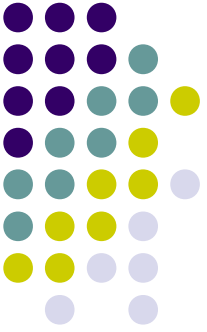
Lecture #7

Multiple testing



Outline

- Purpose
- Types of error
  - FDR
  - FWER
- Procedures
  - Single-step adjustments
    - *Bonferroni* method
    - *Sidak* method
    - *Westfall* and *Young minp* method
    - Permutation method
  - Step-down/step-up adjustments
    - *Holm's* method
    - FDR control

Multiple test purpose

- p-value review for two-sample test
  - When conducting a statistical test, under the null hypothesis (means are equal), the p-value (observed significance) is the chance of getting a test statistic more extreme than the observed test statistic
- When conducting a single statistical test, this probability is a good estimate
- However, when conducting multiple statistical tests, the likelihood of getting a significant p-value increases due to the shear number of independent tests
  - Effect of testing too many genes can result in high false positive rate (over-estimate of effect sizes)
  - For 100 t-tests, the number of significant results occurring by chance at $\alpha=0.05$ is 5
- As a result, we need a method to adjust the p-value or criteria to compensate for the multiple tests and make better estimates of false discovery rates

Types of Error

- *V = # Type I errors [false positives]*
- *T = # Type II errors [false negatives]*
- *m0 = # of true hypotheses*
- *R = # rejected hypotheses*

	# non-rejected hypothesis	# rejected hypothesis	
# true null hypotheses (non-differential genes)	*U*	*V – Type I error*	*m0*
# false null hypotheses (differential genes)	*T – Type II error*	*S*	*m1*
	m-R	*R*	*m*

Types of Error

- FWER
  - Family-wise error rate
    - The probability of at least one type I error (false positive)
    - P(V>0)
- FDR
  - False discovery rate
    - The expected proportion of type I errors (false positives) among the rejected hypotheses
    - E(V/R | R>0)

- Controlling the FDR is more important than controlling the FWER

Controlling Error Rates

- Assuming a set of significant differentially expressed genes

- Strong control of Type I error rate
  - An unknown fraction of genes might be differentially expressed
  - Control over Type I error rate under any combination of true and false null hypotheses

- Weak control of Type I error rate
  - Assume that all of the null hypotheses are true
  - Fail to reject all H_0 (no differential expression)

Bonferroni adjustment

- Single step procedure, only dependent upon number of tests (genes), *m*
- Adjusted $p_i = \min(mp_i, 1)$
- Gives strong control of FWER
- This is one of the more conservative methods
  - p-values are adjusted from significance
  - Sensitivity in differentially expressed genes can be an issue

Sidak single-step adjustment

- Single step procedure, only dependent upon number of tests (genes), *m*
- Adjusted $p_i = 1 - (1 - p_i)^m$
- Gives strong control of FWER
- This is less conservative than the *Bonferroni* adjustment

Westfall and Young minp adjustment

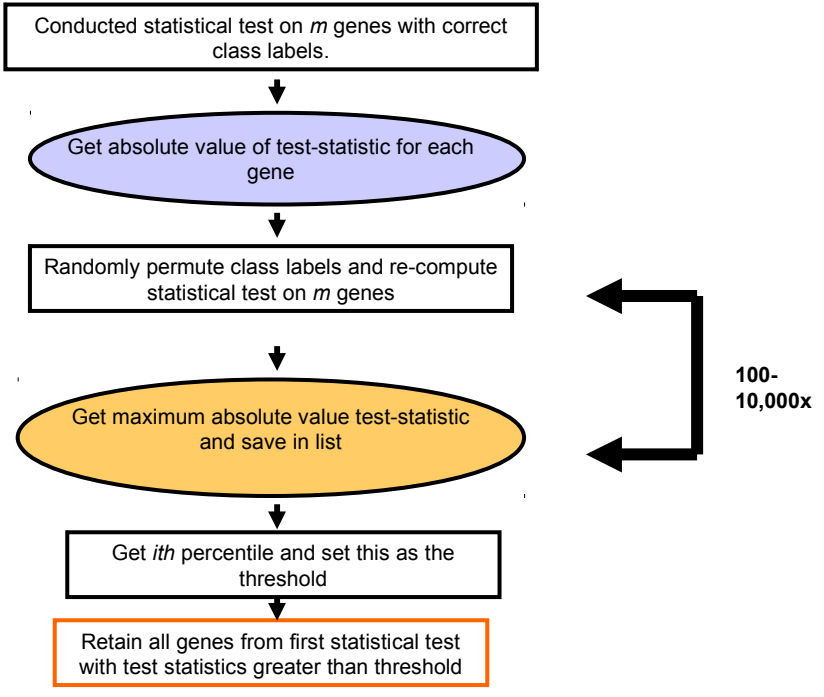
- Re-sampling procedure, dependent upon distribution of the p-values, P_k, where $1 \leq k \leq m$

  m is number of tests (genes)

