

# **Multiple Testing Corrections and FDR Adjustments for Microarray Data Analysis**

Aixiang Jiang, MS  
Pengcheng Lu, MS

Vanderbilt Cancer Biostatistics Division Workshop

3/19/2010

# Multiple Comparison Issue

- Consider a one-way ANOVA with 3 groups( $g=3$ ); Each with 4, 5, and 5 observations separately, so  $N=14$ ;
- $g_1$ : 25,30,20,32;  
 $g_2$ : 30,33,29,40,36;  
 $g_3$ : 32,39,35,41,44;
- ANOVA Table

---

Source	SS	DF	MS	F	P
Between	292.11	2	146.056	6.21	0.016
Within	258.75	11	23.52		
Total	550.86	13			

---

# Multiple Comparison Issue(Cont.)!

- ANOVA can't report individual p values for each group's comparison. The point of multiple comparisons is to analyze a family of comparisons at once.
- The protected t-test is performed only when the overall F is significant;  
Formula:  $F_{comp12} = (\text{mean}(g1) - \text{mean}(g2))^2 / \text{MSE}(1/n1 + 1/n2)$   
Relationship between F and t:  $F = t^2$
- $F_{12} = 4.43$ ,  $F_{23} = 2.25$ ,  $F_{13} = 12.39$
- $P_{12} = 0.06$ ,  $p_{23} = 0.16$ ,  $p_{13} = 0.005$

# Adjusted p value

- The effects of multiple comparisons on the true error rates.

Number of comparisons	1	2	3	4	5
Nominal Type I error	5%	5%	5%	5%	5%
Actual overall Type I error	5%	12.2%	20.3%	28.6%	36.6%

- The p-value is the smallest level of significance that results in rejection of the null hypothesis.
- The adjusted p-value for a particular hypothesis within a collection of hypotheses is the smallest overall (experiment-wise or family-wise) significance level at which the particular hypothesis would be rejected.

Number of comparisons	1	2	3	4	5
Nominal Type I error	5%	?	?	?	?
Actual overall Type I error	5%	5%	5%	5%	5%

# Multiple Testings in Microarray Data

- Suppose one test of finding differentially expressed genes under different conditions has been conducted for one gene. There are  $m$  genes in one microarray experiment so  $m$  tests conducted.
- Let  $p_1, p_2, \dots, p_m$  denote the  $p$ -values corresponding to the  $m$  tests;
- Let  $H_{01}, H_{02}, \dots, H_{0m}$  denote the null hypotheses corresponding to the  $m$  tests.
- Let  $c$  denote a value that will serve as a cutoff for significance:
  - Reject  $H_{0i}$  if  $p_i \leq c$
- When  $m$  is large, false rejected are likely to occur.

# Classification of $m$ Hypothesis tests

	Accept Nulls	Reject Nulls	
	Declare Non-Sig.	Declare Sig.	
True Nulls	U	V	$m_0$
False Nulls	T	S	$m_1$
Total	$m-R$	R	$m$

# Classification of $m$ Hypothesis tests

	Accept Nulls	Reject Nulls	
	Declare Non-Sig.	Declare Sig.	
True Nulls	U	V	$m_0$
False Nulls	T	S	$m_1$
Total	$m-R$	R	$m$

U=number of true negatives

# Classification of $m$ Hypothesis tests

	Accept Nulls	Reject Nulls	
	Declare Non-Sig.	Declare Sig.	
True Nulls	U	V	$m_0$
False Nulls	T	S	$m_1$
Total	$m-R$	R	$m$

$V$  = number of false positives  
 = number of false discoveries  
 = number of type I errors

# Classification of $m$ Hypothesis tests

	Accept Nulls	Reject Nulls	
	Declare Non-Sig.	Declare Sig.	
True Nulls	U	V	$m_0$
False Nulls	T	S	$m_1$
Total	$m-R$	R	$m$

$T$ =number of type II errors

# Classification of $m$ Hypothesis tests

	Accept Nulls	Reject Nulls	
	Declare Non-Sig.	Declare Sig.	
True Nulls	U	V	$m_0$
False Nulls	T	S	$m_1$
Total	$m-R$	R	$m$

