Announcements

- Presentation: tentative 11/30 (Tu)
  - 15 minutes oral presentation on any of the following:
    - Case study: analyzing data set(s) of your choice
    - Journal club: review/critique of a paper on microarray data analysis
    - “Proposal” due: 11/9 (Tu) to cbmb@biostat.ucsf.edu
    - Can work alone or in pairs

- CBMB 1-day short course “Microarrays: Case Studies and Advanced Analysis” on Saturday, 10/23 in Genentech Hall Auditorium, Mission Bay:
  [http://www.biostat.ucsf.edu/cbmb/courses/short04.html](http://www.biostat.ucsf.edu/cbmb/courses/short04.html)

Lecture 2: Array Analysis I
Identify Differentially Expressed Genes in Two-Sample Studies
Ru-Fang Yeh

Analysis of expression data:
- Identify D.E. genes, estimation and testing,
- clustering, and
- discrimination.

Quality assessment
Pre-processing

Microarray experiment

Microarray experiment

Quality Measurement

Image analysis

Pass

Failed

Preprocessing

Analysis

Estimation

Testing

Annotation

Clustering

Discrimination

Biological verification and interpretation

Adapted from J Yang, UCSF
Agenda

What is the question?
- What are the target genes of my knock-out gene?
- Is there a specified group of genes all up-regulated in my treated samples?

⇒ Identify differentially expressed genes in A vs B

- Two-sample statistics
- Assess significance: Multiple testing issues

Log-Ratios or Single Channel Intensities?

- We assume proper normalization is done.
- For two-color spotted arrays, most analysis uses log-ratios $M = \log(R/G)$ as the primary data, i.e., gene expression measurements are relative.
- Alternative approach treats individual channel intensities R and G as primary data, i.e., gene expression measures are absolute (Wolfinger et al, Kerr et al), and
  - makes stronger assumptions
  - requires more complex models (mixed models in place of ordinary linear models) to accommodate correlation between R and G on same spot
  - requires single-channel normalization methods

Fold Change vs. Signal Intensity Plot

- Actual fold-differences measured by real-time PCR are in many cases much greater than those calculated by analysis of arrays
- Simple fold-change rules give no assessment of statistical significance
  ⇒ Need replicates to estimate the variability and construct test statistics

Two components in identifying DE genes

1. Select a statistic which will rank the genes in order of strength of the evidence for differential expression, from strongest to weakest.

2. Choose a critical value for the ranking of statistics above which any value is considered to be significant.
**Statistical Test & p-value**

- Example: The two-sample t-statistic
  \[ T = \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \]
  is used to test equality of the group means \( \mu_1, \mu_2 \)

- The *p-value* \( p^* \) is the probability that, under the null hypothesis (H0: \( \mu_1 = \mu_2 \)), the test statistic is at least as extreme as the observed value \( t^* \).

**Two-sample comparisons with replicates**

Perform statistical tests on normalized (log-transformed) data:

- **Standard t-test**: assumes normally distributed data in each class (questionable), equal variances within classes
- **Welch t-test**: as above, but allows unequal variances
- **Wilcoxon test**: non-parametric, rank-based
- **Permutation test**: estimate the distribution of the test statistic under the null hypothesis by permutations of the sample labels

**When there are few replicates…**

Gene expression data on 18,000 genes for 6 samples (replicates)

<table>
<thead>
<tr>
<th>Genes</th>
<th>mRNA samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M1</td>
</tr>
<tr>
<td>1</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>-0.10</td>
</tr>
<tr>
<td>3</td>
<td>0.15</td>
</tr>
<tr>
<td>4</td>
<td>-0.45</td>
</tr>
<tr>
<td>5</td>
<td>-0.06</td>
</tr>
</tbody>
</table>

\[ M = \log_2(Tumor / Normal) \]

For each spot, we have \( M_1, M_2, \ldots, M_6 \).

**Identify DE genes with Average \( M \)?**

Can we trust average effect sizes \( \bar{M} = \frac{1}{6} \sum_{i=1}^{6} M_i \) alone?

- Averages can be driven by outliers.

