
Project Introduction: Gene Expression Analysis

Utah State University
Bioinformatics: Problems and Solutions
Summer 2006

Traditional questions & extensions

- Which genes are really changing expression level between conditions (disease, tissue, genotype, etc.)?
- How much (and in which direction) are they really changing?
- Where do these genes lie? (chromosomal location)
- How are these genes different from the others? (gene ontology)

Recurring themes in differential expression tests

- Sample size
 - “small n, large p”
 - $n = \#$ of replicates or arrays \leftarrow (small)
 - $p = \#$ of genes \leftarrow (large)
 - affects ability to make useful statistical inference
 - “how big is big enough” – compare test statistic to “sampling distribution” – but this usually depends on large-sample theory
 - two main approaches
 - pool information across genes
 - generate sampling distribution using permutations
- Multiple testing
 - with thousands of genes, some will appear significant just by chance
 - need to adjust P-values somehow (false discovery rate – FDR by Benjamini & Hochberg)

A generalized t-test in a linear model (limma)

- For gene k under “treatment” j on array i :

$$Y_{ijk} = \beta_{k,0} + \beta_{k,1} T_{jk} + \varepsilon_{ijk}, \quad \text{Var}[\varepsilon_{ijk}] = \sigma_k^2$$

expression level (log scale) treatment effect (DE) treatment level (could be more than just 2 levels)

- What if there are more covariates than just treatment? –

use matrix notation for convenience:

$$E[Y_k] = X\beta_k$$

log-scale expression vector design matrix (n x m) covariate effects

Assumptions in linear model (Smyth)

Obtain estimates $\hat{\beta}_k$ and $\hat{\sigma}_k$, and $Var[\hat{\beta}_k] = V_k \hat{\sigma}_k^2$

For covariate w ,

$$\hat{\beta}_{k,w} \mid \beta_{k,w}, \sigma_k^2 \sim N(\beta_{k,w}, V_{k,w} \hat{\sigma}_k^2)$$

$$\hat{\sigma}_k^2 \mid \sigma_k^2 \sim \frac{\sigma_k^2}{d_k} \chi_{d_k}^2, \quad d_k = \text{resid. d.f.} = \underbrace{n_k - m_k}_{\text{k not necessary here}}$$

Then $t_{k,w} = \frac{\hat{\beta}_{k,w}}{\hat{\sigma}_k \sqrt{V_{k,w}}} \sim t_{d_k}$

Hierarchical model to borrow information across genes (Smyth): eBayes

Assume prior distribution $\frac{1}{\sigma_k^2} \sim \frac{1}{d_0 s_0^2} \chi_{d_0}^2$

$(s_0^2 \text{ and } d_0 \text{ estimated from data using empirical Bayes methods})$

(using all of the genes)

Consider the posterior mean $\tilde{\sigma}_k^2 = E[\sigma_k^2 | \hat{\sigma}_k^2] = \frac{d_0 s_0^2 + d_k \hat{\sigma}_k^2}{d_0 + d_k}$

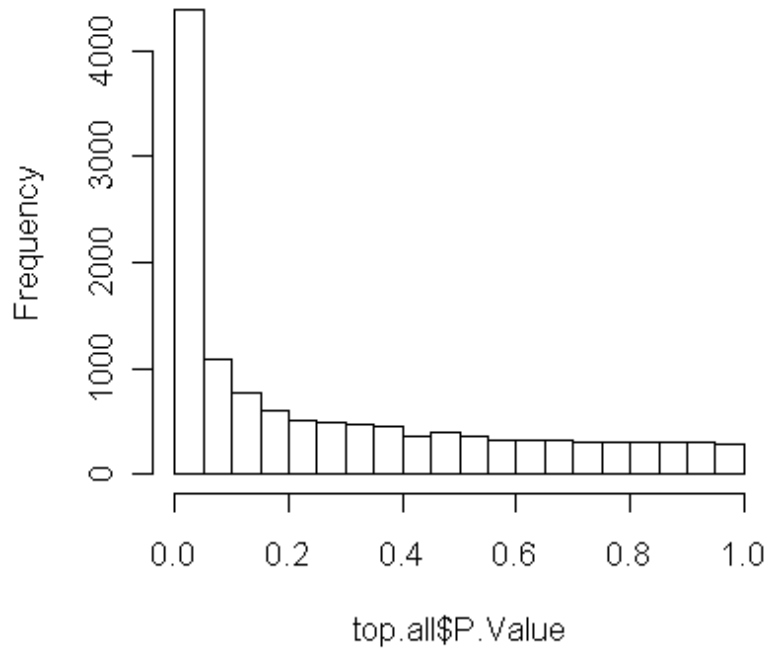
Then the "moderated" t - statistic $\tilde{t}_{k,w} = \frac{\hat{\beta}_{k,w}}{\tilde{\sigma}_k \sqrt{V_{k,w,w}}} \sim t_{d_0+d_k}$