S=number of true positives  
 =number of true discoveries

# Classification of $m$ Hypothesis tests

	Accept Nulls	Reject Nulls	
	Declare Non-Sig.	Declare Sig.	
True Nulls	U	V	$m_0$
False Nulls	T	S	$m_1$
Total	$m-R$	R	$m$

$R$ =number of rejections of null hypotheses

# Classification of $m$ Hypothesis tests

	Accept Nulls	Reject Nulls	
	Declare Non-Sig.	Declare Sig.	
True Nulls	U	V	$m_0$
False Nulls	T	S	$m_1$
Total	$m-R$	R	$m$

$m-R$  = number of non-rejections  
= number of null hypotheses accepted

# Classification of m Hypothesis tests

	Accept Nulls	Reject Nulls	
	Declare Non-Sig.	Declare Sig.	
True Nulls	U	V	$m_0$
False Nulls	T	S	$m_1$
Total	$m-R$	R	m

Unobservable

Observable

# Family-wise Error Rate

- Family-wise error rate is defined as the probability of getting at least one false positive (Type I error) result. So
$$\text{FWER} = \text{Prob}(V > 0)!$$
- Controlling FWER: choosing the significance cut off  $c$  so that FWER is less than or equal to a desired level  $\alpha$ .
- The Bonferroni Method is the simplest way to control FWER.
- Choose  $c = \alpha/m$ . With this  $c$ , the FWER will be no larger than  $\alpha$  for any family of  $m$  tests.
- For example, suppose control  $\text{FWER} < 0.05$ , then comparing  $p$  values from a 3 family-wise tests with cut off  $c$ , where  $c = 0.05/3 = 0.017$ . If a  $p$  value less than 0.017 then claim significance.

# Holm's Method of Controlling FWER

- Holm's Method also controls FWER, but less conservative than the Bonferroni method.
- First order the  $m$   $p$ -values from smallest to largest.
- Find the largest integer  $k$  so that  $p_{(i)} \leq \alpha / (m-i+1)$  for all  $i=1, \dots, k$ .
- Set  $c = p_{(k)}$  (reject the nulls corresponding to the smallest  $k$   $p$ -values).
- If no such  $k$  exists, set  $c = 0$  (declare no significant test).

# An Example of Applying Holm's Method

- Suppose we get 6 p values from 6 pairwise comparison tests(how many groups?).

Test	1	2	3	4	5	6
p value	0.049	0.01	0.001	0.005	0.025	0.043

- Guess which test(s) will be significant if controlling FWER at level 0.05?
- $0.001 \leq 0.05 / (6 - 1 + 1) = 0.008$   $i=1$   
 $0.005 \leq 0.05 / (6 - 2 + 1) = 0.01$   $i=2$   
 $0.01 \leq 0.05 / (6 - 3 + 1) = 0.013$   $i=3$   
 $0.025 > 0.05 / (6 - 4 + 1) = 0.017$   $i=4$   
 $0.043 > 0.05 / (6 - 5 + 1) = 0.025$   $i=5$   
 $0.049 \leq 0.05 / (6 - 6 + 1) = 0.05$   $i=6$

# An Example of Applying Holm's Method

- 3 individual p values are significant when controlling FWER at level 0.05 which correspond to tests 3, 4 and 2.
- Cut off  $c = p_{(3)} = 0.01$ .
- In R, `p.adjust()` is the function aimed for p value adjustment.
- R script:

```
p<-c(0.049, 0.01, 0.001, 0.005, 0.025, 0.043)
adjustP<-p.adjust(p, method="holm")
adjustP
> adjustP
[1] 0.086 0.04 0.006 0.025 0.075 0.086
```

# False Discovery Rate

- FDR is another measure for the erroneous rejection of a number of true null hypotheses, thus it's also related to Type I error.
- FDR was introduced by Benjamini and Hochberg (1995)  
Definition:  $FDR = E(V/R)P(R>0)$ , and  $FDR=0$ , if  $R=0$
- Alternatively, define  $pFDR=E(V/R|R>0)$ . When  $m$  is large,  $P(R>=0)$  is approximately 1 and FDR is approximately equal to pFDR.
- FDR control is less strict than FWER control, so the FDR controlling procedures are more potentially powerful.
- The analysis of microarray data is a case where FDR control suffices, as its purpose is to exact genes that are potential candidates for further investigation, as long as the proportion is small.

