Statistics for Human Genetics and Molecular Biology
Lecture 26: Differential Expression Analysis

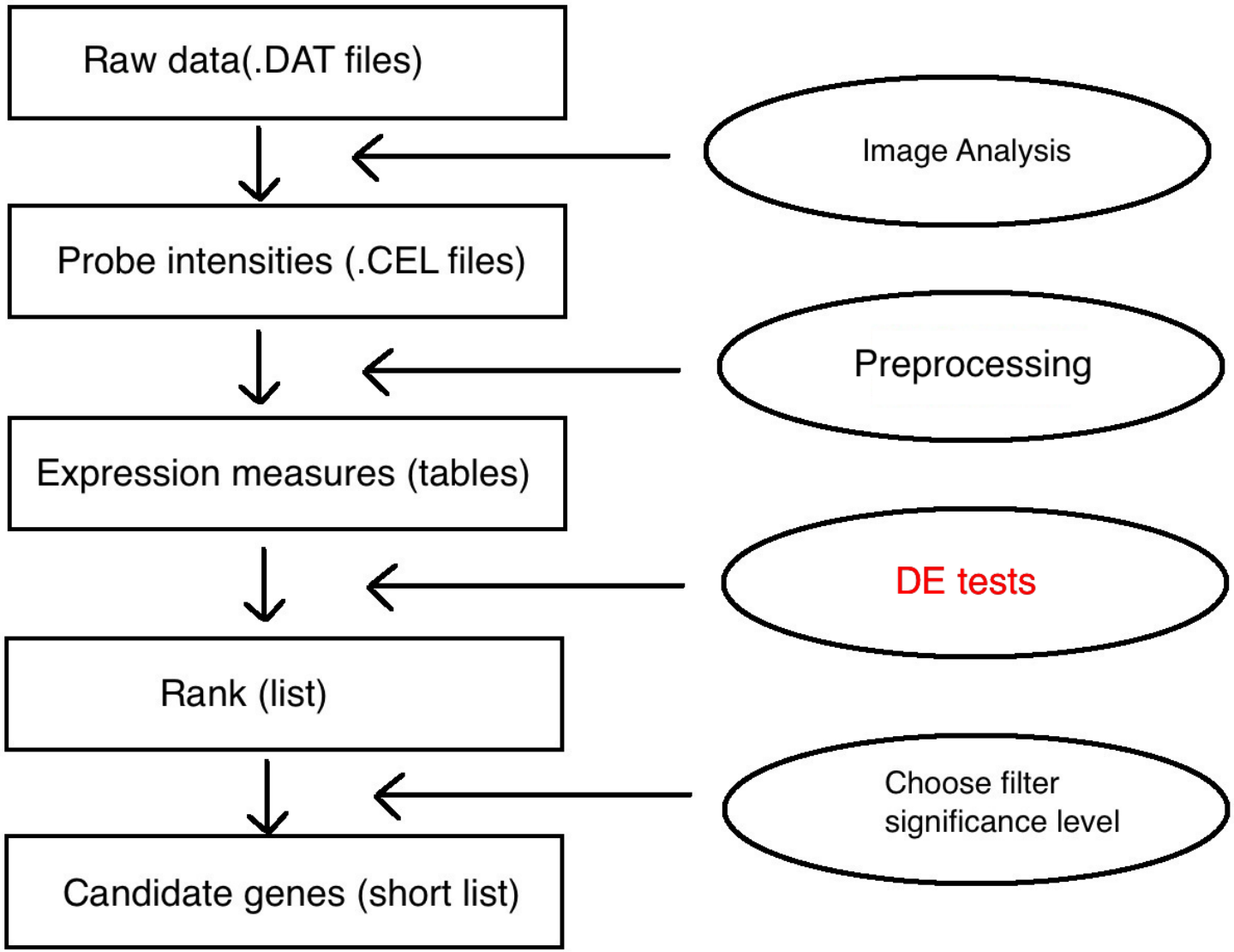
Dr. Yen-Yi Ho (yho@umn.edu)