- Adjusted $p_i = P(\min P_k \leq p_i \mid H)$

- This is more powerful than both the *Bonferroni and Sidak* adjustments

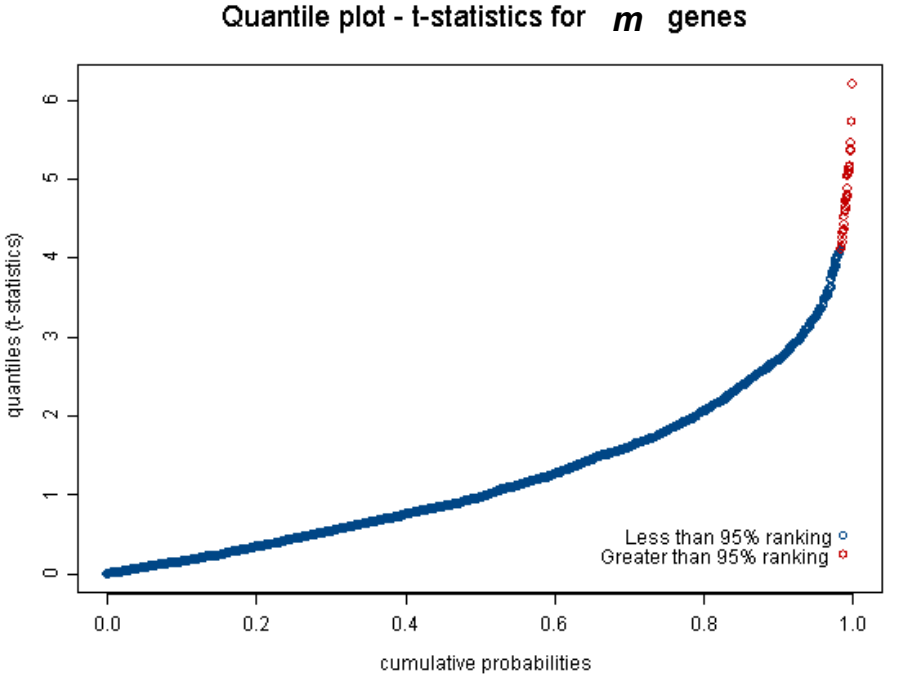
Permutation method



- Computationally expensive to permute data set multiple times
- Permutation iterations are dependent upon number of arrays (i.e. can reach convergence quickly with small studies)

Permutation method (cont.)

- Fraction of significant genes from permutation results

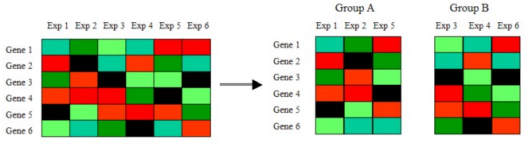


Significance Analysis of Microarray Data (SAM)[3]

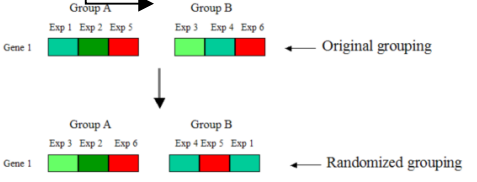
- A type of permutation method for detecting differentially expressed genes
- Can be applied to multiple experimental designs
  - time course, two-sample (paired or unpaired), one-sample, etc.
- For two-sample unpaired example, test statistic calculation is analogous to Student's *t*-test

$$s^2 = \frac{1}{m+n-2}\left(\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right). \qquad T(X,Y) = \frac{\bar{X} - \bar{Y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

- Then class structure is randomly shuffled



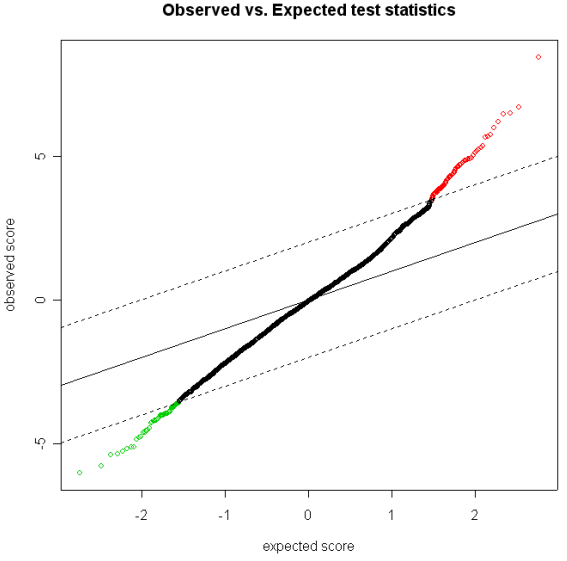
- This process is repeated numerous iterations and the average randomized test statistic is calculated for each gene

