Multiple Testing Issues with Gene Expression Data

Utah State University – Spring 2014
STAT 5570: Statistical Bioinformatics
Notes 3.1
References

- Chapter 15 of Bioconductor Monograph (course text)


Where are we?

Up to now:
- Intro. to microarray technology and estimating gene expression levels (preprocessing)
- Clustering and visualization (sometimes using a specific subset of genes)

Coming up:
- Testing for differential expression (DE)
  - finding a subset of “significant” genes
- Annotation and online resources
- Technologies other than microarrays

Here: what to do with DE test results
Differential Expression (DE) tests – basics

- Have 2 or more groups of samples
  ex: healthy, beg. disease, adv. disease

  Null: Gene expressed same in all groups
  Alt.: Gene not expressed same in all groups (biological relevance?)

- Result:
  Test Stat.: some “standardized” measure of DE
  – like a t-test, maybe
  P-value: some measure of “significance”
# load data and define gene to test
library(affy); library(ALL)
data(ALL) ; gn <- featureNames(ALL)
gn.test <- "39020_at"
t <- gn==gn.test
gn.exprs <- exprs(ALL)[t,81:110]
exprs.vals <- as.vector(gn.exprs)
cell <- c(rep(0,15),rep(1,15))
   # 0 for B-cell; 1 for T-cell

boxplot(exprs.vals~cell,main=gn.test,
       cex.lab=1.5,cex.main=1.5,
       xlab='T-cell indicator',
       ylab='RMA expression')
# Test for significance
a1 <- lm(exprs.vals~cell)
s1 <- summary(a1)
round(s1$coefficients,3)

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 8.432 | 0.149 | 56.768 | 0.000 |
| cell | 0.605 | 0.210 | 2.882 | 0.008 |

NOTE: In practice, we won’t use this simple t-test; we will improve on it later (Notes 3.3).
**Significance and P-values**

- Usually, “small” P-value $\Rightarrow$ claim significance
- Correct interpretation of P-value from a test of significance:
  
  “The probability of obtaining a difference at least as extreme as what was observed, just by chance when the null hypothesis is true.”

- Consider a t-test of $H_0: \mu_0 - \mu_1 = 0$, when in reality, $\mu_0 - \mu_1 = c$ (and SD=1 for both pop.)
- What P-values are possible, and how likely are they?
For each value of c, 1000 data sets (think of as 1000 genes) were simulated where two populations are compared, and the “truth” is \( \mu_0 - \mu_1 = c \). For each data set, the t-test evaluates \( H_0: \mu_0 - \mu_1 = 0 \) (think of as no change in expression level). The resulting P-values for all data sets are summarized in the histograms.

What’s going on here?
```r
set.seed(123)                  # (Don’t worry about this
c.N <- 1000                     # code; it’s just here for
c.c.list <- c(0,0.1,0.15,0.2)   # completeness)
c.k <- length(c.c.list)        
c.p.mat <- matrix(nrow=N,ncol=length(c.c.list))
j <- 0
for(c in c.c.list){
j <- j+1; p <- 1:N
for(i in 1:N){
x <- rnorm(50,mean=c,sd=1)
y <- rnorm(50,mean=0,sd=1)
resp <- c(x,y)
d <- c(rep(0,50),rep(1,50))
s <- summary(lm(resp~d))$coefficients
p.mat[i,j] <- s[2,4]}
}

par(mfrow=c(2,2))
for(i in 1:k){
  hist(p.mat[,i],xlab='P-value',ylab='Density',
       main=paste('Histogram when c =',c.c.list[i]))
}
```
Note:
- Even when there is no difference (c=0), very small P-values are possible
- Even for larger differences (c=0.2), very large P-values are possible
- When we look at a histogram of P-values from our test of DE, we have a mixture of these distributions (because each gene has its own true value for c)
n <- 200
x.mat <- y.mat <- matrix(nrow=n,ncol=k)
for(i in 1:k)
  {
    d <- density(p.mat[,i],n=n, from=0, to=1)
    x.mat[,i] <- d$x
    y.mat[,i] <- d$y/max(d$y)
  }
library(RColorBrewer)
cols <- brewer.pal(4, "PuOr")
par(mfrow=c(1,1))
plot(x.mat[,1],y.mat[,1],xlim=c(0,1),type='l',
    lwd=3, xlab='P-value',col=cols[1], ylim=c(0,1),
    ylab='Density, smoothed and scaled')
for(i in 2:k){lines(x.mat[,i],y.mat[,i],col=cols[i],
    lwd=3, lty=i)}
legend(0.7,1.0,paste('c =',c.list[1]),bty='n')
legend(0.6,0.8,paste('c =',c.list[2]),bty='n')
legend(0.5,0.6,paste('c =',c.list[3]),bty='n')
legend(0.45,0.45,paste('c =',c.list[4]),bty='n')
ALL subset example: observed P-values (simple t-test, comparing 15 B-cell to 15 T-cell)

Remember, this is a mixture of distributions.

A flat histogram would suggest that there really aren’t any: DE genes.

The peak near 0 indicates that: some genes are DE.