Nov 6, 2015

Objectives of Lecture 26

- Simple Differential Expression
- Advanced Differential Expression

Analysis Flow Chart



Goal of Differential Expression (DE) Test

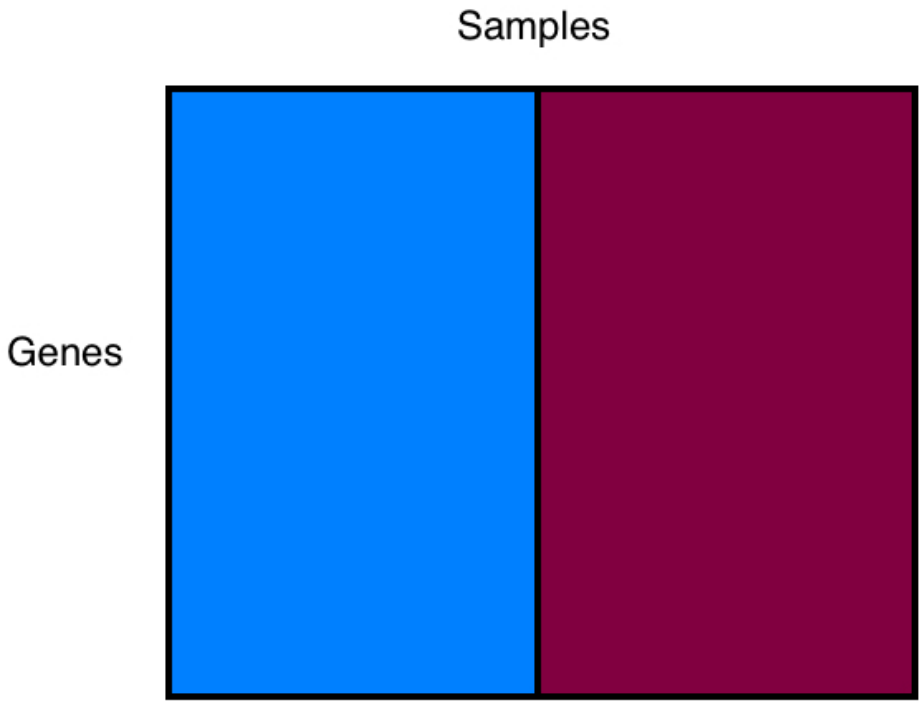
- Goal: Find genes that are expressed differently between conditions.
  - Assign a score for each gene to represent its statistical significance of being different.
  - Rank the genes according to the score.
  - Find a proper threshold for the score for DE.
- Naive Solution:
  - Hypothesis testing (t-tests, anova, linear regression model etc) to get p values as scores
  - use 0.05 as cut off

Example

```
>source("http://www.bioconductor.org/biocLite.R")
>biocLite("Biobase")
>biocLite("genefilter")
>biocLite("ALL")
>library("Biobase")
>library("genefilter")
>library("ALL")
>data("ALL")
>bcell<-grep("^B", as.character(ALL$BT))
>moltyp<-which(as.character(ALL$mol.biol) %in%
c("NEG", "BCR/ABL"))
>ALL_bcrneg<-ALL[, intersect(bcell, moltyp)]
>ALL_bcrneg$mol.biol<-factor(ALL_bcrneg$mol.biol)
```

Simple Differential Expression in Two Populations

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$



Simple Differential Expression in Two Populations

```
>ALLsfilt <- ALL_bcrneg[sds>=sh, ]
>dim(exprs(ALLsfilt))
>table(ALLsfilt$mol.biol)
>tt <- rowttests(ALLsfilt, "mol.biol")
>names(tt)
>ALLsrest <- ALL_bcrneg[sds<sh, ]
>ttrest <- rowttests(ALLsrest, "mol.biol")
>hist(tt$p.value, breaks=50, col="mistyrose",
xlab="p-value", main="Retained")
>hist(ttrest$p.value, breaks=50, col="lightblue",
xlab="p-value", main="Removed")
```