# Benjamini and Hochberg's Method

- First order the  $m$   $p$ -values from smallest to largest.
- Find the largest integer  $k$  so that  $p_{(k)} \leq k \alpha / m$ .
- Set  $c = p_{(k)}$  (reject the nulls corresponding to the smallest  $k$   $p$ -values).
- If no such  $k$  exists, set  $c = 0$  (declare no significant test).

# Applying Benjamini and Hochberg's Method

- Suppose we get 6 p values from 6 pairwise comparison tests as before.

Test	1	2	3	4	5	6
p value	0.049	0.01	0.001	0.005	0.025	0.043

- Guess which test(s) will be significant if controlling FDR at level 0.05?
- $0.001 \leq 1 * 0.05 / 6 = 0.008$   $i=1$   
 $0.005 \leq 2 * 0.05 / 6 = 0.017$   $i=2$   
 $0.01 \leq 3 * 0.05 / 6 = 0.025$   $i=3$   
 $0.025 \leq 4 * 0.05 / 6 = 0.033$   $i=4$   
 $0.043 > 5 * 0.05 / 6 = 0.042$   $i=5$   
 $0.049 \leq 6 * 0.05 / 6 = 0.05$   $i=6$

# The BH Method

- The BH method would consider all these tests are significant when controlling FDR at level 0.05 even though  $0.043 > 0.042$ .
- Cut off  $c = p_{(6)} = 0.049$ .
- R script:

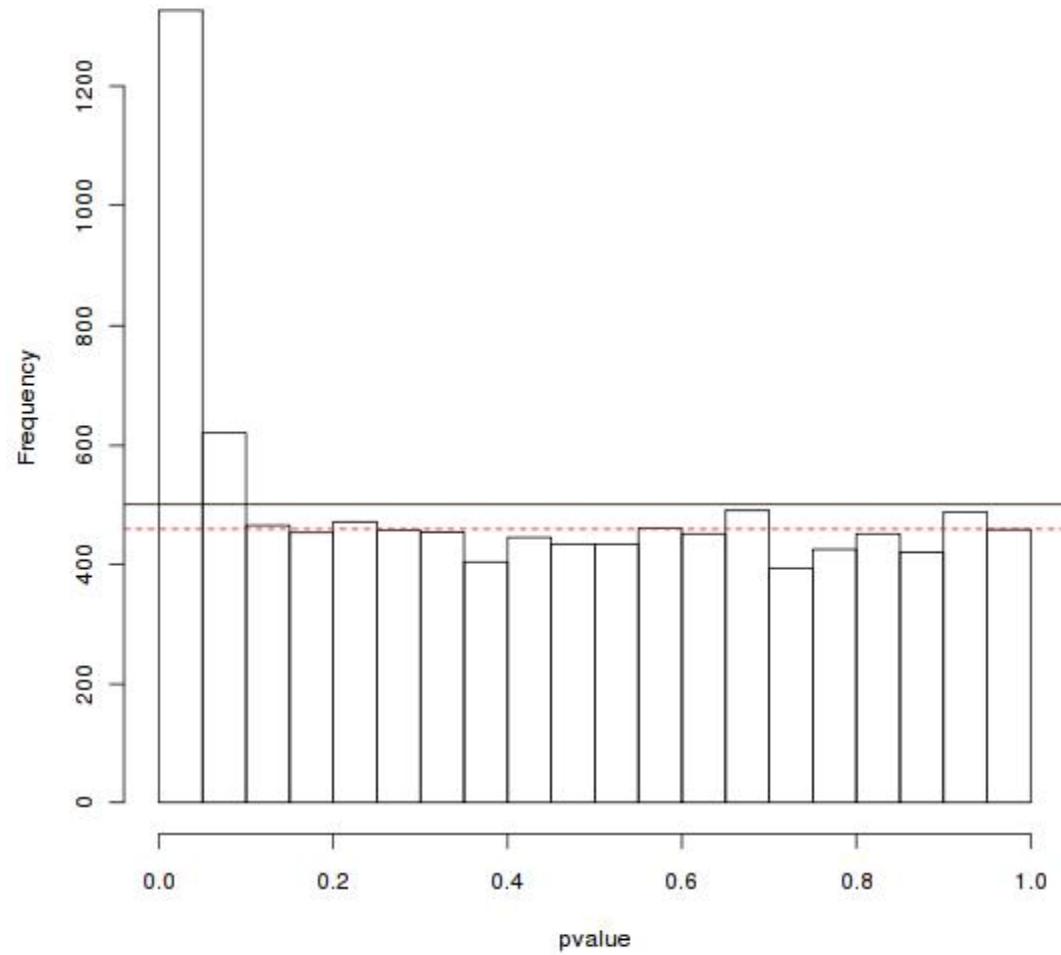
```
p<-c(0.049, 0.01, 0.001, 0.005, 0.025, 0.043)
adjustP<-p.adjust(p, method="BH")
adjustP
> adjustP
[1] 0.049 0.02 0.006 0.015 0.0375 0.049
```

