Statistics for Human Genetics and Molecular Biology
Lecture 27: Differential Expression Analysis

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

Nov 09, 2015

Objectives of Lecture 27

- Simple Differential Expression
- Advanced Differential Expression

Borrowing Strength

- An advantage of having tens of thousands of genes is that we can try to learn about typical standard deviations by looking at all genes
- Empirical Bayes gives us a formal way of doing it

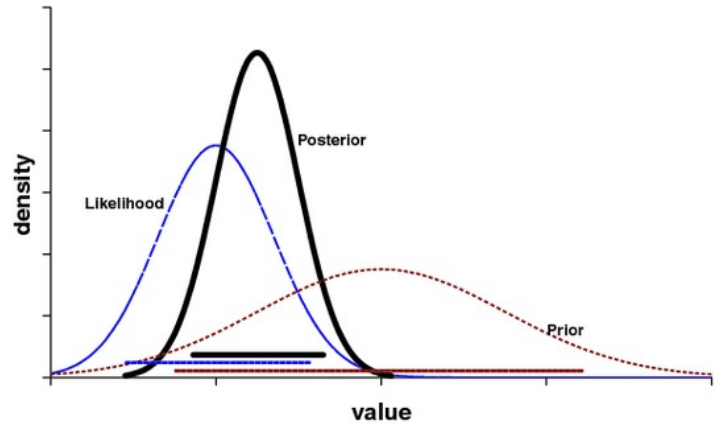
Bayesian Approach

$$\begin{aligned} \theta &\sim G \\ Y|\theta &\sim f(y|\theta) \end{aligned}$$

- G is the prior
- f is the sampling distribution

Posterior Distribution

$$g(\theta|Y) = \frac{f(y|\theta)g(\theta)}{f_G(Y)}$$



Marginal Distribution

$$f_G(Y) = \int f(y|u)g(u)du$$

Posterior Statistics: eBayes and SAM

Posterior variance estimators

$$\tilde{s}_g^2 = \frac{s_g^2 d_g + s_0^2 d_0}{d_g + d_0}$$

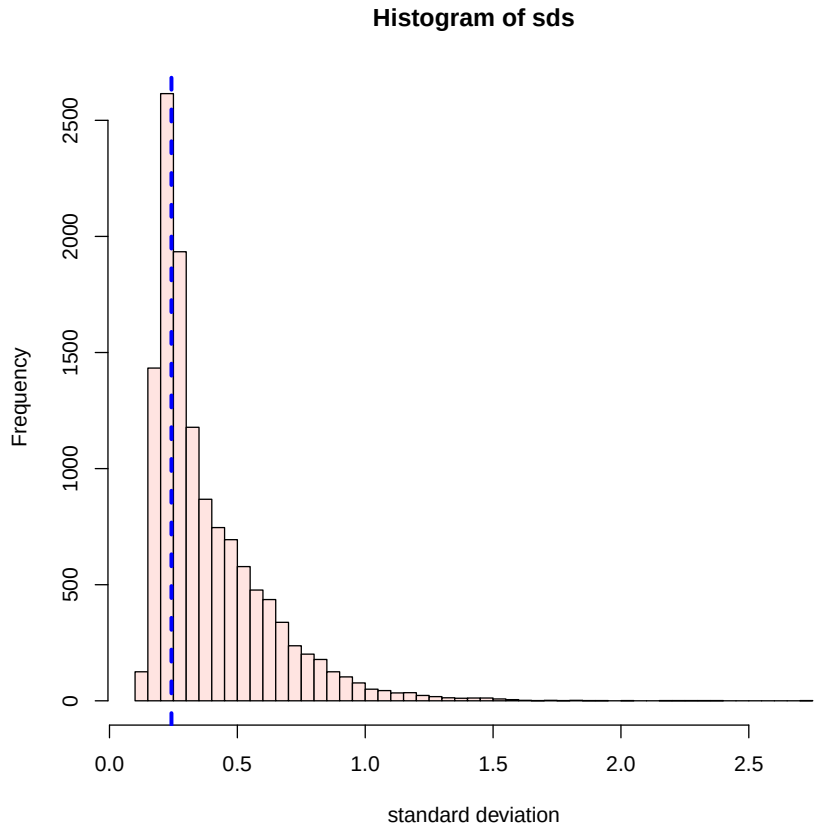
Moderated t-statistics

$$\tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{c_g}}$$

Eliminates large t-statistics merely from very small sd

Shrinkage Estimate

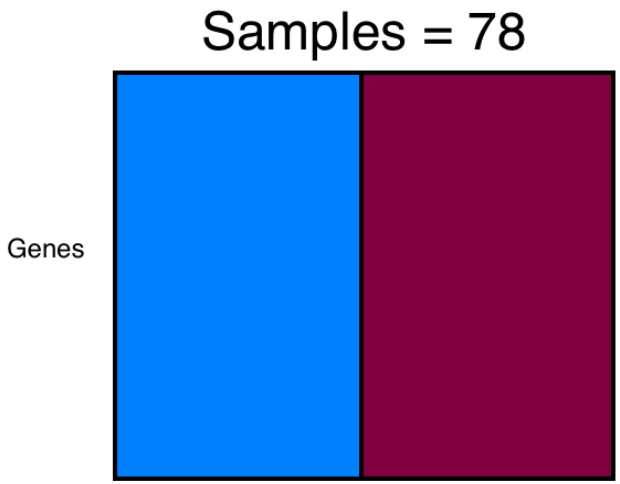
$$\tilde{s}_g^2 = \frac{s_g^2 d_g + s_0^2 d_0}{d_g + d_0}, \qquad \tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{c_g}}$$