Simple Differential Expression in Two Populations

	statistic	dm	p.value
1636_g_at	9.26	1.10	3.76e-14
39730_at	8.69	1.15	4.79e-13
1635_at	7.28	1.20	2.45e-10
1674_at	6.90	1.43	1.28e-09
40504_at	6.57	1.18	5.27e-09
37015_at	6.19	1.03	2.74e-08
40202_at	6.18	1.78	2.79e-08
32434_at	5.78	1.68	1.54e-07
37027_at	5.65	1.35	2.60e-07
39837_s_at	5.50	0.48	4.74e-07

Potential Problems

- Hypothesis testing:
  - Sample sizes are usually small which lead to unstable test results.
- When data are not normal, p values are not accurate
- Multiple comparison problem: Bonferroni vs. False discovery rate (FDR)

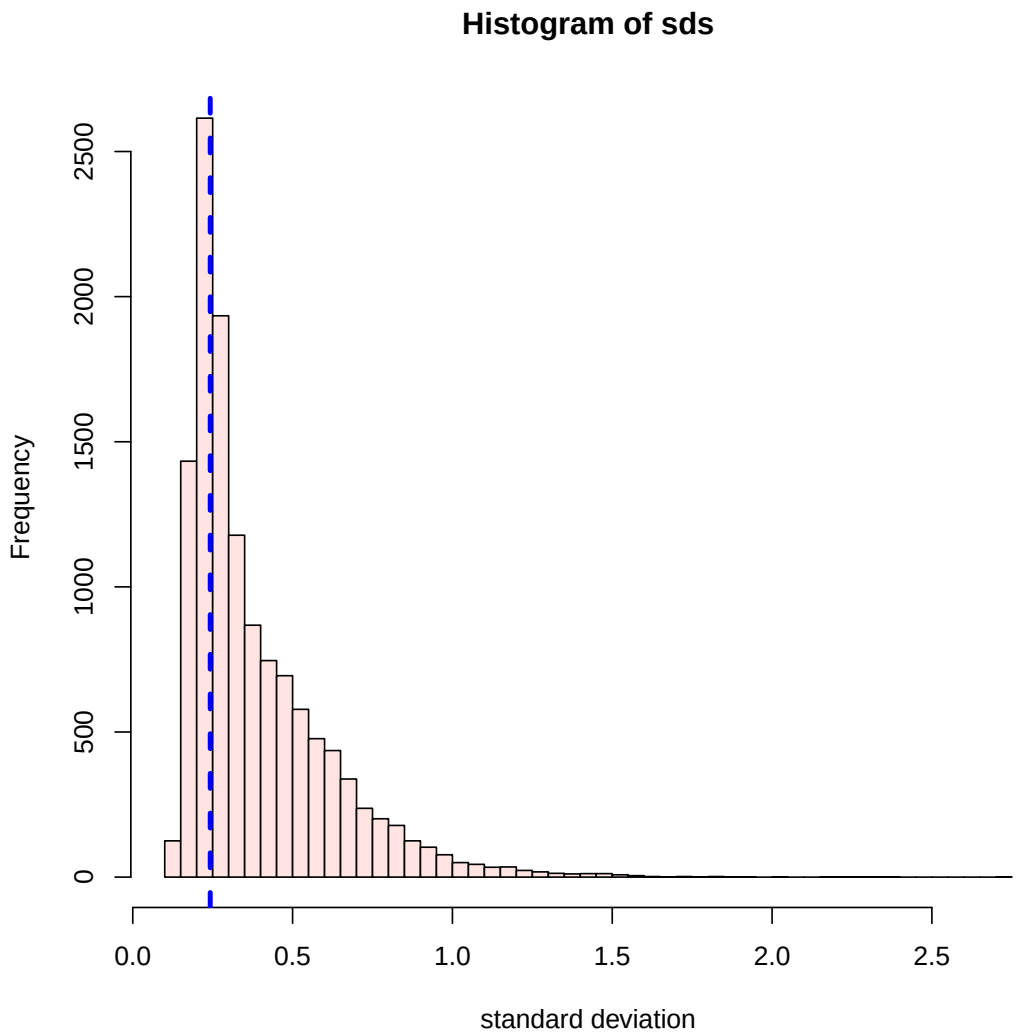
Nonspecific Filtering

- Reduced number of hypothesis testings
- Remove probe sets with low variability
  - control probes: "AFFX"
  - probe sets with low sensitivity to detect expression
  - non differentially expressed genes
- Caution: when number of samples is low in some conditions, might remove differentially expressed genes.

Nonspecific Filtering

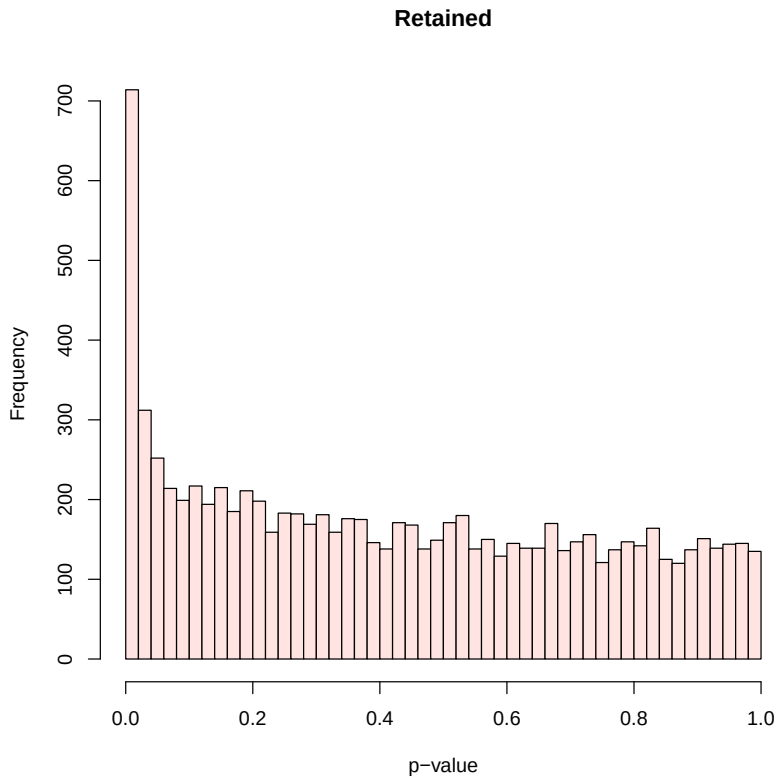
```
>library("genefilter")
>sds = rowSds(exprs(ALL_bcrneg))
>sh = shorth(sds)
>sh
>hist(sds, breaks=50, col="mistyrose", xlab="standard
deviation")
>abline(v=sh, col="blue", lwd=3, lty=2)
```