## Other Methods for Estimating FDR

- The BH method tries to find the cut off  $c=p_{(k)}$  by finding the largest integer  $k$  such that  $p_{(k)} m / k \leq \alpha$ , other methods consider replacing  $m$  here with estimated  $m_0$  to find the cut off  $c=p_{(k)}$ .
- Methods of estimating  $m_0$  :
  - Lowest Slope: Schweder and Spjotvoll (1982), Hochberg and Benjamini (1990), Benjamini and Hochberg (2000).
  - Mixture of Uniform and Beta distributions: Allison et al. (2002), Pounds and Morris (2003).
  - $\lambda$  Threshold: Storey (2002), Storey and Tibshirani (2003)
  - Density Estimation: Langaas, Ferkingstad, Lindqvist (2004)

# Simulated $p$ -value Distribution

Histogram of pvalue



# $q$ -value Method

- Recall that a  $p$ -value for an individual test can be defined as the smallest significance level (tolerable type I error rate) for which we can reject the null the hypothesis.
- The  $q$ -value for one test in a family of tests is the smallest pFDR for which we can reject the null hypothesis for that one test and all others with smaller  $p$ -values.
- Can be called FDR adjusted  $p$ -value.
- A measure attached to each individual gene.

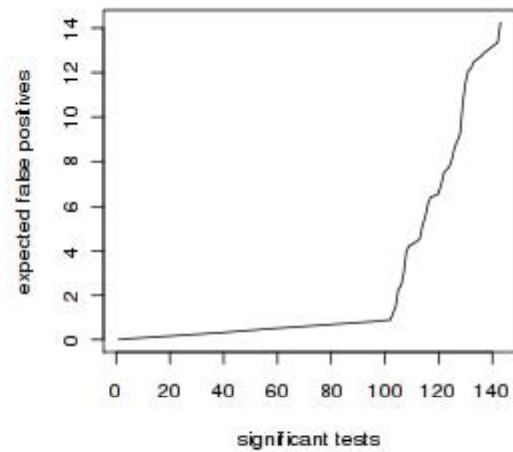
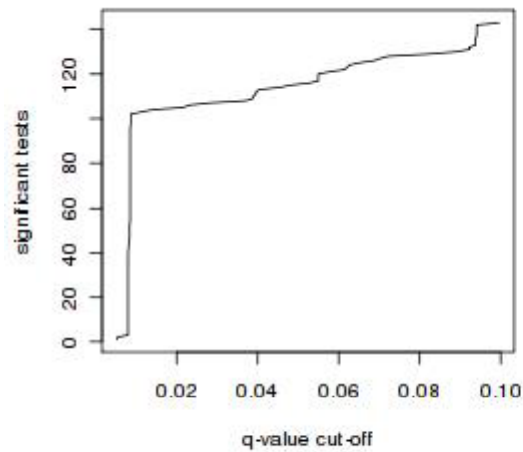
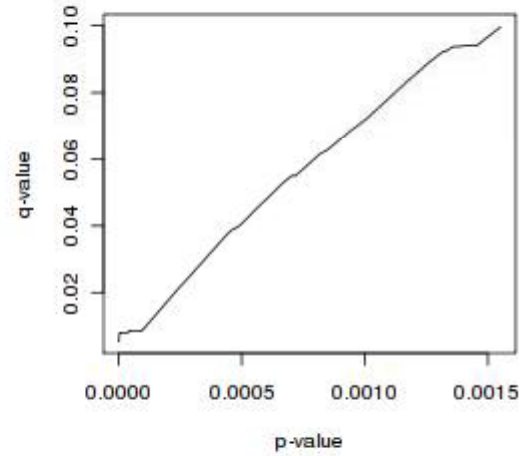
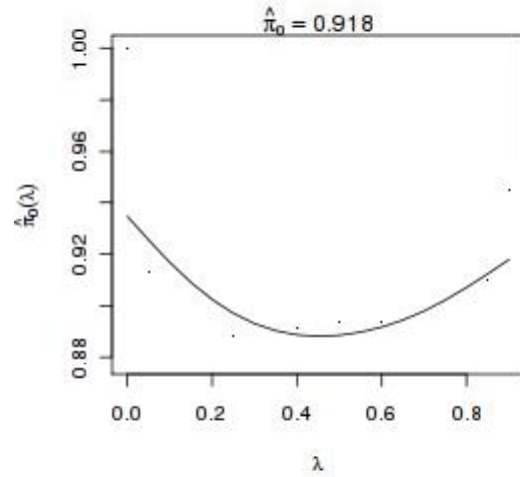
# Computation of $q$ -values