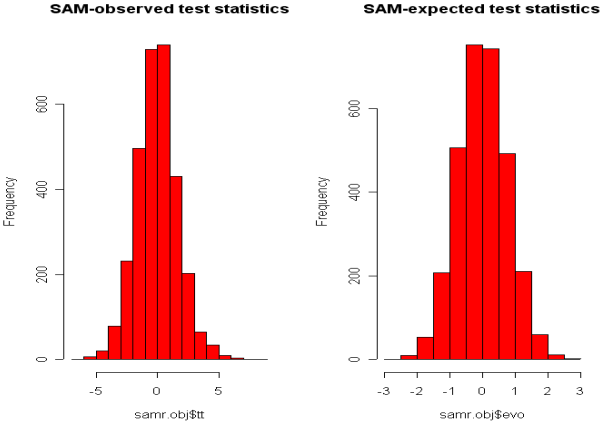


http://compbio.utmem.edu/MSCI814/Module11.htm

SAM determination of differentially expressed genes[3]

- A threshold ‘delta’ value is determined from the distributions of both expected and observed test statistics





- Then the observed test statistic values are plotted against the expected test statistic values
- Those genes with values outside of the specified delta range are considered differentially expressed (red and green points)

Holme's adjustment

- Step-down procedure requires a series of modifications to the parameters for each adjusted value
- Rank the p-values in ascending order, p1 < p2 < …p_m
- Adjusted p_i = max(k=1..j) {(m-k+1)p_k}
- Unlike the single step adjustments, the p-value is not multiplied by the same factor, *m*, but successively smaller factors (e.g. m-1, m-2, etc.)
- Strong control of FWER
- This is less conservative than the *Bonferroni* adjustment

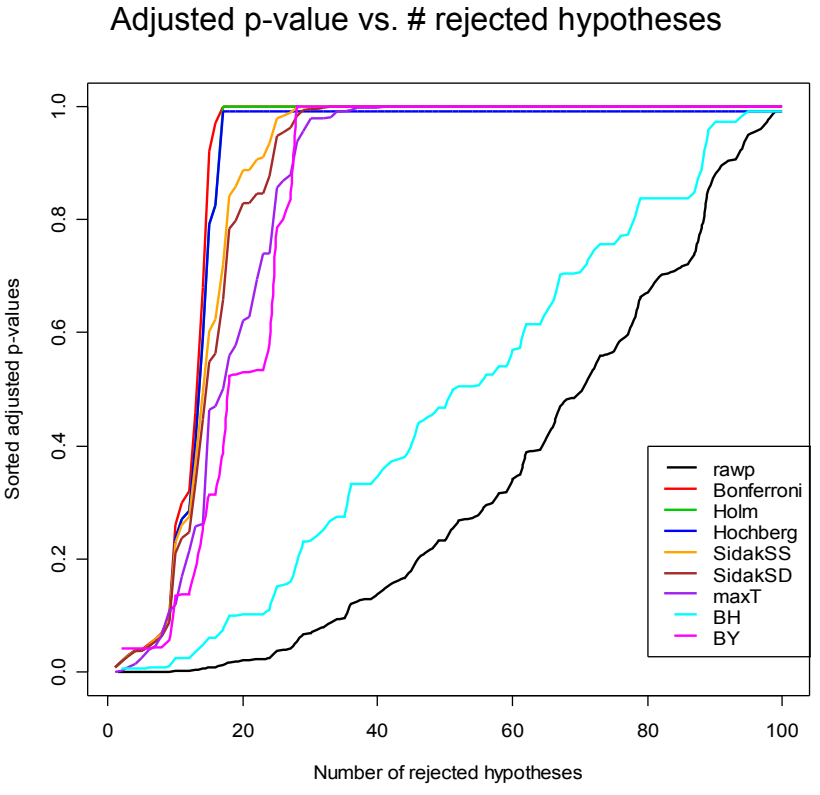
FDR control

- Step-down procedure requires a series of modifications to the parameters for each adjusted value