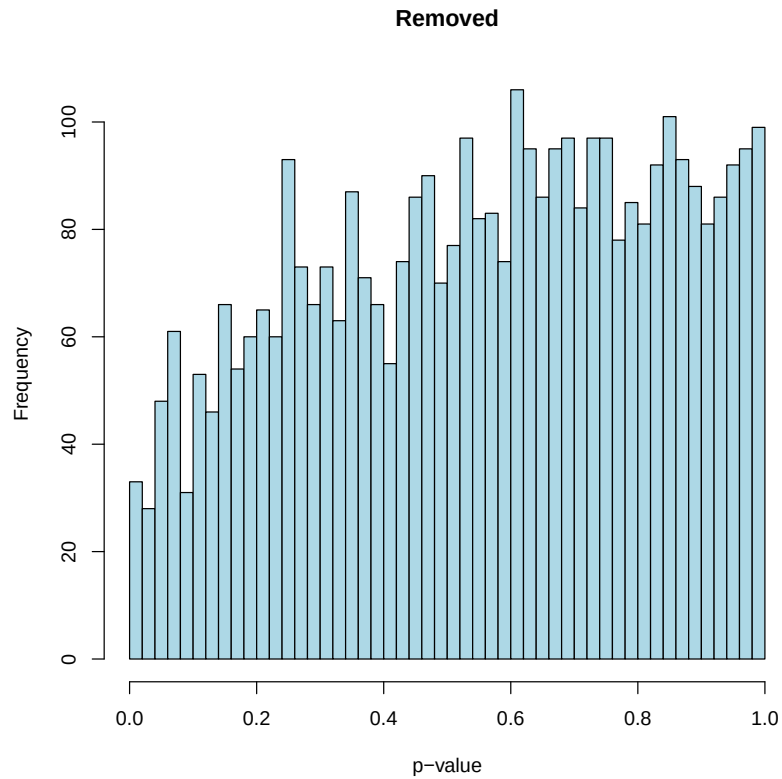
Nonspecific Filtering



Simple Differential Expression in Two Populations

```
>ALLsfilt <- ALL_bcrneg[sds>=sh, ]
>dim(exprs(ALLsfilt))
>table(ALLsfilt$mol.biol)
>tt <- rowttests(ALLsfilt, "mol.biol")
>names(tt)
>ALLsrest <- ALL_bcrneg[sds<sh, ]
>ttrest <- rowttests(ALLsrest, "mol.biol")
>hist(tt$p.value, breaks=50, col="mistyrose",
xlab="p-value", main="Retained")
>hist(ttrest$p.value, breaks=50, col="lightblue",
xlab="p-value", main="Removed")
```