But which ones?
# load data and define genes to test
library(affy); library(ALL)
data(ALL) ; gn <- featureNames(ALL)
gn.exprs <- exprs(ALL[1:1000,81:110])
cell <- c(rep(0,15),rep(1,15))
  # 0 for B-cell; 1 for T-cell
# test for significance
gn.func <- function(exprs.vals)
  {
    a1 <- lm(exprs.vals~cell)
    s1 <- summary(a1)
    return(s1$coefficients[2,4])
  }
p.vec <- apply(gn.exprs,1,gn.func)
# look at results
hist(p.vec,main='Histogram for first 1000 genes',
    xlab='P-value',ylab='density',
    cex.lab=1.5,cex.main=1.5)

NOTE: In practice, we won’t use this simple t-test; we will improve on it later (Notes 3.3).
How to treat these P-values?

- Traditionally, consider some cut-off

  Reject null if P-value < \(\alpha\), for example (often \(\alpha = 0.05\))

- What does this mean?

  \(\alpha\) is the acceptable level of Type I error: 
  
  \[ \alpha = P(\text{reject null | null is true}) \]
Multiple testing

- We do this with many (thousands, often) genes simultaneously – say m genes

<table>
<thead>
<tr>
<th></th>
<th>Fail to Reject Null</th>
<th>Reject Null</th>
<th>Total Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null True</td>
<td>U</td>
<td>V</td>
<td>$m_0$</td>
</tr>
<tr>
<td>Null False</td>
<td>T</td>
<td>S</td>
<td>$m - m_0$</td>
</tr>
<tr>
<td></td>
<td>m-R</td>
<td>R</td>
<td>$m$</td>
</tr>
</tbody>
</table>

# of Type I errors: $V$

# of Type II errors: $T$

# of correct “decisions”: $U + S$
Error rates

- Think of this as a family of $m$ tests or comparisons
- Per-comparison error rate: $\text{PCER} = E[V/m]$
- Family-wise error rate: $\text{FWER} = P(V \geq 1)$
- What does the $\alpha$-cutoff mean here?
  
  Testing each hypothesis (gene) at level $\alpha$
  guarantees:

  $$\text{PCER} \leq \alpha$$

  - let’s look at why
What are P-values, really?

Suppose $T$ is the test stat., and $t$ is the observed $T$.

$$Pval = P(T > t \mid H_0)$$

Assume $H_0$ is true. Let $F$ be the cdf of $T$ and $f$ be pdf:

$$F(t) = P(T \leq t) = \int_{-\infty}^{t} f(t) dt = 1 - Pval$$

What is the distribution of $Y = F(t)$? Let $g$ be pdf of $Y$:

$$\frac{dy}{dt} = F'(t) = f(t), \quad g(y) = f(t) \left| \frac{dt}{dy} \right| = f(t) \left( \frac{1}{f(t)} \right) = 1$$

So $Y = 1 - Pval$ is $Uniform[0,1]$.

Then when $H_0$ is true, $Pval \sim U[0,1]$. 


P-values and $\alpha$ cut-off

- Suppose null is true for all $m$ genes - (so none of the genes are differentially expressed)

- Look at histogram of $m=1000$ $P$-values with $\alpha=0.05$ cut-off - about 50 “significant” just by chance these can be “expensive” errors

```r
set.seed(2); p <- runif(1000)
hist(p,xlab='P-values',main='',
    breaks=c(0:40)/40)
abline(v=0.05,col='red',lwd=3)
```

(Here, $\frac{V}{m} \approx \frac{50}{1000} = 0.05.$)
How to control this error rate?

Look at controlling the FWER:

Testing each hypothesis (gene) at $\alpha/m$ instead of $\alpha$
guarantees:

$\text{FWER} \leq \alpha$

This is called –

Bonferroni correction

but -

this is far too conservative for large $m$

```r
hist(p,xlab='P-values',main='',
     breaks=c(0:40)/40)
abline(v=0.05/1000,col='red',lwd=3)
```
A more reasonable approach

Consider these corrections sequentially:

Let $P_i$ be the P-value for testing gene $i$, with null $H_i$.
Let $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ be the ordered P-values.

Let $k$ be the largest $i$ for which $P_{(i)} \leq \frac{i}{m} \alpha$.

Reject all $H_{(i)}$ for $i = 1, 2, \ldots, k$.

Then for independent test statistics and for any configuration of false null hypotheses, this procedure guarantees: $E[V / R] \leq \alpha$. 
What does this mean?

- $V =$ # of “wrongly-rejected” nulls
- $R =$ total # of rejected nulls
- Think of rejected nulls as “discovered” genes of significance
- Then call $E[V/R]$ the FDR - False Discovery Rate
- This is the Benjamini-Hochberg FDR correction – sometimes called the marginal FDR correction
Benjamini-Hochberg adjusted P-values

Let $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ be the ordered P-values.

Let $P_{(i)}^{(adj)} = P_{(i)} \cdot \frac{m}{i}$.

If any $P_{(i)}^{(adj)} > 1$, reset it to 1.