- Let  $q_{(i)}$  denote the  $q$ -value that corresponds to the  $i^{\text{th}}$  smallest  $p$ -value  $p_{(i)}$ .
- $q_{(i)} = \min \{ p_{(k)} m_0 / k : k = i, \dots, m \}$ .
- To produce a list of genes with estimated FDR  $\leq \alpha$ , include all genes with  $q$ -values  $\leq \alpha$ .
- R package “qvalue”:  
qobj<-qvalue(pvalue)  
qplot(qobj)  
qp<-data.frame(qobj\$pvalues, qobj\$qvalues)  
qp.sort<-qp[order(qp\$qobj.pvalues),]  
head(qp.sort, 20)

- `> head(qp.sort, 20)`

	qobj.pvalues	qobj.qvalues
734	5.949756e-07	0.005364775
79	1.168986e-06	0.005364775
11	3.499588e-06	0.007996216
4263	4.058007e-06	0.007996216
909	5.279558e-06	0.007996216
975	5.424583e-06	0.007996216
3882	6.253702e-06	0.007996216
6800	9.110427e-06	0.007996216
991	9.654950e-06	0.007996216
5301	1.194058e-05	0.007996216
36	1.289034e-05	0.007996216
343	1.301804e-05	0.007996216
807	1.375233e-05	0.007996216
670	1.413852e-05	0.007996216
5684	1.445377e-05	0.007996216
67	1.659658e-05	0.007996216
222	1.774918e-05	0.007996216
1027	1.829619e-05	0.007996216
66	1.844333e-05	0.007996216
8800	1.949099e-05	0.007996216

# Graphs for $q$ -value Method



# Global FDR vs. local fdr

- Global FDR: For a collection of simultaneous hypothesis tests, FDR is the expected proportion of type I errors made using a given rejection rule. It assigns the same FDR to a set of genes in tail area.
  - > Advantage: easier to estimate.
- Local fdr:  $fdr(z) \equiv Pr\{\text{null}|z\} = p_0 f_0(z)/f(z)$ ,
  - where  $f_0(z)$  = the density of  $Z$  for unaffected genes
  - $f(z)$  = the density of  $Z$  for all genes
  - > Advantage: more precise

Ref: Efron, B. et al. (2001) J. Am. Stat. Soc., 96, 1151–1160.