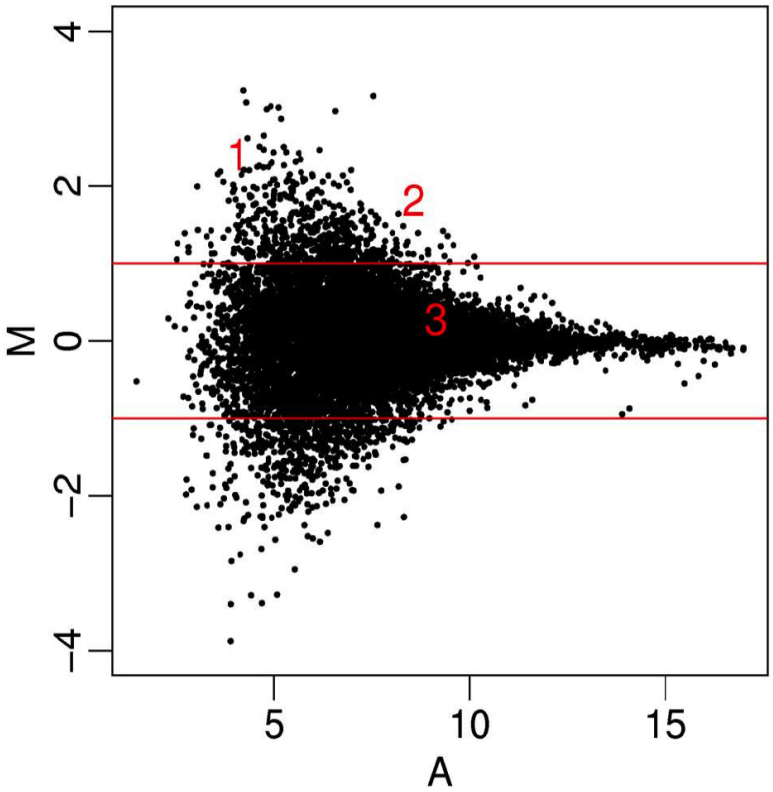

Retained
Frequency
0
100
200
300
400
500
600
700
0.0
0.2
0.4
0.6
0.8
1.0
p−value

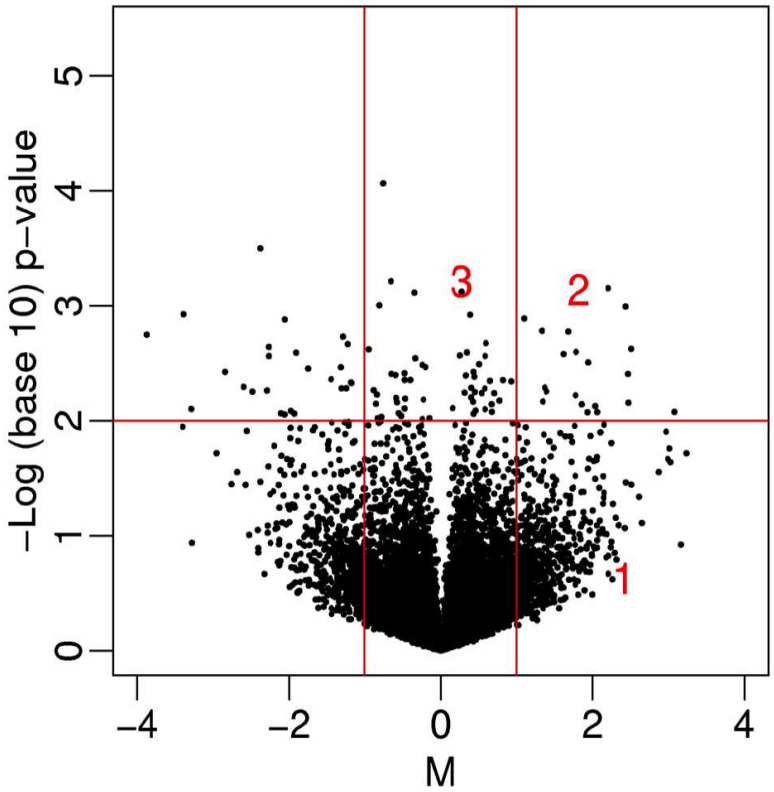

Removed
Frequency
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
p−value

Volcano Plot

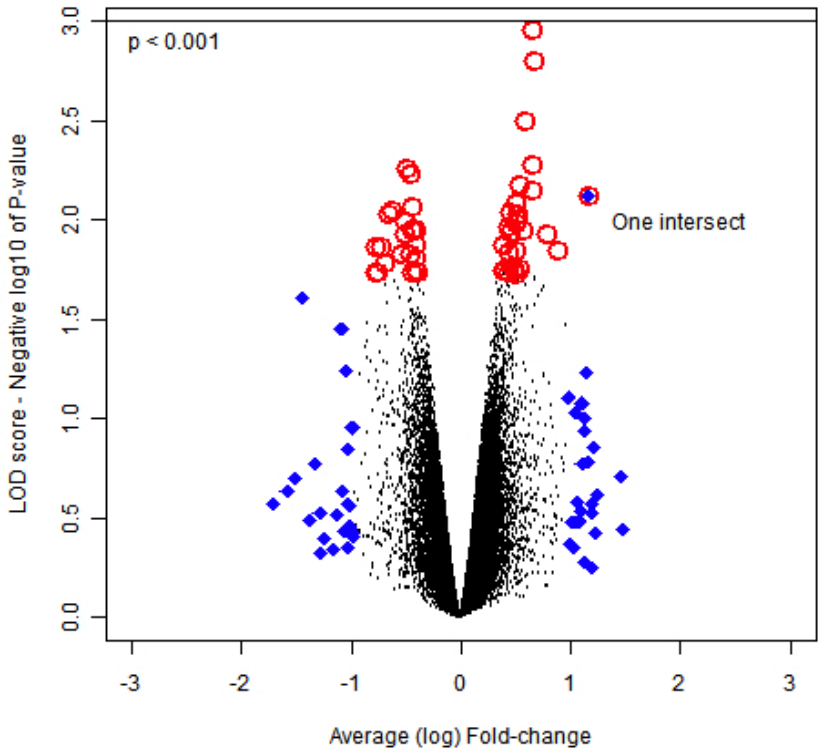
- A diagnostic plot to visualize the test results
- Scatter plot of statistical significance (log p values) versus biological significance (log fold-changes)
- Ideally the two should agree with each other