If any $P_{(i)}^{(adj)} > P_{(i+1)}^{(adj)}$, reset it to $P_{(i+1)}^{(adj)}$

(starting at the end of the list, checking backwards).

Then $P_{(1)}^{(adj)} \leq P_{(2)}^{(adj)} \leq \ldots \leq P_{(m)}^{(adj)}$ are the ordered BH - FDR - adjusted P-values.
An extension: the q-value

- **P-value for a gene:**
  the probability of observing a test stat.
  more extreme when null is true

- **q-value for a gene:**
  the expected proportion of false positives
  incurred when calling that gene significant

- Compare (with slight abuse of notation):

\[
pval = P(T > t \mid H_0 \text{ true}) \quad qval = P(H_0 \text{ true} \mid T > t)
\]
Estimating the q-value

Let \( p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)} \) be the ordered P-values.

For \( \lambda = 0 \) to 0.95 by 0.01: \( \hat{\pi}_0(\lambda) = \frac{\#(p_j > \lambda)}{m(1 - \lambda)} \).

Let \( \hat{f} \) be the natural cubic spline with 3 df of \( \hat{\pi}_0(\lambda) \) on \( \lambda \).

Let \( \hat{\pi}_0 = \hat{f}(1) \). (\( \pi_0 = m_0/m \) is prop.of genes that are "truly null." )

Calculate \( \hat{q}(p_{(m)}) = \hat{\pi}_0 p_{(m)} \).

For \( i = m - 1, m - 2, \ldots, 1 \) calculate \( \hat{q}(p_{(i)}) = \min \left( \frac{\hat{\pi}_0 p_{(i)} m}{i}, \hat{q}(p_{(i+1)}) \right) \).
Interpretation

- P-value is a measure of significance in terms of the false positive rate: \( V/m \)

- q-value is a measure of significance in terms of the FDR (false discovery rate): \( E[V/R] \)
What other adjustments are there?

- More than we could talk about here:

  \[ pFDR = E[V/R \mid R>0] \]
  \[ gFWER(k) = P(V \geq k) \]
  \[ TPPFP(\alpha) = P(V/R > \alpha) \]

  maxT – based on ordered test statistics
  minP – based on ordered P-values
  many more … two-step, etc.

  (recall V = # of false disc., R = # of rejected nulls)

- Other ideas: estimating the FDR, estimating the proportion or number of false nulls
Return to example: first 1000 genes

# (use ALL results from slide 12 code;
# simple t-test, comparing B-cell to T-cell)
rawP <- p.vec
adjP <- p.adjust(p.vec,method='BH')
par(mfcol=c(2,2))
  # NOTE this is different from mfrow
hist(rawP,main='Raw P-values',
     cex.main=1.5)
hist(adjP,main='FDR-adjusted P-values',
     cex.main=1.5)

# See methods automatically available
p.adjust.methods

[1] "holm"       "hochberg"    "hommel"
[4] "bonferroni" "BH"        "BY"
[7] "fdr"        "none"
Comparison: raw, FDR-adj., q-values

In general, q-values tend to be less than FDR-adjusted p-values.
```r
par(mfrow=c(2,2))
library(RColorBrewer)
c.vec <- brewer.pal(4,"Blues")
t.raw <- rawP < 0.05; t.bonf <- rawP < 0.05/length(rawP)
t.FDR <- adjP < 0.05
use.col <- rep(c.vec[1],length(rawP))
plot(rawP, adjP, pch=16, cex=1.5, col=use.col, xlab='Raw P-value', ylab='FDR-adjusted P-values')
abline(0,1)
legend('bottomright',c('Non-significant',
paste('.05 cutoff: ',sum(t.raw)),
paste('.05 FDR: ',sum(t.FDR)),
paste('.05 Bonferroni: ',sum(t.bonf))),
col=c.vec,pch=16,pt.cex=1.5,bty='n')

# Compare these FDR-adjusted P-values with q-values
library(qvalue)
qvals <- qvalue(p.vec)$qvalues
plot(qvals,adjP,col='red',pch=16,cex=1.5, xlab='q-values', ylab='FDR-adjusted P-values')
abline(0,1)
```
Which error rate?

- Type I: call gene ‘candidate’ when it’s not
  - PCER / FWER / FDR / etc.
- Type II: fail to identify true candidate
- Relative value (I vs. II) depends on perspective
  - Wasted effort
  - Lost opportunity
- How to reconcile?
  - Sample size → power → low Type II
  - Statistical method → low Type I
Current Areas of Research

- Controlling error rates with multiple dependent tests
- Controlling error rates in multiple structured hypotheses (e.g., nested or conditional tests)
- Choosing an appropriate family
  - (within which collection of tests should error rates be controlled?)
Summary

- Tests of differential expression
  - Null: gene is not DE
  - Alt: gene is DE
  - Test Stat. → P-value

- How to treat P-values: uniform random variables

- Multiple comparison procedures
  - simple cut-off → too liberal
  - Bonferroni correction → too conservative
  - FWER
  - FDR and q-values
  - others – we may return to this topic: good 6570 projects