# Estimate global false discovery rate in R

Package	Type of FDR	Input Data	Package Authors
fdrtool	fdr and Fdr simultaneously.	p-values, z-scores, correlations, and t-scores.	K. Strimmer O. Muralidharan, with many suggestions from B. Efron
mixfdr	fdr and Fdr simultaneously.	z-scores.	S. Pounds
BUM / SPLOSH	fdr and Fdr simultaneously.	p-values.	P. Broberg
SAGx	fdr and Fdr simultaneously.	p-values.	A. Dabney and J. D. Storey
qvalue	frequentist Fdr.	p-values.	M. Guedj and G. Nuel K. S. Pollard, Y. Ge, S. Taylor, S. Dudoit
nFDR	frequentist Fdr.	p-values.	C. Dalmaso R. Tibshirani, G. Chu, T. Hastie, Balasubramanian Narasimhan
multtest	frequentist Fdr.	p-values.	
LBE	frequentist Fdr.	p-values.	
samr	frequentist Fdr.	expression data	
p.adjust in stats	FDR	p-values.	Gordon K. Smyth

# Estimate local false discovery rate in R

Package	Type of FDR	Input Data	Package Authors
fdrtool	fdr and Fdr simultaneously.	p-values, z-scores, correlations, and t-scores.	K. Strimmer O. Muralidharan, with many suggestions from B. Efron
mixfdr BUM / SPLOSH	fdr and Fdr simultaneously.	z-scores.	S. Pounds
SAGx	fdr and Fdr simultaneously.	p-values.	P. Broberg
locfdr	local fdr.	z-scores.	B. Efron, B. B. Turnbull and B. Narasimhan
nomi	local fdr.	z-scores.	G. McLachlan, R. W. Bean and L. B.-T. Jones
LocalFDR	local fdr.	p-values.	J. Aubert
kerfdr	local fdr.	p-values.	M. Guedj and G. Nuel
twilight	local fdr.	p-values.	S. Scheid
localFDR	local fdr.	p-values.	J.G. Liao
Cofdr	local fdr.	expression data	A. Jiang, P. Lu, W. Wu, Y. Shyr

link for COfdr: <http://biostat.mc.vanderbilt.edu/wiki/pub/Main/AixiangJiang/SampleRcodecontrolOnlyfdr111308.r>

# Efron's locfdr: definition

Assumption: Mixture density

$$f(z) = p_0 f_0(z) + p_1 f_1(z), \text{ where } p_1 = 1 - p_0$$

(accepted by most of FDR or fdr estimation methods based on test statistic)

The Bayes posterior probability that a case is null given  $z$ , by definition the local false discovery rate, is

$$\text{fdr}(z) \equiv \Pr\{\text{null}|z\} = p_0 f_0(z) / f(z)$$

Assume  $p_0 = 1$  (should be 0.90 or greater)  $\rightarrow$

$$\text{fdr}(z) = f_0(z) / f(z)$$

## Efron's locfdr: procedure

- Estimate  $f(z)$  from observed  $z$ -values by natural spline or multinomial fit to histogram counts
- Estimate empirical null distribution:
  - 1). Fit  $f(z)$  to the histogram counts by Poission regression
  - 2). Obtain  $\delta_0$  and  $\sigma_0$  from  $f(z)$  assuming  $f_0(z) \sim N(\delta_0, \sigma_0^2)$

$$\delta_0 = \arg \max\{f(z)\} \text{ and } \sigma_0 = \left[ -\frac{d^2}{dz^2} \log f(z) \right]_{\delta_0}^{-\frac{1}{2}},$$