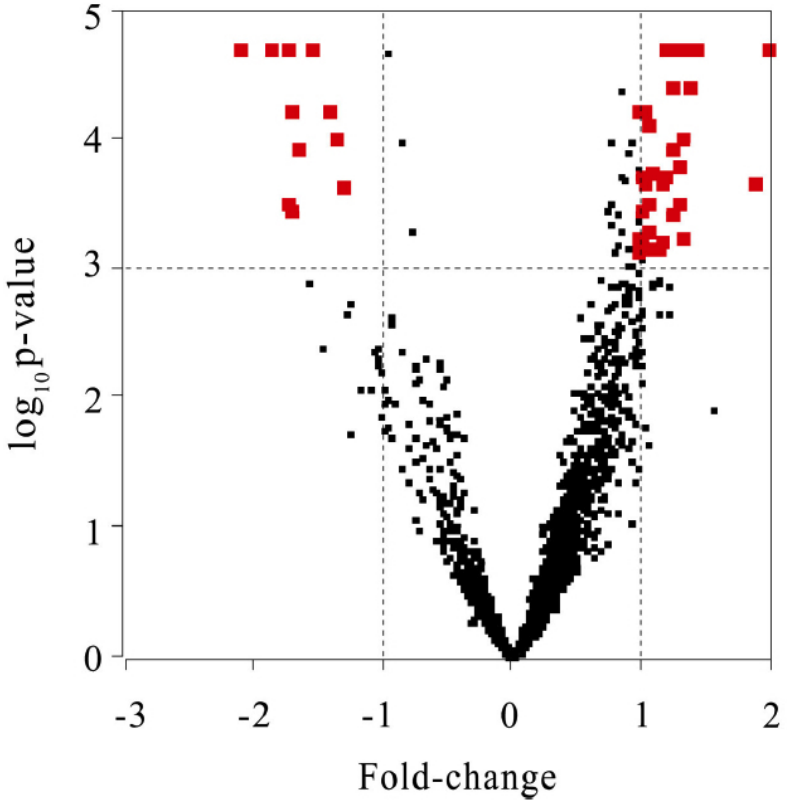
MA and Volcano Plots





Volcano Plots: Bad Versus Good





When sample size is small, SD estimates in t-test are unstable.

