# An Introduction to *Hybrid Testing*

Stan Pounds, Demba Fofana

October 25, 2011

## 1 Introduction

One of the most challenging matters in multiple testing is to consider assumptions in testing procedures. People dealing with multiple testing scenario are in general overwhelmed by the size of the data, and then turn a blind eye on the assymptions. Turning a blind eye on assumptions can result on biased results or inaccuracy. The problem of multiple testing always occurs in gene expression data analysis. To solve the problem of multiple testing we have developed a method of testing that takes assumptions into considerations. So, for example when dealing with k-group comparison when the data are normally distributed our procedure gives high weight to ANOVA test and when the data are not normally distributed the procedure gives higher weigth to Kruskal Wallis test than to F-test. When dealing with regression analysis our procedure uses Pearson test if the data are normally distributed and Spearman test when the data are not normally distributed. For two-group comparison a t-test is used when the data are normally distributed and ranksum test is used when the data are not normally distributed. In two-group comparison the equality of variances is taken into consideration. Our methodology is applied to microarray data analysis for three-group comparison and Regression analysis.

## 2 Data Requirements

For regression analysis our procedure deals with complete data; preprocessing the data is then needed when there are missing data. When analysing data in k-group comparison the analyst need to specify clearly the different groups to compare. We have given some examples that will help understand how to procede. Suppose we have expression data and phenotype data; we need then to combine these data sets into an expression set, we call it express.set. The expression data are in file correlation.data and the phenotype data are also in correlation.data. Here is how to procede in order to conduct a regression study.

```
> library(HybridMTest)
> data(correlation.data)
> Y<-exprs(correlation.data)
```

```
> x<-pData(correlation.data)
> test.specs<-cbind.data.frame(label=c("pearson","spearman","shapiro"),
+                              func.name=c("row.pearson","row.spearman","row.slr.shapiro"),
+                              x=rep("x",3),
+                              opts=rep("",3))
> ebp.def<-cbind.data.frame(wght=c("shapiro.ebp","(1-shapiro.ebp)"),
+                           mthd=c("pearson.ebp","spearman.ebp"))
> corr.res<-hybrid.test(correlation.data,test.specs,ebp.def)
> head(corr.res)

  pearson.stat pearson.pval pearson.ebp spearman.stat
1    0.1691322   0.32408047   0.5733143    0.12188687
2    0.1378469   0.42270896   0.7304976    0.07632409
3    0.1563656   0.36245200   0.6256931    0.09859064
4    0.3080026   0.06762317   0.3960809    0.31198919
5    0.2266644   0.18372621   0.4389164    0.20168608
6    0.1362335   0.42821043   0.7418425    0.20657700
  spearman.pval spearman.ebp shapiro.stat shapiro.pval
1    0.47884669    0.4165712    0.9332133  0.031356622
2    0.65817864    0.4747819    0.9401584  0.051379824
3    0.56727867    0.4190165    0.9406454  0.053206026
4    0.06396674    0.2862117    0.9566950  0.169710037
5    0.23816781    0.4077764    0.9824013  0.823178934
6    0.22673574    0.4071157    0.9066853  0.005224356
  shapiro.ebp wgt.mean.ebp  best.pval  best.ebp
1  0.11546141    0.4346690 0.47884669 0.4082477
2  0.15068827    0.5133152 0.65817864 0.4156794
3  0.16181081    0.4524590 0.56727867 0.4118789
4  0.57744066    0.3496546 0.06762317 0.2730540
5  1.00000000    0.4389164 0.18372621 0.3707854
6  0.04496325    0.4221661 0.22673574 0.3862690
```

# 3 Microarray Data Analysis

Since multiple testing is widely used in microarray gene expression data, we are using microarray data analysis to illustrate our procedures. However our procedure can be used for other types of data analysis for multiple testing. The prerequisite is that the user needs to understand empirical Bayes probabilities, False Discovery Rate(FDR). Empirical Bayes probabilities are known also by local FDR; genes which empirical Bayes probabilities are less than a given cut off point are said to be expressed. As mentioned in Efron et al., empirical Bayes probabilities take multiple testing into consideration.