<table>
<thead>
<tr>
<th>Gene</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M6</th>
<th>Mean</th>
<th>SD</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2.5</td>
<td>2.7</td>
<td>2.5</td>
<td>2.8</td>
<td>3.2</td>
<td>2</td>
<td>2.61</td>
<td>0.40</td>
<td>16.10</td>
</tr>
<tr>
<td>B</td>
<td>0.01</td>
<td>0.05</td>
<td>-0.05</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0.03</td>
<td>0.25</td>
</tr>
<tr>
<td>C</td>
<td>0.5</td>
<td>0.7</td>
<td>-0.5</td>
<td>-0.8</td>
<td>20</td>
<td>0.01</td>
<td>3.31</td>
<td>8.19</td>
<td>0.41</td>
</tr>
<tr>
<td>D</td>
<td>0.5</td>
<td>0.2</td>
<td>0.1</td>
<td>-0.3</td>
<td>0.3</td>
<td>0.13</td>
<td>0.27</td>
<td>1.19</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0.1</td>
<td>0.11</td>
<td>0.1</td>
<td>0.11</td>
<td>0.09</td>
<td>0.10</td>
<td>0.01</td>
<td>33.09</td>
<td></td>
</tr>
</tbody>
</table>
Identify DE genes with $t$-statistic?

Can we trust the t-statistic $t = \frac{M}{s\sqrt{1/6}}$ alone?

- $t$ statistics can be driven by tiny variances.

<table>
<thead>
<tr>
<th>Gene</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M6</th>
<th>Mean</th>
<th>SD</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2.5</td>
<td>2.7</td>
<td>2.5</td>
<td>2.5</td>
<td>3.2</td>
<td>2</td>
<td>2.61</td>
<td>0.40</td>
<td>16.10</td>
</tr>
<tr>
<td>B</td>
<td>0.01</td>
<td>0.05</td>
<td>-0.05</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0.03</td>
<td>0.25</td>
</tr>
<tr>
<td>C</td>
<td>2.5</td>
<td>2.7</td>
<td>2.5</td>
<td>1.8</td>
<td>20</td>
<td>1</td>
<td>5.08</td>
<td>7.34</td>
<td>1.69</td>
</tr>
<tr>
<td>D</td>
<td>0.5</td>
<td>0</td>
<td>0.2</td>
<td>0.1</td>
<td>-0.3</td>
<td>0.3</td>
<td>0.13</td>
<td>0.27</td>
<td>1.19</td>
</tr>
<tr>
<td>E</td>
<td>0.1</td>
<td>0.11</td>
<td>0.1</td>
<td>0.11</td>
<td>0.09</td>
<td>0.10</td>
<td>0.01</td>
<td>0.01</td>
<td>15.95</td>
</tr>
</tbody>
</table>

3) Penalized-$t$

Trying to find a compromise between solely using $t$ and solely using mean. There are several similar solutions of the following form:

$$t^* = \frac{\bar{M}}{(s + a)/\sqrt{n}}$$

where $s$ = standard deviation.

Question: how to estimate $a$?
- minimizes the coefficient of variation(cv) of the absolute $t$-values (SAM). Tusher et al (2001)

Other Statistics (cont.)

d) $B$-statistic: log posterior odds ratios

Odds ratios is the ratio between the probability that a given gene IS differentially expressed over the probability that a given gene IS NOT differentially expressed. Lonnstedt & Speed (2002). Equivalent (for the purpose of ranking genes) to penalized-$t$ of form

$$t^* = \frac{\bar{M}}{\sqrt{(a + s^2)/n}}$$

where $a$ is estimated from the mean and variance of $s^2$

e) Moderated $t$-statistics (G Smyth 2004, Limma):

$$t^* = \frac{\bar{M}}{\tilde{s}/\sqrt{n}}$$

where $\tilde{s}^2 = \frac{s^2 d + s_0^2 d_0}{d + d_0}$ is the shrunken estimate of standard deviation
Other Statistics (cont.)

f) Generalize t-statistics with robust estimates of $\overline{M}$ and $s$.

g) “Single-channel” methods modeling absolute gene-expression levels:
   - Newton et al 2001: log-intensities ~ Gamma
   - Wolfinger et al 2001: linear mixed model on log-intensities

h) Composite methods: DEDS (Yang, Xiao & Segal 2004)

Finding differentially expressed genes

<table>
<thead>
<tr>
<th>Method</th>
<th>Taking variability into account</th>
<th>Using all genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Median</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>t-statistic</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Moderated t-statistic (limma)</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>SAM</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>B-statistic (log odds)</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>

Steps to generate a list of DE genes?