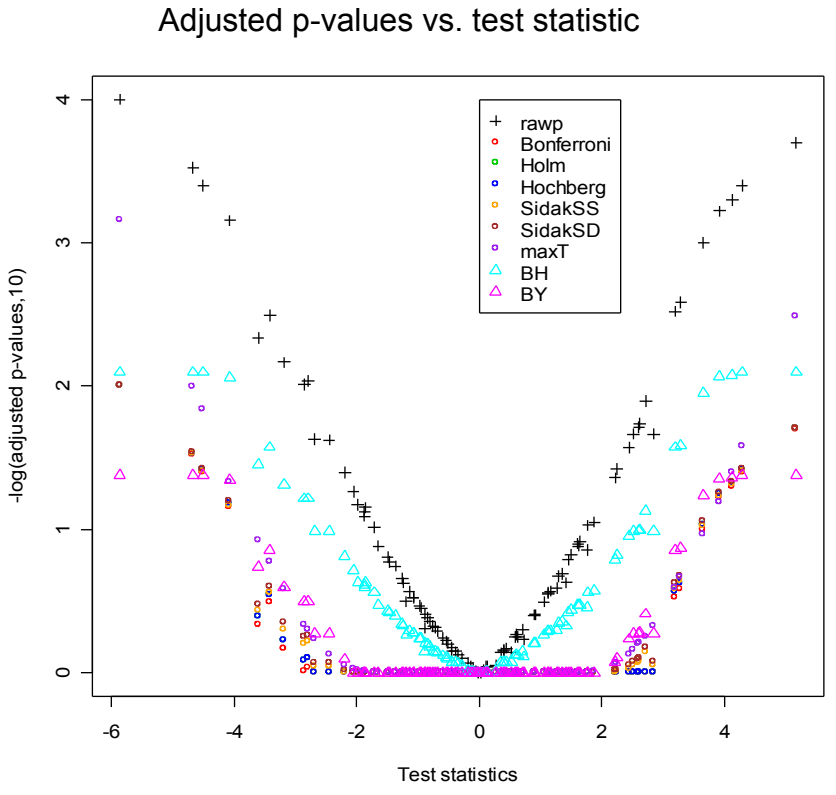
- Rank the p-values in ascending order, p1 < p2 < …p_m

- Adjusted p_i = min(k=i..m) {(mp_k/k)}

Multiple adjustment comparison



Multiple adjustment comparison



References

- 1) Xu R and Li X. (2003) A comparison of parametric vs. permutation methods with applications to general and temporal microarray gene expression data. *Bioinformatics*. **19**, 1284-1289.

- 2) Pan W. (2003) On the permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics*. **19**, 1333-1340.

- 3) http://compbio.utmem.edu/MSCI814/Module11.htm

R Code

```
library(Biobase);       library(annotate);      library(golubEsets);    library(multtest);
data(geneData);         data(golub);
dat1 <- geneData
dat2 <- golub[1:100,]
ann.dat2 <- golub.cl    # class labels

t.test.all.genes <- function(x,s1,s2) {
     x1 <- x[s1]
     x2 <- x[s2]
     x1 <- as.numeric(x1)
     x2 <- as.numeric(x2)
     t.out <- t.test(x1,x2, alternative="two.sided",var.equal=T)
     out <- as.numeric(t.out$p.value)
     return(out)
}
# s1 and s2 are dimensions of the two samples
# run function on each gene in the data frame
rawp <- apply(dat2,1,t.test.all.genes,s1=ann.dat2==0,s2=ann.dat2==1)

# apply multiple test correction using some permutation and step-down/up methods
library(multtest}

# another option for a t-test and non-parameteric tests, using minP adjustment method
# p-value results are sorted in ascending order (be aware)
resP<-mt.minP(dat2,ann.dat2,test="t",side="abs")$rawp

# apply multiple test correction using non-permuted methods
library(base)
p <- c(0.01,0.04,0.77,0.34)
p.cor <- p.adjust(p,method="holm")
```

R Code

```
# get first 100 genes of golub data with class labels
data(golub)
smallgd<-golub[1:100,]
classlabel<-golub.cl

# calculate multiple adjusted p-values with various methods
procs<-c("Bonferroni","Holm","Hochberg","SidakSS","SidakSD","BH","BY")
res2<-mt.rawp2adjp(rawp,procs)

# nice function to calculate the number of rejected hypotheses using Westfall and Young maxT adjustment
res<-mt.maxT(smallgd,classlabel)
mt.reject(cbind(res$rawp,res$adjp),seq(0,1,0.1))$r

# see mt.plot() for plots from the lecture

# SAM
dat <- golub
sam.ann <- classlabel+1 #the class labels must be 1 and 2 (not 0 and 1)
data=list(x=dat,y=sam.ann,
     geneid=as.character(1:nrow(x)),genenames=paste("g",as.character(1:nrow(x)),sep="") , logged2=F)
samr.obj<-samr(data,  resp.type="Two class unpaired", nperms=100)

# look at distributions of observed and expected test statistics
par(mfcol=c(1,2))
hist(samr.obj$tt,col='red',main='SAM-observed test statistics')
hist(samr.obj$evo,col='red',main='SAM-expected test statistics')

# plot the observed vs. expected genes using a delta of +/-2
delta=2
samr.plot(samr.obj,delta)
title(main='Observed vs. Expected test statistics')
```