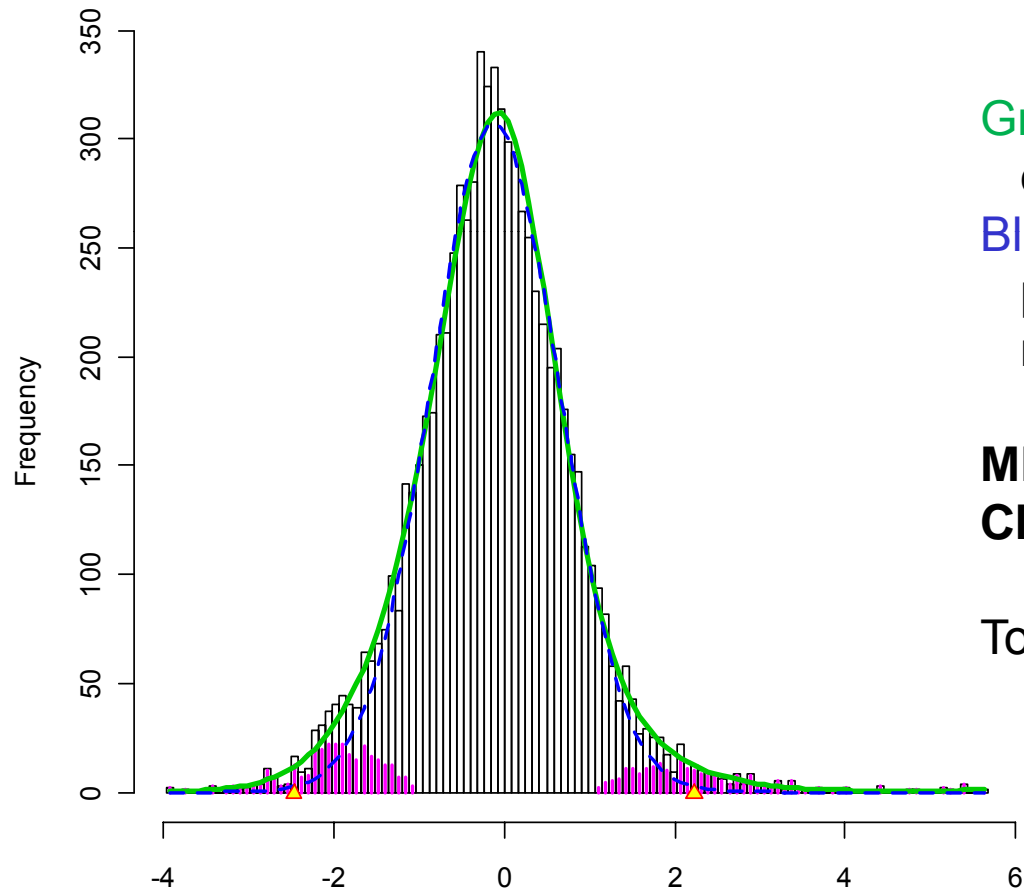
# Efron's locfdr: example

- Data: hivdata with A vector containing 7680 z-values  
-> Reference van't Wout, et. al., Cellular gene expression upon human immuno-deficiency virus type 1 infection of CD4+-T-Cell lines, Journal of Virology 77, 1392-1402.

- R code

```
library(locfdr)
## HIV data example
data(hivdata)
w <- locfdr(hivdata)
```

# locfdr: output



**Green solid line** -> the spline-based estimate of the mixture density  $f$ .  
**Blue dashed line** -> the null subdensity  $p_0 f_0$ , estimated by default by MLE maximum likelihood (nulltype=1)

**MLE**: maximum likelihood estimation  
**CME**: central matching estimation

To get fdr values -> `w$fdr`

MLE: delta: -0.116 sigma: 0.754 p0: 0.934  
CME: delta: -0.083 sigma: 0.712 p0: 0.895

## Challenge of most of FDR and fdr estimation methods

- How to estimate  $p_0$  ?
  - Which is the proportional of genes or features that are from null distribution.
- How do we know if the estimation is good?

## Our approach: CO<sub>f</sub>dr

- Avoid estimation of  $p_0$
- Use control group only to build up statistic null distribution
- Use Bootstrap to increase range of null distribution (compared to permutation)
- Formula:  $\text{fdr}(z) = f_0(z)/f(z)$

## Cofdr: data of an example

- Data: AML(Acute Myelogenous Leukemia) and ALL (Acute Lymphocytic Leukemia) expression data.  
-> Ref: Golub,T.R. et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537

# Cofdr basic R function:gett

#function to calculate Welch t values

```
gett=function(vdata,n1,n2){
```

```
#assuming vdata is a vector
```

```
#assuming the first n1 elements are from one sample,
```

```
#and the last n2 elements are from another group
```

```
#with un-equal variance assumption
```

```
...}
```

```
##### use gett, assume we have "dat" and "info" already, and both of them
```

```
##### have same array names
```

```
controls = rownames(subset(info, info$group="ALL"))
```

```
cases = rownames(subset(info, info$group="AML"))
```

```
## use ALL as control for an example
```

```
tvalues = apply(cbind(dat[,controls],dat[,cases]),1,gett, n1=length(controls),  
               n2=length(cases))
```

# Cofdr basic R function: getBoost

## a function to get t values of bootstrap sampling on control groups

```
getBootst=function(data,B){
```

```
#note: data is a data set of control or normal samples
```

```
#B is the number of bootstrap samples' t values
```

```
#bootstrap sampling, get t values
```

```
.....}
```