↑
represents added information from using all genes

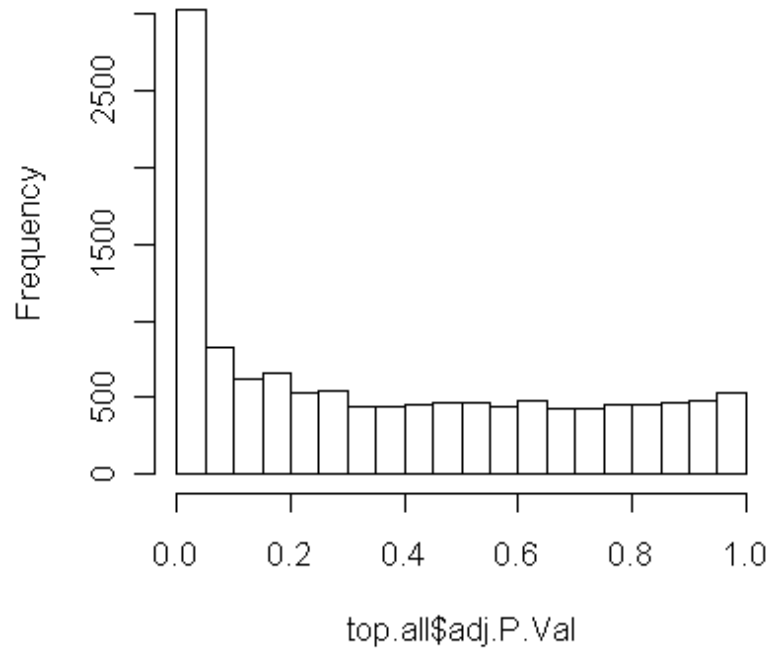
Differential expression: a quick example in R

```
# load data
library(ALL); data(ALL)
# define comparison (based on knowledge of samples)
eset <- ALL # these are normalized expression levels
sampleNames(eset)
trt <- c(rep(0,95),rep(1,33)) # 0=B, 1=T
# test for differential expression (DE)
library(limma)
design <- cbind(Intercept=1,trt=trt)
fit <- lmFit(eset@exprs,design)
e.fit <- eBayes(fit)
# Visualize results
top.all <- topTable(e.fit,n=nrow(eset@exprs),
                    coef=2,adjust="BH")
hist(top.all$P.Value,main='raw P-value')
hist(top.all$adj.P.Val,main='adj. P-value')
sum(top.all$adj.P.Val<0.05)
```

raw P-value



adj. P-value



Making a final report: a quick example

```
# Report for top 25 genes
top.25 <- topTable(e.fit,n=25,coef=2,adjust="BH")
gn.25 <- as.character(top.25$ID)
library(annaffy); aaf.handler() # annotation types
anncols <- aaf.handler()[c(1,5,6,11,12)] # pick columns
anntable <- aafTableAnn(gn.25,"hgu95av2",anncols)
add.table <- aafTable("Log Fold-Change"=top.25$M,
  "eBayes t"=top.25$t, "FDR-Adjusted P-Value"
  =top.25$adj.P.Val, signed=T)
new.table <- merge(anntable,add.table)
fname <- "C:\\folder\\ALL.top.25.html"
saveHTML(new.table,fname,
  title="Summary of Top 25 Significant Genes")
browseURL(fname)
# Look at tab-delimited format (for spreadsheet use)
fname <- "C:\\folder\\ALL.top25.txt"
saveText(new.table,fname)
```

Summary of Top 25 Significant Genes

Probe	Chromosome	Chromosome Location	PubMed	Gene Ontology	Log Fold-Change	eBayes t	FDR-Adjusted P-Value
38319_at	11	-117710476	33	transmembrane receptor activity protein binding cytoplasm protein complex assembly cell surface receptor linked signal transduction membrane integral to membrane T cell receptor complex T cell activation positive thymic T cell selection protein heterodimerization activity	4.65504	35.302	4.6283e-64
38147_at	X	123205728	39	SH3/SH2 adaptor activity cytoplasm cellular defense response intracellular signaling cascade cell-cell signaling	3.15474	26.3685	5.25779e-50
				regulation of progression through cell cycle regulation of progression through cell cycle nucleotide binding nucleotide binding			

“gene” location info.

“gene” function info.

“gene” DE results

Background

- Barley (*Hordeum vulgare*)
 - grain used for animal feed (poultry, e.g.) and human use (bread, beer)
 - Utah is among the top 12 U.S. producers
- Several cultivars (strains or genotypes): Kindred, Peruvian, Beka, ...
- Susceptible to pathogens causing leaf blotch: *Septoria passerinii* & *Septoria tritici*

Gene Expression Experiment

- Affymetrix barley1 arrays
- Gene expression observed for ~23,000 genes at different time points after pathogen inoculation

Treatment	Array ID's at Time			
Description	0	5	12	24
Kindred & water	158, 189	113, 190	114, 191	157, 192
Kindred & <i>Septoria passerinii</i>	159, 193	117, 194	160, 195	140, 196
Kindred & <i>Septoria tritici</i>	97, 197	120, 198	98, 199	99, 200
Peruvian & water	161, 201	122, 202	123, 203	204, 141
Peruvian & <i>Septoria passerinii</i>	162, 205	127, 206	163, 207	131, 208

Questions & Problems

■ Questions:

- ❑ Which genes are differentially expressed between cultivars?
- ❑ Where are these genes?
- ❑ How are they different from other genes?

■ Problems:

- ❑ Chromosomal location not automatically included in an annotation package for barley – but other sources possible:
 - hvuhomology package for R
 - **barleybase.org**, plantgdb.org, gramene.org
 - NetAffx from affymetrix.com
- ❑ Gene ontology information not available for all probe sets

Project description

- Analyze these barley data, focusing on:
 - graphical display of results
 - visualization of chromosomal locations of significant genes
 - summarization of possible over-representation of gene ontology terms among differentially expressed genes
- All team members will need to contribute, culminating in poster presentation

Issues to Consider

- Quality checks
- Summarization of spot intensities
- Choice of test for differential expression
- Multiple testing adjustments
- Filtering choices
- Use of annotation data
- Visualization of results

(We will briefly touch on each of these)