Differential Expression

Chao-Jen Wong

Fred Hutchinson Cancer Research Center

January 28, 2010

1. Differential Expression

2. Moderated t-statistics

3. Linear Models

4. Using the limma Package

Outline

1. Differential Expression
2. Moderated t-statistics
3. Linear Models
4. Using the limma Package

- Identify differentially expressed genes associated with biological or experimental conditions.
- Many different gene-by-gene approaches: fold-change, *t*-statistics, empirical Bayesian, moderate *t*-statistics, ROC, etc.
- Primarily concerned with two-class problems.
- Data with n samples and p probes ($p >> n$).

A	A	A	A	A	B	B	B	B	B
$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$	$x_{1,6}$	$x_{1,7}$	$x_{1,8}$	$x_{1,9}$	$x_{1,10}$
$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$	$x_{2,6}$	$x_{2,7}$	$x_{2,8}$	$x_{2,9}$	$x_{2,10}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{p,1}$	$x_{p,2}$	$x_{p,3}$	$x_{p,4}$	$x_{p,5}$	$x_{p,6}$	$x_{p,7}$	$x_{p,8}$	$x_{p,9}$	$x_{p,10}$

Subsetting and non-specific filtering

`ALLfilt_bcrneg`: B-cell tumors found to carry out BCR/ABL mutation and those with no cytogenetic abnormalities, NEG.

non-specific filtering

```
> library(ALL)
> library(hgu95av2.db)
> data(ALL)
> bcell <- grep("^B", as.character(ALL$BT))
> types <- c("NEG", "BCR/ABL")
> moltyp <- which(as.character(ALL$mol.biol) %in% types)
> ALL_bcrneg <- ALL[, intersect(bcell, moltyp)]
> ALL_bcrneg$BT <- factor(ALL_bcrneg$BT)
> ALL_bcrneg$mol.biol <- factor(ALL_bcrneg$mol.biol)
> library(genefilter)
> filt_bcrneg <- nsFilter(ALL_bcrneg,
+                        require.entrez=TRUE,
+                        require.GOBP=TRUE,
+                        remove.dupEntrez=TRUE,
+                        feature.exclude="^AFFX",
+                        var.cutoff=0.5)
> ALLfilt_bcrneg <- filt_bcrneg$eset
```

Outline

1. Differential Expression
2. **Moderated t-statistics**
3. Linear Models
4. Using the limma Package

Getting Dataset

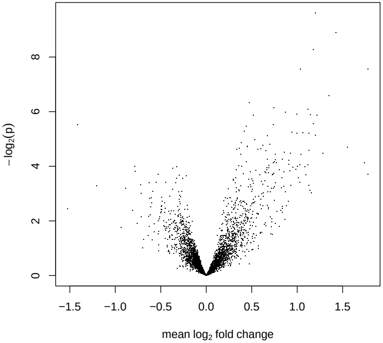
Alternatively, load the `ALLfilt_bcrneg` dataset from the *day2* package.

Data preparation

```
> library(day2)
> library(Biobase)
> data(ALLfilt_bcrneg)
```

Fold-change versus t-test

```
> library(genefilter)
> tt <- rowttests(ALLfilt_bcrneg, "mol.biol")
> plot(tt$dm,  -log10(tt$p.value), pch=".",
+      xlab=expression(mean~log[2]~fold~change),
+      ylab=expression(-log[2](p)))
```



Fold-change and t-test

t-statistics:

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

Drawback:

- The variance in small samples might be noisy.
- Genes with small fold-change might be significant from statistical, not biological point of view.