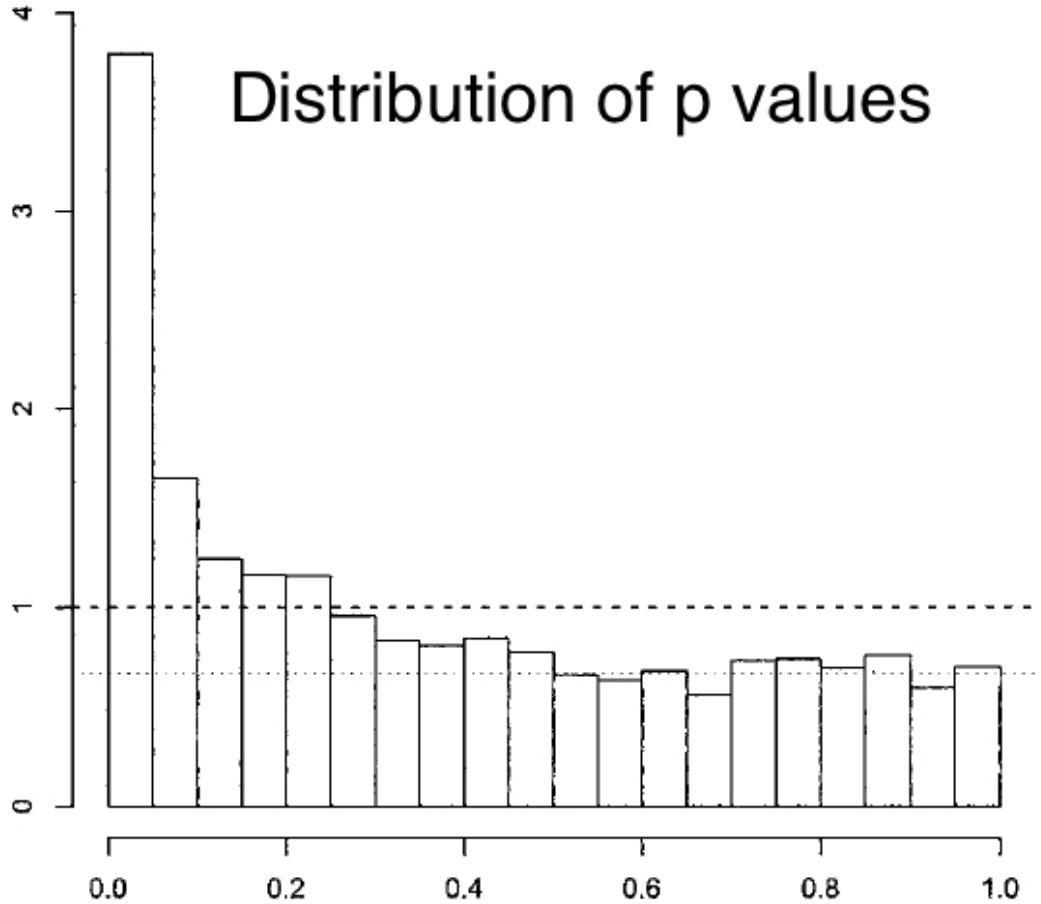
Volcano Plot

```
>plot(tt$dm, -log10(tt$p.value), pch=".", xlab =
expression(mean log[2] fold change), ylab =
expression(-log[10](p)))
```

Simple Differential Expression in Two Populations

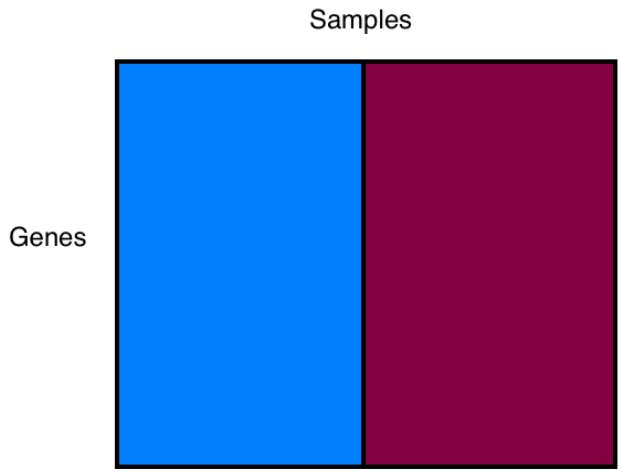
	statistic	dm	p.value
1636_g_at	9.26	1.10	3.76e-14
39730_at	8.69	1.15	4.79e-13
1635_at	7.28	1.20	2.45e-10
1674_at	6.90	1.43	1.28e-09
40504_at	6.57	1.18	5.27e-09
37015_at	6.19	1.03	2.74e-08
40202_at	6.18	1.78	2.79e-08
32434_at	5.78	1.68	1.54e-07
37027_at	5.65	1.35	2.60e-07
39837_s_at	5.50	0.48	4.74e-07

Multiple Testing Correction: False Discovery Rate (FDR)



Simple Differential Expression in Two Populations

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$



When sample size is small t-test has less power to detect differences and

SD estimates are unstable.

Exercise: Homework 8

1. Download CEL files from GSE18088 at gene expression omnibus (http://www.ncbi.nlm.nih.gov/geo/)
2. Perform simple differential expression analysis for the data.

Hint:

```
library(oligo)
library(siggenes)
library(limma)
library(pd.hg.u133.plus.2)
library(hgu133plus2.db)
```