Advanced Differential Expression

$$Y = \beta_0 + \beta_1 \times \text{BCR/ABL}$$

```
>library("limma")
>design = cbind(mean = 1, diff = cl)
>design[1:5,]
>fit = lmFit(exprs(ALLset1), design)
>fit = eBayes(fit)
>topTable(fit, coef=2)
```



$$\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \end{pmatrix}$$

Top Table

	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
156	1636_g_at	1.10	9.20	9.03	4.88e-14	1.23e-10	21.29
1915	39730_at	1.15	9.00	8.59	3.88e-13	4.89e-10	19.34
155	1635_at	1.20	7.90	7.34	1.23e-10	1.03e-07	13.91
163	1674_at	1.43	5.00	7.05	4.55e-10	2.87e-07	12.67
2066	40504_at	1.18	4.24	6.66	2.57e-09	1.30e-06	11.03
2014	40202_at	1.78	8.62	6.39	8.62e-09	3.63e-06	9.89
1262	37015_at	1.03	4.33	6.24	1.66e-08	6.00e-06	9.27
437	32434_at	1.68	4.47	5.97	5.38e-08	1.70e-05	8.16
1269	37027_at	1.35	8.44	5.81	1.10e-07	3.08e-05	7.49
1366	37403_at	1.12	5.09	5.48	4.27e-07	1.08e-04	6.21

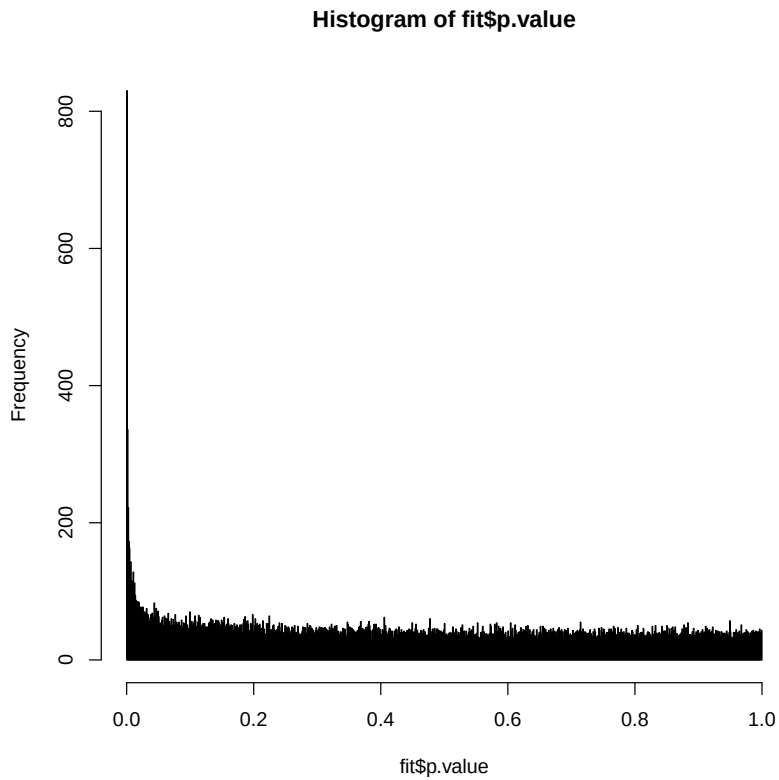
Annotation

```
>library("hgu95av2.db")
>ALLset1Syms = unlist(mget(featureNames(ALLset1), env
= hgu95av2SYMBOL))
>topTable(fit, coef = "diff", adjust.method = "fdr",
sort.by = "p", genelist = ALLset1Syms)
>plot(-log10(tt$p.value), -log10(fit$p.value[,
"diff"]), xlab = "-log10(p) from two-sample t-test",
ylab = "-log10(p) from moderated t-test (limma)",
pch=".")
>abline(c(0, 1), col = "red")
```

