STAT588/BIOL588: Genomic Data Science
Lecture 14: Differential Expression Analysis

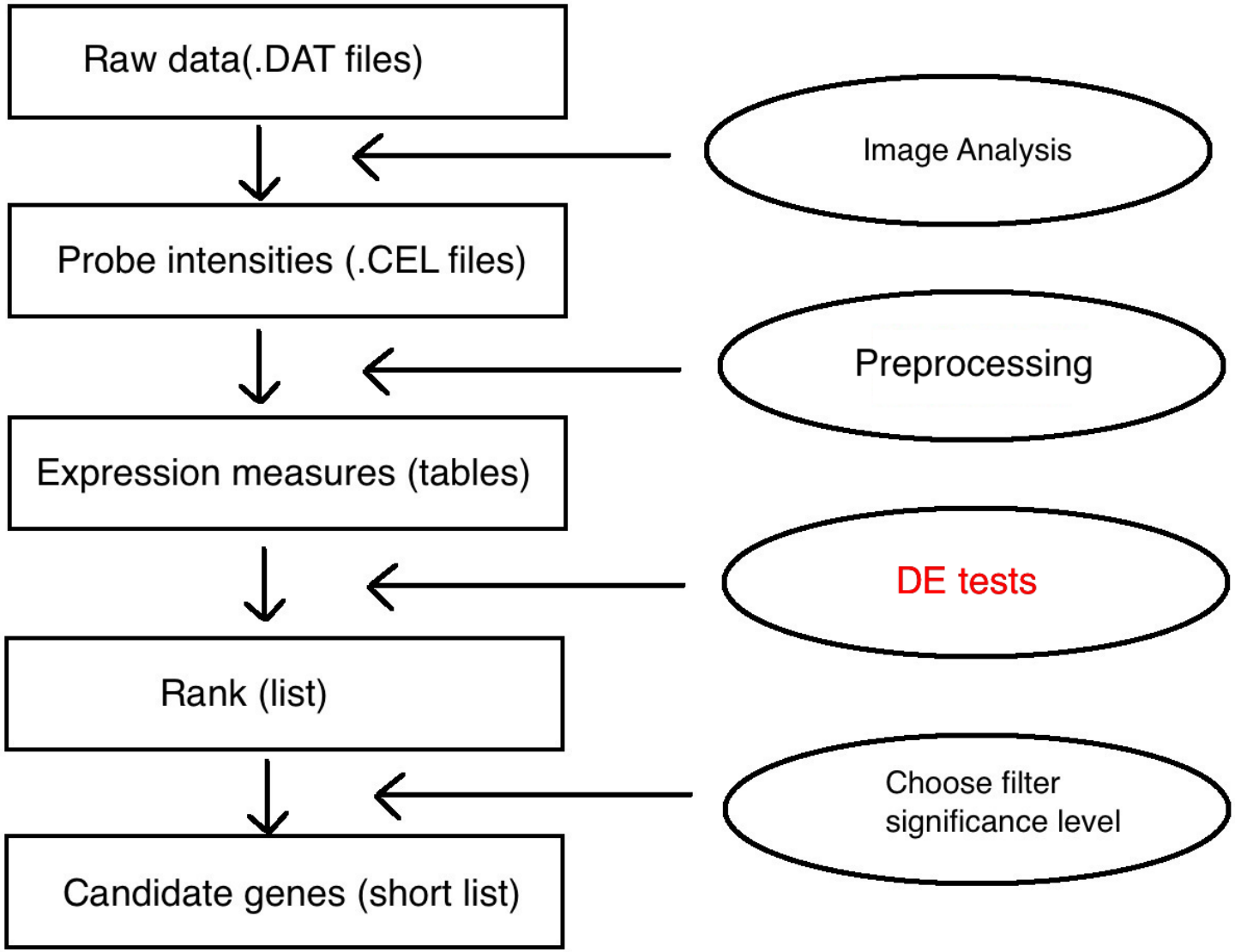
Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