Let's give another example for k-group comparison scenario.

```
> library(HybridMTest)
> data(GroupComp.data)
```

```
> brain.express.set <- exprs(GroupComp.data)
> brain.pheno.data  <- pData(GroupComp.data)
>  brain.express.set[1:5, 1:8]

       Y 1      Y 2      Y 3      Y 4      Y 5      Y 6
1 3.328627 3.490428 7.808851 6.267770 7.451358 8.945424
2 3.214868 4.102643 9.073030 6.503689 7.369979 5.239628
3 7.376696 7.403122 4.238445 5.368776 6.971200 5.751302
4 8.722889 9.944269 8.942526 9.260197 4.681205 9.584824
5 7.684094 8.830104 4.182050 7.824086 8.039448 7.950749
       Y 7      Y 8
1 5.030438 5.403578
2 8.250280 5.843255
3 7.612831 6.227919
4 4.903050 7.205041
5 8.129559 6.463341

>   head(brain.pheno.data)

     grp
Y 1 grp1
Y 2 grp1
Y 3 grp1
Y 4 grp1
Y 5 grp1
Y 6 grp1

> test.specs<-cbind.data.frame(label=c("anova","kw","shapiro"),
+                              func.name=c("row.oneway.anova","row.kruskal.wallis","row.kgrp
+                              x=rep("grp",3),
+                              opts=rep("",3))
> ebp.def<-cbind.data.frame(wght=c("shapiro.ebp","(1-shapiro.ebp)"),
+                           mthd=c("anova.ebp","kw.ebp"))
> Kgrp.res<-hybrid.test(GroupComp.data,test.specs,ebp.def)
> head(Kgrp.res)

  anova.stat   anova.pval    anova.ebp   kw.stat
1  1.4044783 2.514820e-01 2.981217e-01  2.328232
2  0.8461295 4.328757e-01 1.000000e+00  1.808631
3 10.0989499 1.228159e-04 2.455672e-04 17.487984
4 16.7224426 8.554658e-07 1.717332e-06 24.371043
5 42.5883642 2.544631e-13 2.796332e-12 40.246356
6  1.0030499 3.713241e-01 5.425748e-01  1.949219
       kw.pval       kw.ebp shapiro.stat shapiro.pval
1 3.121985e-01 3.953826e-01    0.9455597  0.001534488
2 4.048188e-01 6.705833e-01    0.9535278  0.004483816
3 1.594162e-04 3.081395e-04    0.9884533  0.671483770
```

```
4 5.103819e-06 9.214304e-06    0.9749330  0.104208652
5 1.822279e-09 3.792653e-09    0.9688176  0.041173210
6 3.773396e-01 5.558066e-01    0.9349729  0.000402740
  shapiro.ebp wgt.mean.ebp    best.pval     best.ebp
1 0.009423449 3.944661e-01 3.121985e-01 3.145583e-01
2 0.011079522 6.742331e-01 4.048188e-01 6.705833e-01
3 1.000000000 2.455672e-04 1.228159e-04 1.041002e-04
4 0.086642691 8.564746e-06 5.103819e-06 9.214304e-06
5 0.071479712 3.521755e-09 1.822279e-09 3.792653e-09
6 0.002215387 5.557772e-01 3.773396e-01 5.020097e-01

>
```

# 4 Reference


1. Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: a pPractical and Powerful Approach to Multiple Testing. J.R. Statist. Soc. B, 57, 289-300.

2. Efron, B. (2004) Large-Scale Simultaneous Hypothesis Testing: the choice of a null hypothesis. JASA

3. Efron B., Tibshirani R., Story J.D and Tusher V. Empiricaal Bayes Analysis of a Microarray Experiment.2001, vol. 96, 1151-1160

4. Efron, B.,Tibhirani, R., 2002. Empirical Bayes Methods and False Discovery Rates for Microarrays. Wiley-Liss, Inc. 23,70-86

5. Pounds, S.,Rai, S., 2009. Assumption adequacy averaging as a concept for developing more robust methods for differential gene expression analysis. Elsevier 53, 1604-1612.

6. Pounds, S., Cheng, C., 2004. Improving false discovery rate estimation. Bioinformatics 20, 1737-1745.

7. Strimmer, K., 2008. Fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics 24, 1461-1462.