \cdot \sqrt{\frac{1}{MSE} \frac{1}{h_i} + \frac{1}{h_j}}
\]

If significant, continue with the 2 stretches of \( k = a-1 \) means, as was done in REGWR, but always use \( u = q_{\alpha} (k, df) \) as the critical value.

For \( k \leq a-2 \), the SNK "threshold" will be smaller than the REGWR. But the SNK controls the FDR, not the FWER in the "strong" sense.
One-way ANOVA with 5 treatment groups
SNK method for controlling FDR on pairwise comparisons tests.

Three EUs are randomly assigned to each treatment.
This is a balanced completely randomized design (CRD).

SAS statements for testing pairwise comparisons using the SNK adjustment (controls FDR):

```
proc glm data=mine;
  class group;
  model y=group;
  means group/SNK;
run;
```

The GLM Procedure
Class Level Information
Class Levels Values
group 5 1 2 3 4 5

Number of Observations Read 15
Number of Observations Used 15

The GLM Procedure
Dependent Variable: y

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>459.06666667</td>
<td>114.7666667</td>
<td>31.30</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>10</td>
<td>36.66666667</td>
<td>3.6666667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>14</td>
<td>495.73333333</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>group</td>
<td>4</td>
<td>459.06666667</td>
<td>114.7666667</td>
<td>31.30</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>group</td>
<td>4</td>
<td>459.06666667</td>
<td>114.7666667</td>
<td>31.30</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
The GLM Procedure

Student-Newman-Keuls Test for y

NOTE: This test controls the Type I experimentwise error rate under the complete null hypothesis but not under partial null hypotheses.

Alpha 0.05
Error Degrees of Freedom 10
Error Mean Square 3.666667

Number of Means 2 3 4 5
Critical Range 3.4836233 4.2859233 4.7832114 5.1455143

The GLM Procedure

Student-Newman-Keuls Test for y

Means with the same letter are not significantly different.

S

N

K

G

r

o

u

p

i

n

S

G

m

b

i

n

G

r

o

u

p

i

n

A 30.333 3 5
B 23.667 3 2
C 19.333 3 4
C 18.333 3 3
D 14.000 3 1