STAT599/BIOL599: Genomic Data Science

Objectives of Lecture 14

- Simple Differential Expression
- Advanced Differential Expression

Analysis Flow Chart



Goal of Differential Expression (DE) Test

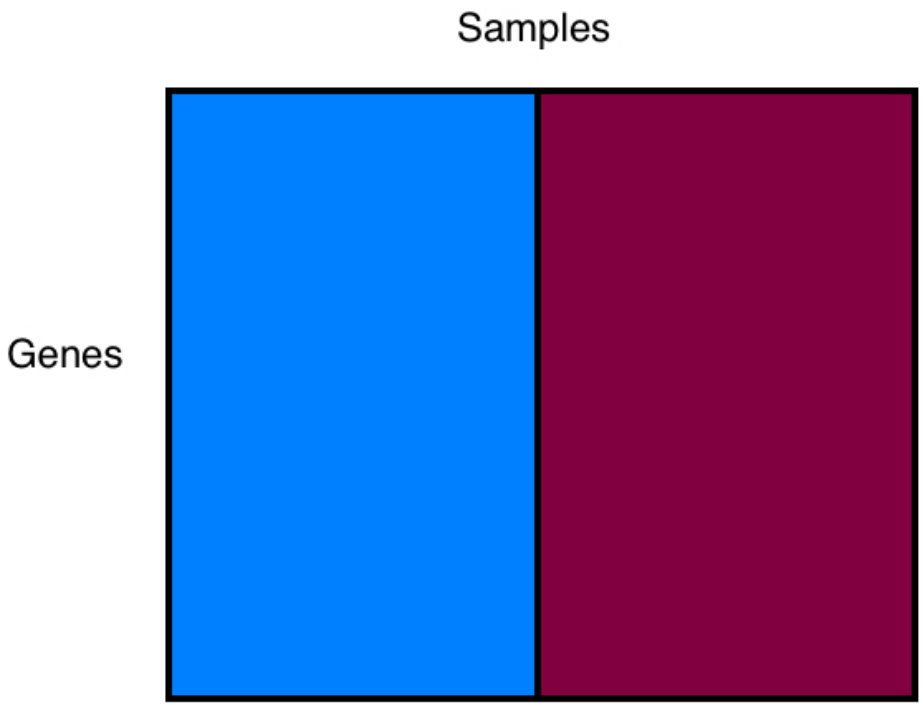
- Goal: Find genes that are expressed differently between conditions.
  - Assign a score for each gene to represent its statistical significance of being different.
  - Rank the genes according to the score.
  - Find a proper threshold for the score for DE.
- Naive Solution:
  - Hypothesis testing (t-tests,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

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



Exercise: Simple Differential Expression in Two Populations

Perform t-test for every probe set in the ALL dataset and present top genes in this table.

	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

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)