- For every gene, calculate $S_i = f(M1, M2, M3, M4, M5, M6)$ e.g. mean, t-statistics
- Statistics of interest $S_1, S_2, \ldots, S_{18000}$
- Rank and select the top n genes
- A list of candidate DE genes
Assessing significance

a) Diagnostic plots: q-q plot, histogram.
b) Testing: p-values, adjusted p-values.

Histogram and qq-plot for t-statistics. Looking for points off the straight line

Steps to generate a list of DE genes?

SCIENTIFIC: To determine which genes are differentially expressed between two sources of mRNA (T, C).

STATISTICAL: To assign appropriately adjusted p-values to thousands of genes, and/or make statements about false discovery rates.

Univariate hypothesis testing E.g. For a single gene

We wish to test the null hypothesis $H$ that the gene is not differentially expressed.

Traditionally, if you have a large data set or know the distribution, you can read the p-value off a table. Alternatively, you can do a non-parametric permutation test.

Is this p-value correct?

Testing: motivation

Many tests: a simulation study

Why do we adjust?

- Simulations of this process for 10,000 genes with 6 log-ratios.
- All the gene expression values were simulated independent and identically distributed (i.i.d) from a Normal (0,1) distribution, i.e. NOTHING is differentially expressed.
- Clearly we can’t just use standard p-value thresholds (.05, .01).
- We will expect $10000*0.01 = 100$ genes with p-value < 0.01.

<table>
<thead>
<tr>
<th>Index</th>
<th>t-stat</th>
<th>Unadj.P</th>
<th>Adj.P</th>
</tr>
</thead>
<tbody>
<tr>
<td>8167</td>
<td>3.88</td>
<td>0.0001</td>
<td>0.93</td>
</tr>
<tr>
<td>7440</td>
<td>-3.57</td>
<td>0.0003</td>
<td>0.93</td>
</tr>
<tr>
<td>8184</td>
<td>3.57</td>
<td>0.0003</td>
<td>0.93</td>
</tr>
<tr>
<td>548</td>
<td>3.52</td>
<td>0.0004</td>
<td>0.93</td>
</tr>
<tr>
<td>2295</td>
<td>3.49</td>
<td>0.0006</td>
<td>0.93</td>
</tr>
<tr>
<td>5806</td>
<td>-3.41</td>
<td>0.0009</td>
<td>0.98</td>
</tr>
<tr>
<td>3180</td>
<td>-3.32</td>
<td>0.0011</td>
<td>0.98</td>
</tr>
<tr>
<td>1840</td>
<td>-3.26</td>
<td>0.0015</td>
<td>0.98</td>
</tr>
<tr>
<td>6852</td>
<td>3.17</td>
<td>0.0016</td>
<td>0.98</td>
</tr>
</tbody>
</table>
Multiple Hypothesis Testing

Hypotheses

\[ H_0 \]: # true null hypotheses (non-diff. genes)
\[ H_a \]: # false null hypotheses (diff. genes)

\[ \begin{array}{ccc}
U & V & m_0 \\
T & S & m_1 \\
m - R & R & m
\end{array} \]

\[ \text{Type I error} \]

Type I Error Rates (False Positives)

- **Family-Wise Error Rate (FWER)**
  \[ Pr(V > 0) = Pr( \text{At least one false positive} ) \]

- **False Discovery Rate (FDR)** -- The FDR (Benjamini & Hochberg 1995) is the expected proportion of type I errors among the rejected hypotheses.
  \[ \text{FDR} = E(Q), \]
  With \[ Q = \begin{cases} V/R, & \text{if } R > 0 \\ 0, & \text{if } R = 0 \end{cases} \]

Multiple Testing: Controlling a Type I Error Rate

**AIM:**

For a given type I error rate \( \alpha \), use a procedure to select a set of “significant” genes that guarantee a type I error rate \( \leq \alpha \).