Top Table

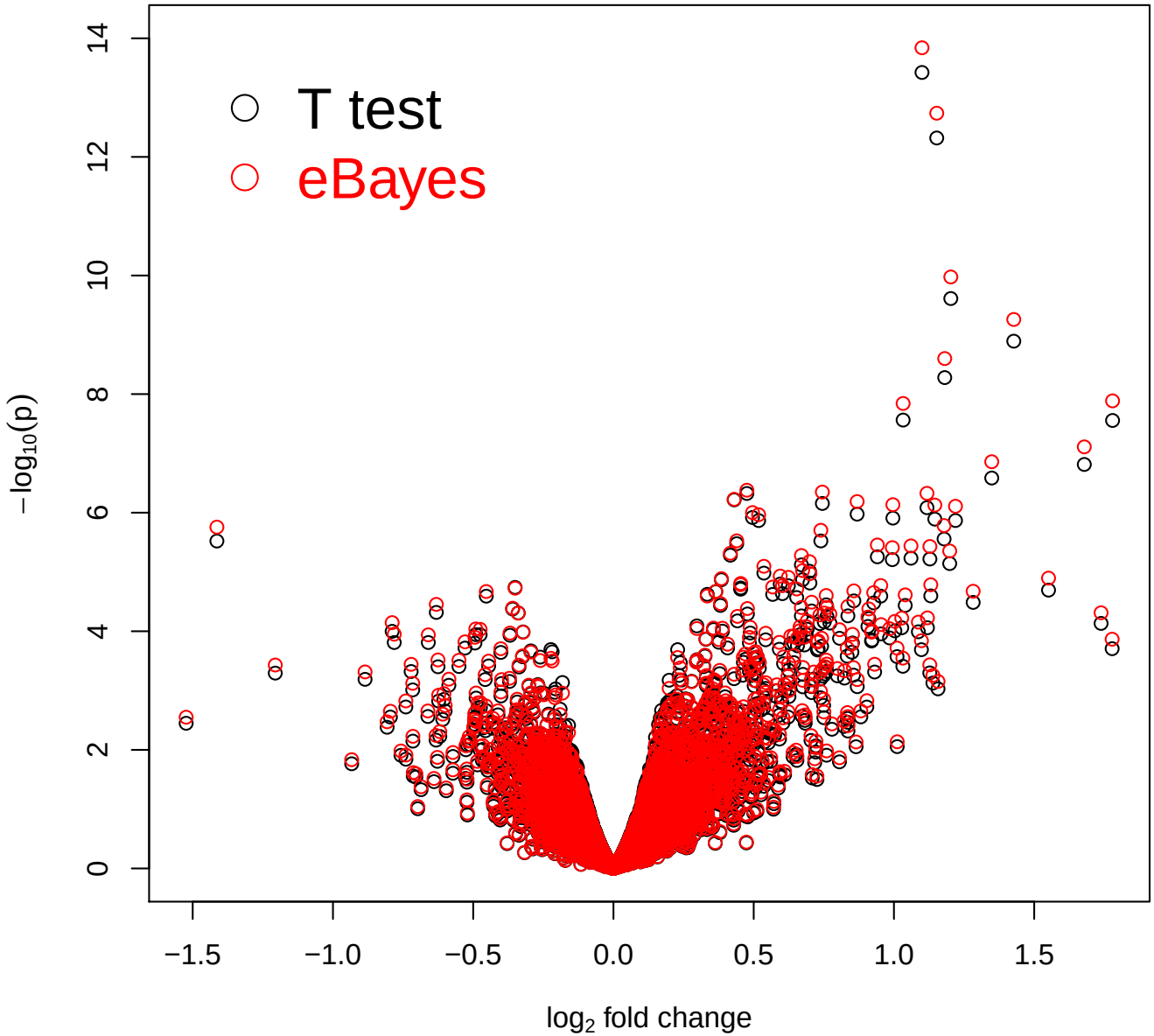
	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
156	ABL1	1.10	9.20	9.03	4.88e-14	1.23e-10	21.29
1915	ABL1	1.15	9.00	8.59	3.88e-13	4.89e-10	19.34
155	ABL1	1.20	7.90	7.34	1.23e-10	1.03e-07	13.91
163	YES1	1.43	5.00	7.05	4.55e-10	2.87e-07	12.67
2066	PON2	1.18	4.24	6.66	2.57e-09	1.30e-06	11.03
2014	KLF9	1.78	8.62	6.39	8.62e-09	3.63e-06	9.89
1262	ALDH1A1	1.03	4.33	6.24	1.66e-08	6.00e-06	9.27
437	MARCKS	1.68	4.47	5.97	5.38e-08	1.70e-05	8.16
1269	AHNAK	1.35	8.44	5.81	1.10e-07	3.08e-05	7.49
1366	ANXA1	1.12	5.09	5.48	4.27e-07	1.08e-04	6.21