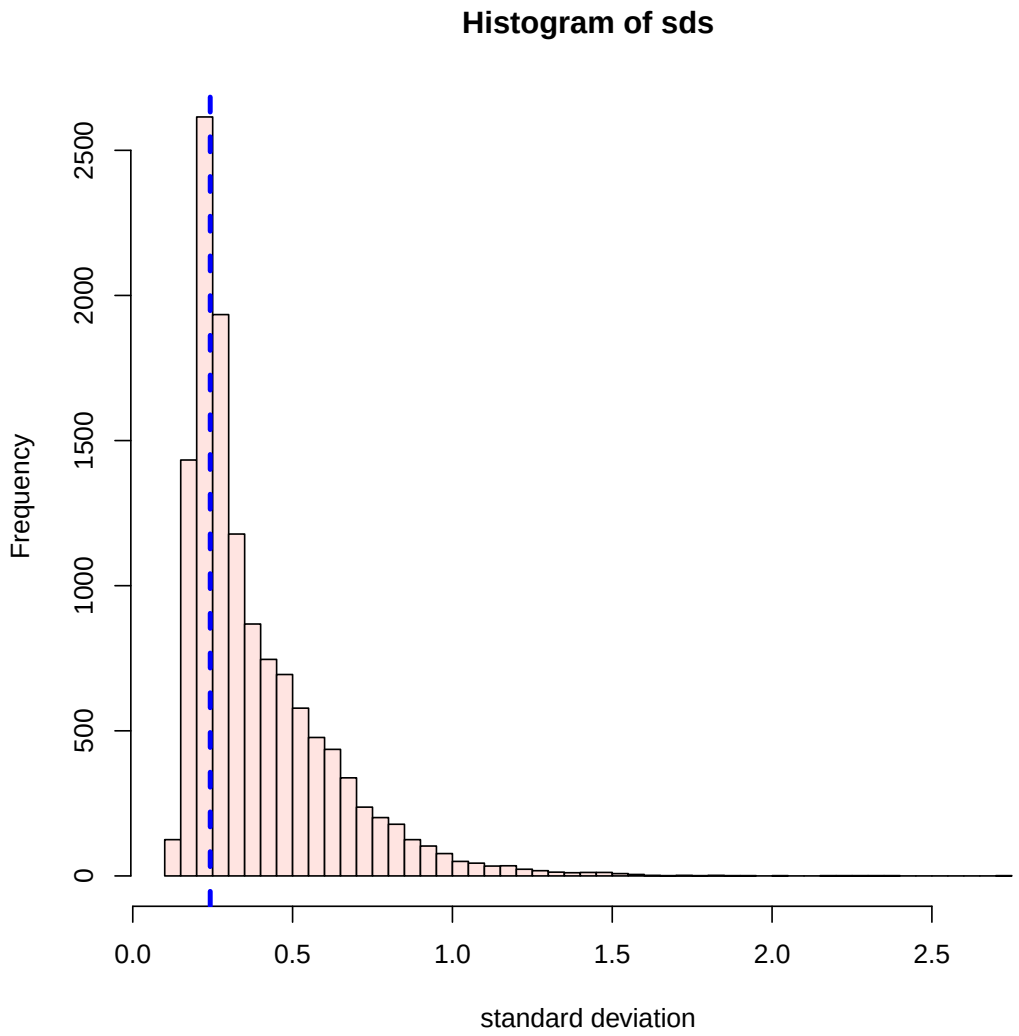
Nonspecific Filtering

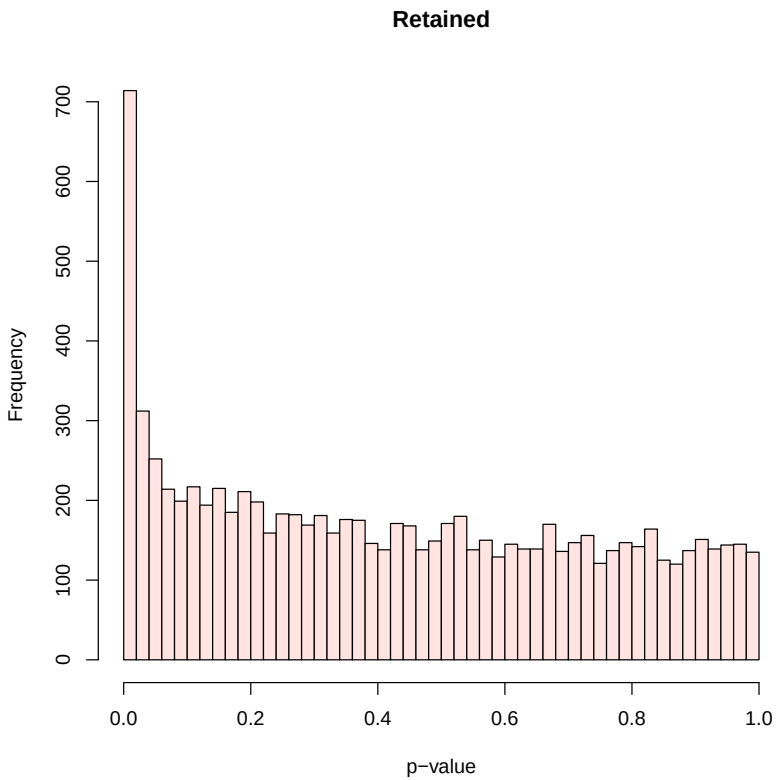
- Reduced number of hypothesis testings
- Remove probe sets with low variability
  - control probes: "AFFX"
  - probe sets with low sensitivity to detect expression
  - non differentially expressed genes
- Caution: when number of samples is low in some conditions, might remove differentially expressed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