##### use getBoost, assume we have “dat” and “info” already, and both of them

##### have same array names

```
controls = rownames(subset(info, info$group="ALL"))
```

## use ALL as control for an example

```
b=1000
```

```
bst=getBootst(data=dat[,controls],B=b)
```

# Cofdr basic R function: getfdrs

##a function to estimate Cofdr: control only fdr

```
getfdrs=function(obst, bootst, bre=120) {
```

```
##obst: a vector of observed welch t values, for estimation of f(t)
```

```
##bootst: a matrix of bootstrap welch t values, for estimation of f0(t)
```

```
##bre: number of breaks for histogram
```

```
##f(t) and f0(t) are estimated based on combined distribution: across genes
```

```
...}
```

##### use getfdrs

```
fdrs=getfdrs(obst=tvalues,bootst=bst)
```

## Compare Cofdr to other methods: use ALL as control

fdr cutoff	OCplus fdr	BHfdr	twilight fdr	COfdr	locfdr
$1.00 \times 10^{-6}$	3	0	0	10	0
$5.00 \times 10^{-6}$	4	1	0	20	0
$1.00 \times 10^{-5}$	5	2	0	23	0
$5.00 \times 10^{-5}$	9	4	0	38	0
$1.00 \times 10^{-4}$	10	12	0	46	0
$5.00 \times 10^{-4}$	29	47	0	93	0
0.001	38	95	0	124	0
0.005	87	252	332	204	0
0.01	125	341	455	265	2
0.05	449	762	789	478	19

	OCplus fdr	BHfdr	twilight fdr	COfdr	locfdr
Training dataset					
ALL → ALL	27	27	27	27	27
AML → ALL	0	0	0	0	0
ALL → AML	0	0	0	0	0
AML → AML	11	11	11	11	11
Accuracy	100.00%	100.00%	100.00%	100.00%	100.00%
Testing dataset					
ALL → ALL	20	20	20	20	20
AML → ALL	0	0	0	0	0
ALL → AML	1	2	2	0	2
AML → AML	13	12	12	14	12
Accuracy	97.06%	94.12%	94.12%	100.00%	94.12%

ALL → ALL indicates ALL samples classified as ALL,  
 AML → ALL indicates AML samples classified as ALL, etc

## Compare Cofdr to other methods: use AML as control

fdr cutoff	OCplus fdr	BHfdr	twilight fdr	COfdr	locfdr
1.00 X 10 <sup>-6</sup>	0	0	0	0	0
5.00 X 10 <sup>-6</sup>	1	1	0	1	0
1.00 X 10 <sup>-5</sup>	3	2	0	1	0
5.00 X 10 <sup>-5</sup>	4	4	0	2	0
1.00 X 10 <sup>-4</sup>	6	12	0	5	0
5.00 X 10 <sup>-4</sup>	13	47	0	24	0
0.001	25	95	0	39	0
0.005	52	252	332	140	0
0.01	97	341	455	189	2
0.05	339	762	789	436	19

	OCplus fdr	BHfdr	twilight fdr	COfdr	locfdr
Training dataset					
ALL → ALL	27	27	27	27	27
AML → ALL	0	0	0	0	0
ALL → AML	0	0	0	0	0
AML → AML	11	11	11	11	11
Accuracy	100.00%	100.00%	100.00%	100.00%	100.00%
Testing dataset					
ALL → ALL	20	20	20	20	20
AML → ALL	0	0	0	0	0
ALL → AML	1	2	2	0	2
AML → AML	13	12	12	14	12
Accuracy	97.06%	94.12%	94.12%	100.00%	94.12%

*ALL* → *ALL* indicates ALL samples classified as ALL,  
*AML* → *ALL* indicates AML samples classified as ALL, etc

# Summary

- Concept of false discovery rate
- Difference between global FDR and local fdr
- Several ways to estimate global false discovery rate in R, e.g. p.adjust and qvalue.
- Several ways to estimate local false discovery rate in R, e.g. locfdr and COfdr.

# Acknowledge

- Yu Shyr, PhD
- William Wu, PhD
- Lynne Berry. PhD

Thank you very much!