P values

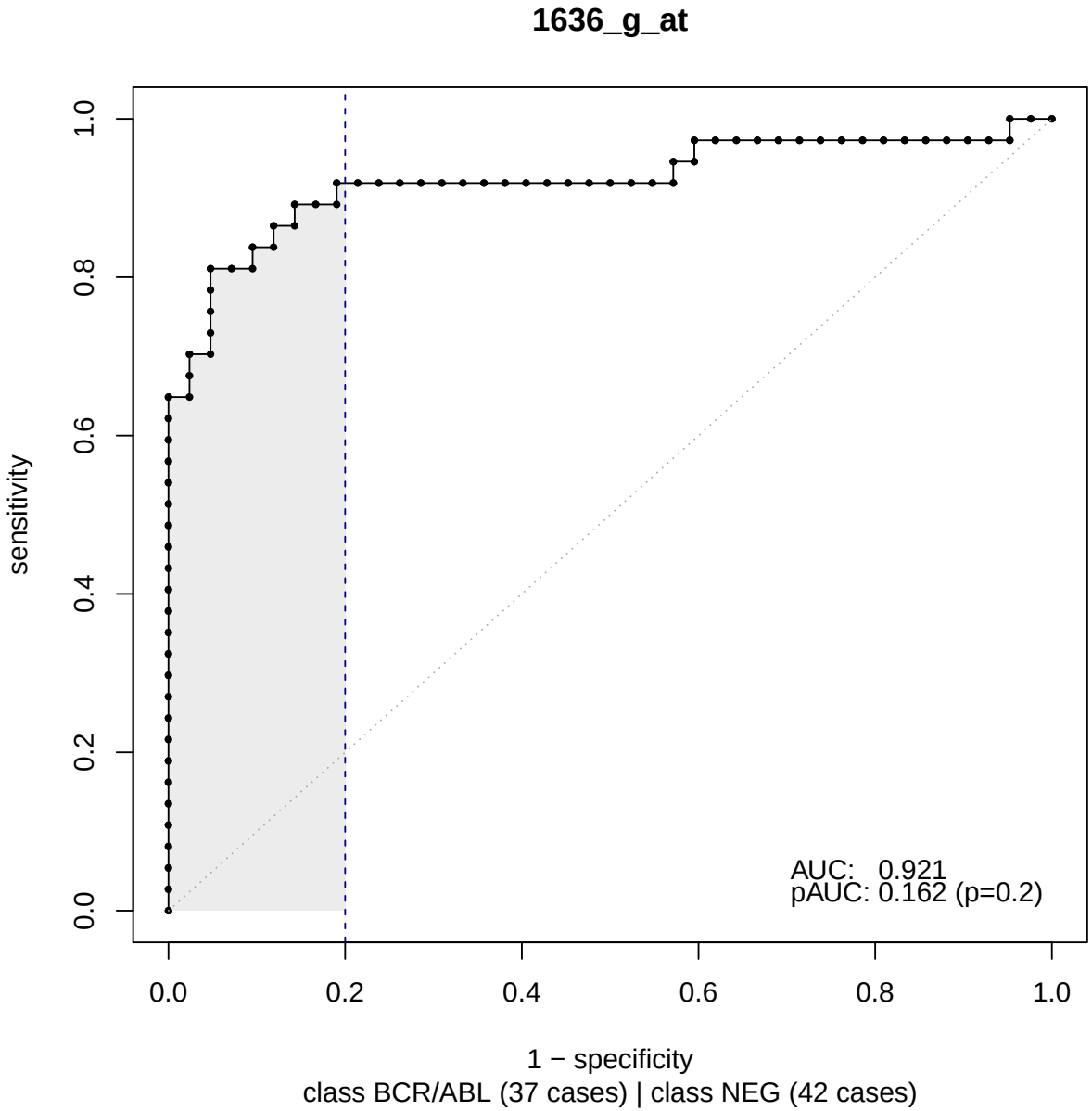


A p value of 0.05 implies that 5% of all the tests will result in false positives.
A FDR adjusted p value (or q value) of 0.05 implies that 5% of significant tests can be false positives.

Volcano Plot



Receiver Operator Chracteristic(ROC) Cuves



Summary

- When sample size is small, sd estimates are unstable and t-test is less powerful.
- eBayes and SAM provide more powerful moderate t-statistics which borrow strength from all genes.

Exercise: Homework 8 (3) (4) (5)

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

Hint:

```
library(oligo)
library(siggenes)
library(limma)
library(pd.hg.u133.plus.2)
library(hgu133plus2.db)
library(hgu133a.db)
library(ROCR)
```