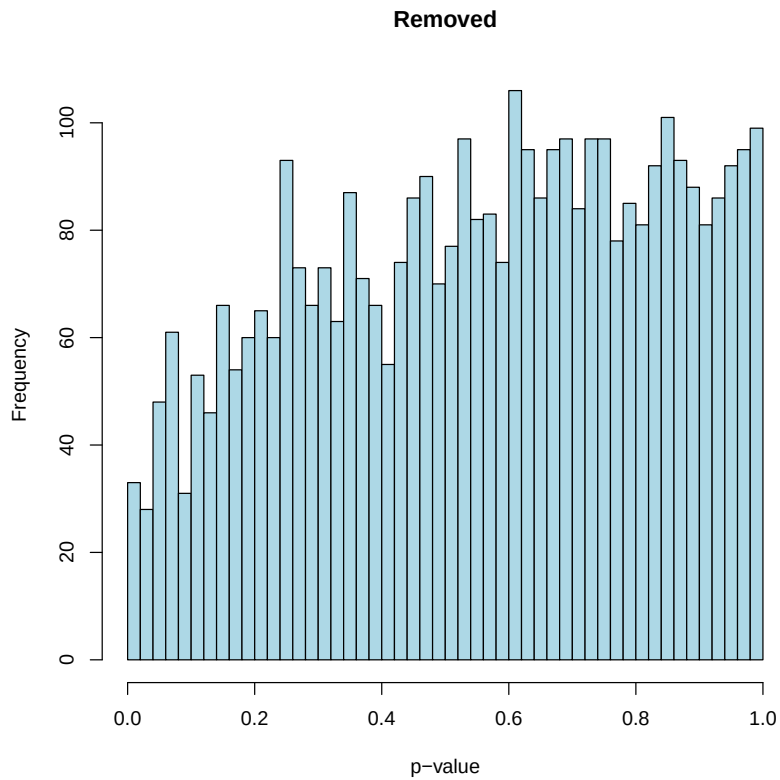
Nonspecific Filtering

```
>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



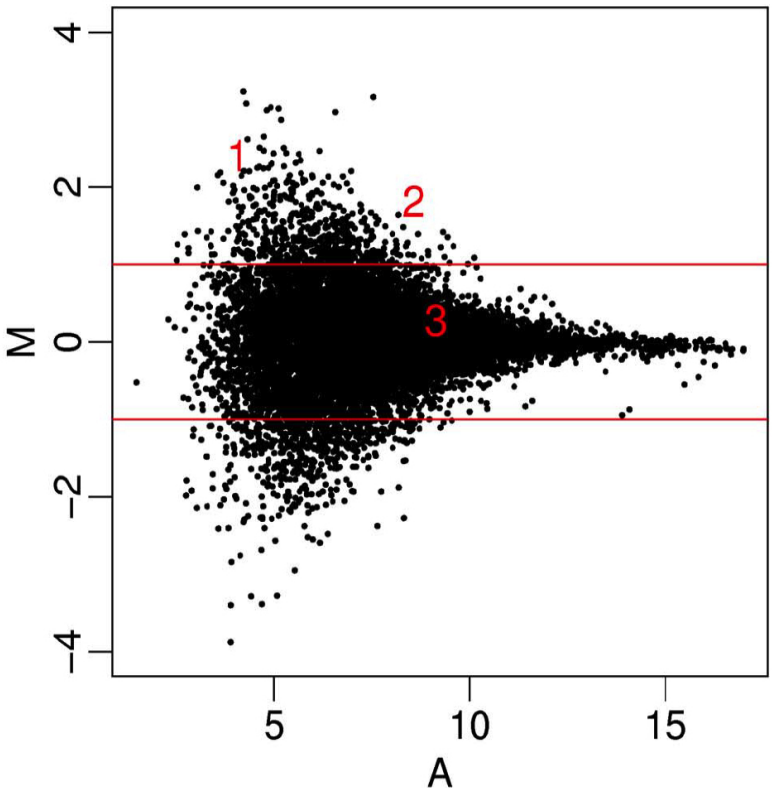

Retained
Frequency
0
100
200
300
400
500
600
700
0.0
0.2
0.4
0.6
0.8
1.0
p−value

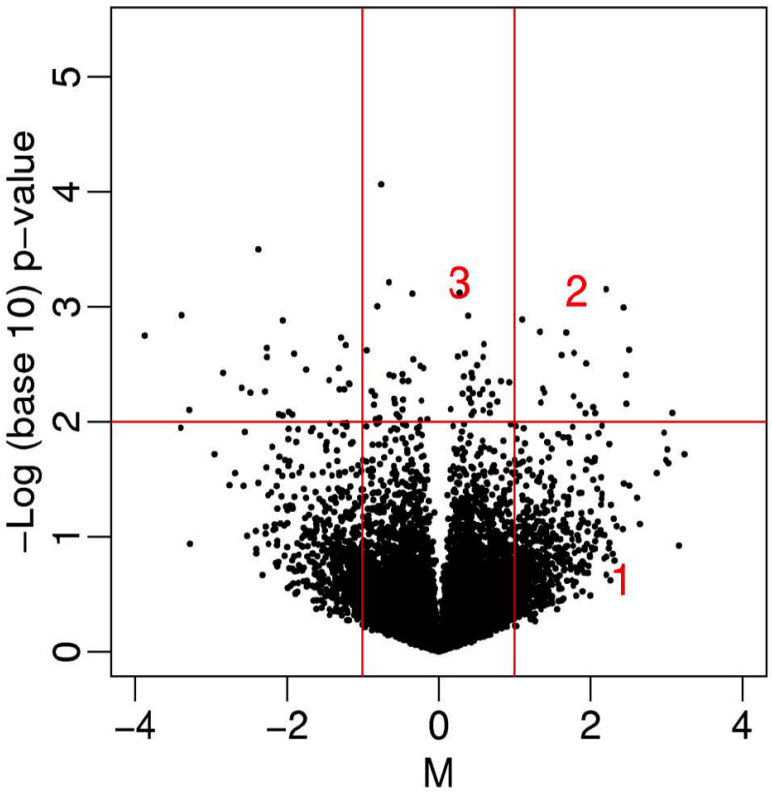

Removed
Frequency
