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>...</th>
<th>EU 2-1</th>
<th>EU 2-2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>219_at2676_at</td>
<td>173.8</td>
<td>593.0</td>
<td>...</td>
<td>582.2</td>
<td>320.3</td>
<td>...</td>
</tr>
<tr>
<td>2619_at618_at</td>
<td>417.5</td>
<td>387.0</td>
<td>...</td>
<td>552.0</td>
<td>542.8</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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 $(0.95)^{22810} = 0.00008$
      -> 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.
Main Idea:

1) The distribution of $p$-values coming from null genes are uniformly distributed $(0,1)$

2) The distribution of $p$-values coming from genes that are differentially expressed is stochastically smaller than the uniform $(0,1)$
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 = $\frac{p(k)^m}{k} = \frac{22.8}{99} = 23\%$
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:
\[
\frac{p_{(k)} m_0}{k} = \frac{0.001 \times (22,810 \times 0.62)}{99} = 14\%
\]
Estimating proportion of null genes

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.

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

- large bias (many non-null genes)
- small variability

- small bias (mostly null genes)
- large variability
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.