Adjusted p-values: Controlling the FWER

- **Bonferroni** correction: \( m \alpha \); most conservative adjustment. assume independence among genes.
- **Sidák**: \( 1-(1-\alpha)^m \)
- **minP** (Westfall & Young):
  \[ \tilde{P}_g = Pr( \min_{k=1, \ldots, m} P_k \leq p_g | H_0) \]
  estimated through permutation; allow dependency between genes.
- **maxT**: replace \( P_g \) by test statistics \( T_g \); min by max. Less computationally intensive than minP.
- Step-down
- Step-up

Choosing all genes with adjusted p-value \( \tilde{P}_g \leq \alpha \) controls the FWER at level \( \alpha \).
**FWER: Method Comparison**

Golub (1999) data --
3051 genes
27 ALL vs 11 AML samples

**Controlling the FDR (Benjamini/Hochberg)**

- Order unadjusted p-values: \( p_{r_1} \leq p_{r_2} \leq \ldots \leq p_{r_m} \).
- To control \( FDR = E(V/R) \) at level \( \alpha \), reject the hypothesis
  \( H_{r_j} \) for \( j = 1, \ldots, j^* \).
  \( j^* = \max \{ j : p_{r_j} \leq (j/m)\alpha \} \).
- Adjusted p-values:
  \[ \hat{p}_{r_j} = \min_{k=j, \ldots, m} \left\{ \min \left( \frac{m}{k} p_{r_k}, 1 \right) \right\} . \]
- Interpretation: expect 5% false positives among genes with < 0.05 FDR-adjusted p-values.

**Estimation of the FDR (SAM, Storey/Tibshirani 2003)**

Idea: Depending on the chosen cutoff-value for the test statistic \( T_g \),
estimate the expected proportion of false positives in the resulting
gene list through a permutation scheme.

1. Estimate the number \( m_0 \) of non-diff. genes.

2. Estimate the expected number of false positives under the
complete null hypothesis, \( E(V_0) \), through resampling.
Then, \( E(V) = \frac{m_0}{m} E(V_0) \) (because only the non-diff. genes may yield
false positives).

3. Estimate \( FDR = E(V/R) \) by \( \frac{E(V)}{R} \).

**q-value** of a gene = min estimated FDR at which it appears significant

**SAM-FDR: Example**

宓Consider the distribution of
p-values: A gene with \( p > 0.5 \)
is likely to be not differentially
expressed.

宓As p-values of non-
diff. genes should be uniformly
distributed in \( [0, 1] \), the number
\( 2 \times \# \{ g | p_g > 0.5 \} \) can be taken
as an estimate of \( m_0 \).

宓In the Golub example with
3051 genes, \( \hat{m}_0 = 1592 \).
FWER or FDR?

- Choose FWER if high confidence in ALL selected genes is desired (for example, selecting candidate genes for RT-PCR validation). Loss of power due to strong control of type-I error.

- Use more flexible FDR procedures if certain proportions of false positives are tolerable (e.g. gene discovery, selecting candidate co-regulated gene sets for GO/pathway analysis).
Multiple Testing References

- R/Bioconductor: http://www.bioconductor.org
  - multtest
  - Siggenes
- SAM: http://www-stat.stanford.edu/~tibs/SAM/

Conclusion

- How many genes are DE in my A vs B?
  - Assigning absolute significance levels (p-values) on the basis of probability models is problematic:
    - hard to check the normality assumption of log-ratios
    - Log-ratios of different genes are correlated in unknown structure
    - High-level of multiple testing means that very small nominal p-values are required -- dependent heavily on extreme tail of test statistic distribution
    - Nice statistical asymptotic theory (such as central limit theorem) unlikely to apply
  - Interpret p-values carefully (as always)
  - DE is continuous, not binary…
  - It’s more realistic to rank the genes in order of evidence for DE