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
p−value

Volcano Plot

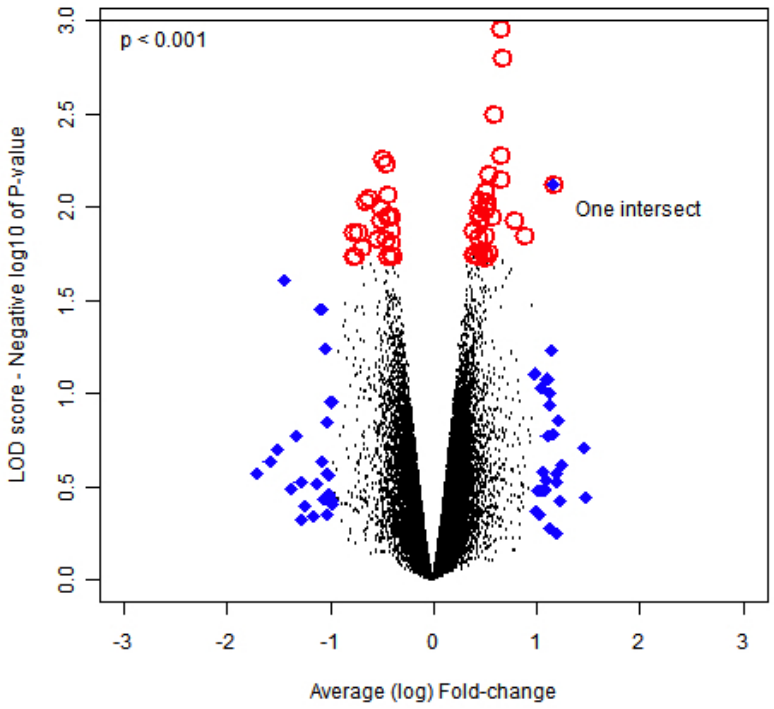
- A diagnostic plot to visualize the test results
- Scatter plot of statistical significance (log p values) versus biological significance (log fold-changes)
- Ideally the two should agree with each other

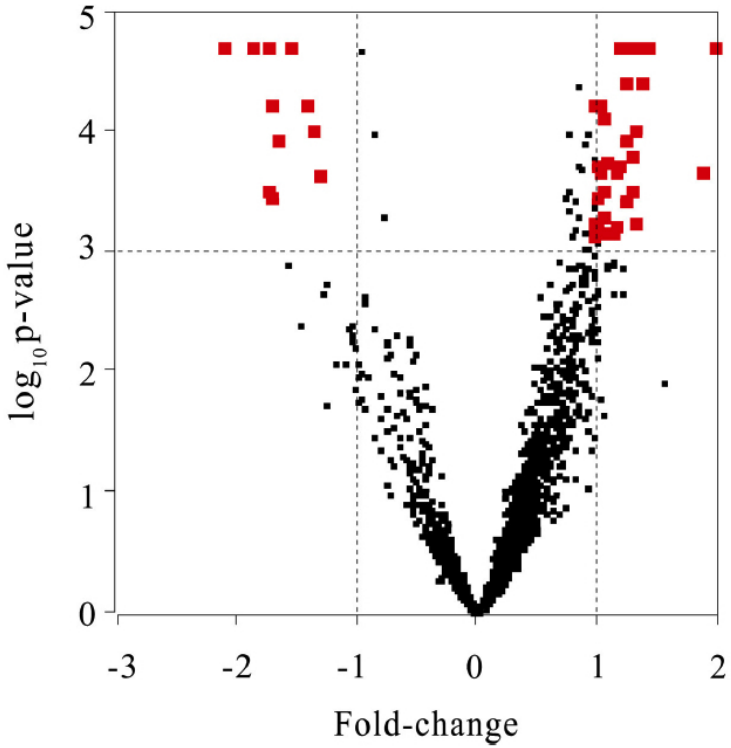
