Microarray Analysis

Chao-Jen Wong
Fred Hutchinson Cancer Research Center

9-10 December, 2010
Basics

Moderated $t$-tests

Using limma

$p$-value Correction

Resources
Introduction

- Identify differentially expressed genes associated with biological or experimental conditions.
- Primarily concerned with two-class problems.
- Data with $n$ samples and $p$ probes ($p \gg n$).

<table>
<thead>
<tr>
<th>A</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th></th>
<th>B</th>
<th>B</th>
<th>B</th>
<th>B</th>
<th>B</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{1,1}$</td>
<td>$x_{1,2}$</td>
<td>$x_{1,3}$</td>
<td>$x_{1,4}$</td>
<td>$x_{1,5}$</td>
<td></td>
<td>$x_{1,6}$</td>
<td>$x_{1,7}$</td>
<td>$x_{1,8}$</td>
<td>$x_{1,9}$</td>
<td>$x_{1,10}$</td>
<td></td>
</tr>
<tr>
<td>$x_{2,1}$</td>
<td>$x_{2,2}$</td>
<td>$x_{2,3}$</td>
<td>$x_{2,4}$</td>
<td>$x_{2,5}$</td>
<td></td>
<td>$x_{2,6}$</td>
<td>$x_{2,7}$</td>
<td>$x_{2,8}$</td>
<td>$x_{2,9}$</td>
<td>$x_{2,10}$</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>$x_{p,1}$</td>
<td>$x_{p,2}$</td>
<td>$x_{p,3}$</td>
<td>$x_{p,4}$</td>
<td>$x_{p,5}$</td>
<td></td>
<td>$x_{p,6}$</td>
<td>$x_{p,7}$</td>
<td>$x_{p,8}$</td>
<td>$x_{p,9}$</td>
<td>$x_{p,10}$</td>
<td></td>
</tr>
</tbody>
</table>
Approaches

- Gene-by-gene hypothesis testing
  - Treating each gene independently of others.
  - Goal: find statistically significant associations of biological conditions.
  - Genes are deemed to be interesting if the $p$-value is small.
  - Method: $t$-tests, moderated $t$-tests, ROC, $F$-test.

- Machine learning
$t_g = \frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2 - \sigma_y^2}}$

Drawback:

- Parametric assumptions hard to justify with few arrays.
- The variance in small samples might be noisy.
- Genes with small fold-change might be significant from statistical, not biological point of view.
Moderated $t$-statistics

- Rather than estimating within-group variability for each gene, pool the global information from all other genes.
- Advantage: eliminate occurrence of accidentally large $t$-statistics due to accidentally small within-group variance.
Moderated $t$-statistics

Using empirical Bayesian approach to estimate:

- Overall estimate variation $s_0^2$.
- Per-gene deviation variation $s_g^2$.
- Shrinkage variation

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

- Contrast estimator $\hat{\beta}_g$ – the difference in means between two classes.
- Moderated $t$-statistics:

$$\tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{\nu_g}}$$
Using limma

1. Define a design matrix to establish parameters of linear model matrix.
2. Fit a linear model for each gene based on the given design matrix (and a contrast matrix): lmFit().
3. Use function eBayes to get moderated $t$-statistics and relevant statistics.
Deriving linear models

Suppose we define a design matrix as the following:

<table>
<thead>
<tr>
<th>sample $i$</th>
<th>(intercept)</th>
<th>mol.biolNEG</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEG</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BCR/ABL</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>NEG</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Each gene $Y_j$ for all sample $i$, the expression level can be expressed by

$$
\begin{bmatrix}
Y_{NEGi,j} \\
Y_{BCR/ABL_i,j}
\end{bmatrix} = \begin{bmatrix} 1 & 1 \\
1 & 0
\end{bmatrix} \begin{bmatrix} \beta_{\text{intercept}} \\
\beta_{\text{mol.biolNEG}}
\end{bmatrix} + \epsilon
$$

$$
\Rightarrow \beta_{\text{mol.biolNEG}} = Y_{BCR/ABL_i,j} - Y_{NEGi,j} + \epsilon
$$

$$
y_j = \beta_{\text{intercept}} + \beta_{\text{mol.biolNEG}} a_{ij} + \epsilon
$$

$$
\Rightarrow y_j = \beta_0 + \beta_1 a_{ij} + \epsilon
$$
Using limma

Step 1:

**Code: define design matrix and contrast model**

> library(limma)
> design <- model.matrix(~mol.biol, ALLfilt_bcrneg)
>

Step 2:

**Code: linear models and eBayes**

> fit1 <- lmFit(exprs(ALLfilt_bcrneg), design)
> fit2 <- eBayes(fit1)
> topTable(fit2, coef=2, adjust.method="BH",
+ number=5)
Suppose we define a design matrix as the following:

<table>
<thead>
<tr>
<th>sample</th>
<th>mol.biolBCR</th>
<th>mol.biolNEG</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCR/ABL</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>BCR/ABL</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>BCR/ABL</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>NEG</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>NEG</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>NEG</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

\[ y_i = \beta_1 a_{ij} + \beta_2 b_{ij} + \varepsilon_i \]
Using limma

Step 1:

**Code: define design matrix and contrast model**

```r
> library(limma)
> design <- model.matrix(~0+mol.biol, ALLfilt_bcrneg)
> colnames(design) <- c("BCR_ABL", "NEG")
> contr <- makeContrasts(BCR_ABL-NEG, levels=designs)
> # contr <- c(1, -1)
```

Step 2:

**Code: linear models and eBayes**

```r
> fit <- lmFit(exprs(ALLfilt_bcrneg), design)
> fit1 <- contrasts.fit(fit, contr)
> fit2 <- eBayes(fit1)
> topTable(fit2, adjust.method="BH", number=5)
```
$t$-tests vs. moderated $t$-tests

- In larger sample size, there is not big difference between the ordinary and the moderated tests.
- For smaller sample size the difference will be larger.

The empirical Bayes moderation is more useful in cases with fewer replicates.
t-tests vs. moderated t-tests
$t$-tests vs. moderated $t$-tests

6 samples -- 3 for each group
**p-value corrections**

- Basic idea: reduce critical value used to reject.
- Trade-off between sensitivity and specificity.
- Approaches implemented in the *multtest* package:
  - Criteria for error rate control include family-wise error rate (FWER) and false discovery rate (FDR).
  - Permutation-based maxT methods.
Lab activity

- Chapter 6 and 7 in *Bioconductor Case Studies*.
- Goals: get familiar with functions provided by *Bioconductor* packages to perform differential expression analysis.
Resources