Moderate t-statistics

Using Bayesian approach to estimate:

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

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

  where $\frac{d_0}{d_0+d_g}$ is weight coefficient associated with all probes and $\frac{d_g}{d_0+d_g}$ is associated with gene g.
- Contrast estimator $\hat{\beta}_g$ – the difference in means between two classes.
- Moderate t-statistics:

  $$\tilde{t_g} = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{\nu_g}}$$

Outline

1. Differential Expression
2. Moderated t-statistics
3. **Linear Models**
4. Using the limma Package

Deriving linear models

Suppose we define a design matrix as the following:

sample i	(intercept)	mol.biolBCR
NEG	1	0
BCR/ABL	1	1
$\vdots$	$\vdots$	$\vdots$

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

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

$$\Rightarrow \beta_{mol.biolBCR} = Y_{BCR/ABL_i,j} - Y_{NEG_i,j} + \epsilon$$

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

$$\Rightarrow y_j = \mu + \beta a_{ij} + \epsilon$$

Define parameters in linear models

Define the linear model by

$$y_i = \mu + \beta a_{ij} + \varepsilon,$$

where $a_{ij} = 1$ if sample $i \in \{BCR/ABL\}$

```
> model.matrix(~ mol.biol,
+              ALLfilt_bcrneg)
```

```
      (Intercept) mol.biolNEG
01005           1           0
01010           1           1
03002           1           0
04007           1           1
04008           1           1
04010           1           1
04016           1           1
06002           1           1
08001           1           0
08011           1           0
08012           1           1
08024           1           1
09008           1           0
09017           1           1
```

Outline

1. Differential Expression
2. Moderated t-statistics
3. Linear Models
4. **Using the limma Package**

Using limma

1. Use design matrix to establish parameters of the model model.matrix.
2. Use linear model to fit the contrast parameters: lmFit().
3. Use function eBayes to get moderate t-statistics and relevant statistics.

Using limma

Step 1:

code: define design matrix and contrast model

```
> library(limma)
> #design <- model.matrix( ~mol.biol, ALLfilt_bcrneg)
> cl <- as.numeric(ALLfilt_bcrneg$mol.biol=="BCR/ABL")
> design <- cbind(intercept=1, mol.biolBCR=cl)
```

Step 2:

Code: linear models and eBayes

```
> fit1 <- lmFit(exprs(ALLfilt_bcrneg), design)
> #fit1 <- contrasts.fit(fit1, contr)
> fit2 <- eBayes(fit1)
```

Using limma

Code: getting top genes

```
> topTable(fit2, coef=2, adjust.method="BH",
+          number=5)

           ID    logFC  AveExpr
1117  1635_at 1.202675 7.897095
3050  1674_at 1.427212 5.001771
2171 40504_at 1.181029 4.244478
2816 40202_at 1.779378 8.621443
799  37015_at 1.032702 4.330511
            t      P.Value    adj.P.Val
1117 7.408878 1.017739e-10 3.910154e-07
3050 7.059429 4.898793e-10 9.410581e-07
2171 6.705277 2.368917e-09 3.033793e-06
2816 6.354009 1.107794e-08 1.064036e-05
799  6.299154 1.406498e-08 1.080753e-05
             B
1117 13.998069
3050 12.530820
2171 11.058580
2816  9.617537
799   9.394541
```

Reference

- G.K. Smyth, Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.
- G. K. Smyth, *limma: Linear Models for Microarray Data*, Bioconductor package vignette, 2005.
- Y. Benjamini and Y. Hochbert, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B*, 57(1): 289-300, 1995.

Lab activity

1. Chapter 7. Read and do the exercise in section 7.3 to 7.5.
2. Activity: Expend your package by adding functions that generate top genes.

   Input *ExpressionSet* (i.e., `ALLfilt_bcrneg` and a cut off value for adj.P.Val (0.01) that defines differenctially expressed genes.

   Output A data.frame containing differentially expressed genes and their corresponding statistics.

Solutions

```
> myFunc <- function(eset, p.cutoff=0.01) {
+   design <- model.matrix( ~mol.biol, eset)
+   fit1 <- lmFit(exprs(eset), design)
+   fit2 <- eBayes(fit1)
+   tstats <- topTable(fit2, coef=2, adjust.method="BH",
+             number=dim(fit2)[1])
+   top <- tstats[tstats$adj.P.Val < p.cutoff, ]
+ }
> top <- myFunc(ALLfilt_bcrneg, 0.01)
```