MA and Volcano Plots





Volcano Plots: Bad Versus Good





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

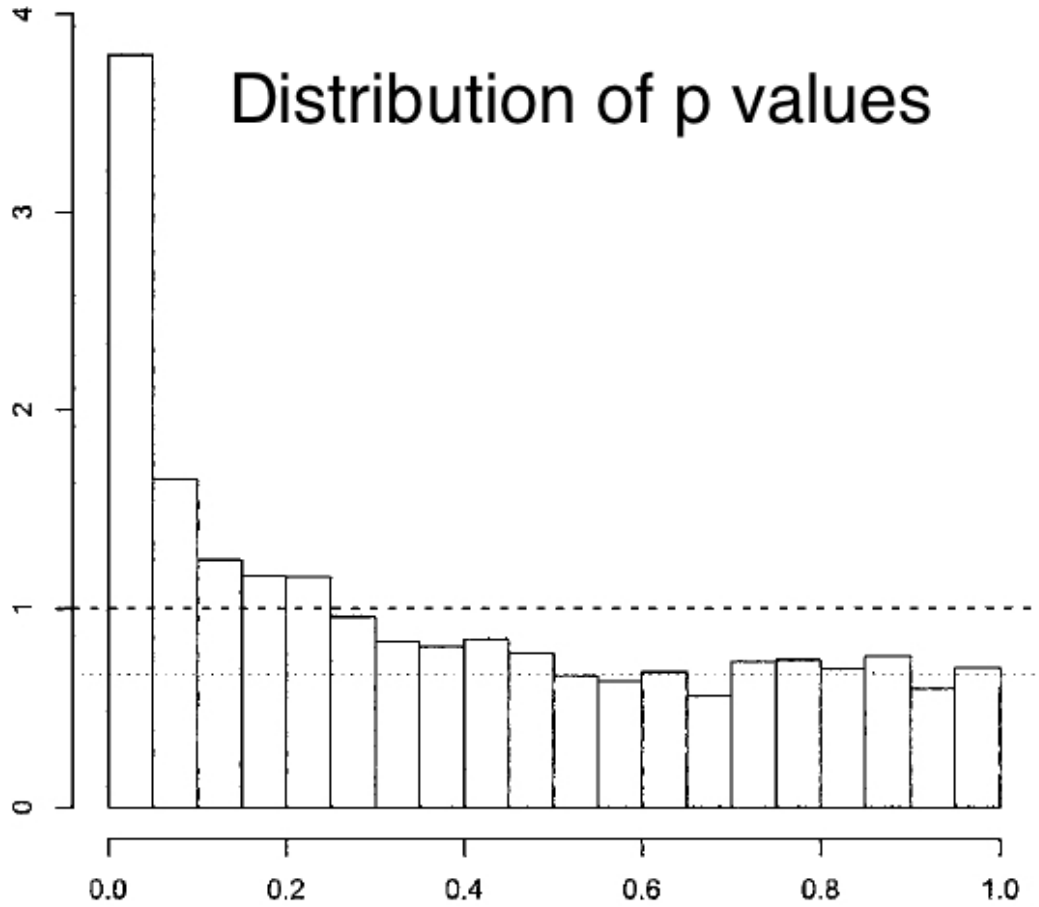
Make a Volcano Plot in R

Exercise: Make a Volcano plot in R

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)



Multiple Comparison Adjustment

