CAP 5510: Introduction to Bioinformatics

Giri Narasimhan
ECS 254; Phone: x3748
giri@cis.fiu.edu
www.cis.fiu.edu/~giri/teach/BioinfS11.html
## Microarray Data

<table>
<thead>
<tr>
<th>Gene</th>
<th>Expression Levels</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sample A</td>
<td>Sample B</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CONTROL</td>
<td>TREATMENT</td>
<td></td>
</tr>
<tr>
<td>Gene1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Microarray Analysis

- Is Gene X upregulated? Downregulated? Had no change in expression levels?
  - Genes are represented by probes
  - Experiments may have repeats

- NULL HYPOTHESIS
  - There is no change in gene expression levels for Gene X between Control and Treatment
Accept/Reject $H_0$ (Null Hypothesis)?

- **P-value thresholds**
  - P-value is probability of data assuming $H_0$ holds
  - P-value threshold of 0.05 means probability of error when $H_0$ is rejected is 5%

- **Fold change**
  - If no repeats are done

- **t-Test**
  - Parametric
  - Non-parametric
    - Wilcoxon rank sum
# Hypothesis Testing Logic

## Hypothesis Choice

<table>
<thead>
<tr>
<th>Decision</th>
<th>H0</th>
<th>H1</th>
</tr>
</thead>
<tbody>
<tr>
<td>H0</td>
<td>Correctly Accept (TN)</td>
<td>Type II Error (FN) β</td>
</tr>
<tr>
<td>H1</td>
<td>Type I Error (FP) α</td>
<td>Correctly Reject (TP)</td>
</tr>
</tbody>
</table>

## Typical Values:
- Type I error of 0.05
- Type II error of 0.2
Problem with Hypothesis Testing

- Not testing just one gene
- If multiple genes are tested, then t-Test assumes each test is independent
- Are the tests independent? 
  - No!
- Need Correction
  - P-values need to be adjusted
  - Bonferroni or other correction methods needed
  - Achieved by controlling Type I error
Type I Error of 0.05 means that there is a 5% error in prediction of FN by t-Test.

**IMPLICATIONS?**

- If N=1000 genes & d=40 are differentially expressed (DE), then ...
  - $960 \times 0.05 = 48$ FPs
  - There are more FPs than TPs
  - Type I error and correcting for multiple hypothesis testing are connected
Multiple Test Corrections

- **Bonferroni correction**
  - Use type I error = \( \alpha / g = \text{FWER} = 0.05/1000 \)
    - Family-wise Error (FWER)
    - Too Conservative! Also reduce true positives!

- **Other less conservative corrections possible**
  - Sidak correction, Westfall-Young correction, ...

- **Using False Discovery Rate (FDR) [Benjamini & Hochberg ’95, Storey ’02 & ’03]**
  - Earlier: 5% of all tests will result in FPs
  - With FDR adjusted p-value (or q-value): 5% of **significant** tests will result in false positives.
Consider example shown. Let $N = 839$. Marked item has $p$-value 0.01 and $q$-value 0.0141. **P-value threshold** of 0.01 implies a 1% chance of false positives. Thus, we expect $839 \times 0.01 = 8.39$ FPs. Since item has rank 52, we expect to have 8 or 9 of these to be FPs.

**Q-value threshold** of 0.0141 implies a 1.41% of all spots with $q$-value less than this to be FPs. Thus, we expect $52 \times 0.0141 = 0.7332$ FPs, i.e., less than one FP.