```
> library("qvalue")
> mt = p.adjust(tt$p.value, method="BH")
> fdr<-qvalue(tt$p.value)$qvalues
> Table<-cbind(tt, fdr)
> Table[1:10,]
> o<-order(Table$fdr)
> Table<-Table[o,]
> Table[1:10,]
```

Top Table

	statistic	dm	p.value	fdr
1636_g_at	9.26	1.10	0.00	0.00
39730_at	8.69	1.15	0.00	0.00
1635_at	7.28	1.20	0.00	0.00
1674_at	6.90	1.43	0.00	0.00
40504_at	6.57	1.18	0.00	0.00
37015_at	6.19	1.03	0.00	0.00
40202_at	6.18	1.78	0.00	0.00
32434_at	5.78	1.68	0.00	0.00
37027_at	5.65	1.35	0.00	0.00
39837_s_at	5.50	0.48	0.00	0.00

Annotation

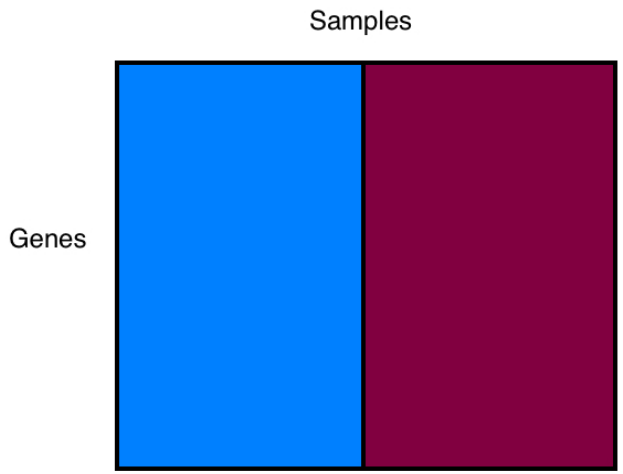
```
> g<-rownames(Table)
> biocLite("hgu95av2.db")
> library("hgu95av2.db")
> syms <- unlist(mget(g, hgu95av2SYMBOL))
> Table<-cbind(g, syms, Table)
> colnames(Table)<-c("probe", "gene symbol", "t",
"log2FC", "p.value", "fdr")
> Table[1:10,]
```

Top Table

	probe	gene symbol	t	log2FC	p.value	fdr
1	1636_g_at	ABL1	9.26	1.10	3.76e-14	2.65e-10
2	39730_at	ABL1	8.69	1.15	4.79e-13	1.69e-09
3	1635_at	ABL1	7.28	1.20	2.45e-10	5.74e-07
4	1674_at	YES1	6.90	1.43	1.28e-09	2.26e-06
5	40504_at	PON2	6.57	1.18	5.27e-09	7.42e-06
6	37015_at	ALDH1A1	6.19	1.03	2.74e-08	2.80e-05
7	40202_at	KLF9	6.18	1.78	2.79e-08	2.80e-05
8	32434_at	MARCKS	5.78	1.68	1.54e-07	1.35e-04
9	37027_at	AHNAK	5.65	1.35	2.60e-07	2.04e-04
10	39837_s_at	ZNF467	5.50	0.48	4.74e-07	3.34e-04

Simple Differential Expression in Two Populations

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



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

SD estimates are unstable.