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 (small)
    - \( p = \) # of genes (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:

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

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 $\text{Var} [\hat{\beta}_k] = V_k \hat{\sigma}_k^2$

For covariate $w$,

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

$$\hat{\sigma}_k^2 | \sigma^2_k \sim \frac{\sigma^2_k}{d_k} \chi^2_{d_k} \text{, } d_k = \text{resid. d.f.} = n_k - m_k$$

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

$k$ not necessary here
Hierarchical model to borrow information across genes (Smyth): eBayes

Assume prior distribution \( \frac{1}{\sigma_k^2} \sim \frac{1}{d_0 s_0^2} \chi^2_{d_0} \)  

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

(\text{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)
Making a final report: a quick example

```r
# 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

<table>
<thead>
<tr>
<th>Probe</th>
<th>Chromosome</th>
<th>Chromosome Location</th>
<th>PubMed</th>
<th>Gene Ontology</th>
<th>Log Fold-Change</th>
<th>eBayes t</th>
<th>FDR-Adjusted P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>38319_at</td>
<td>11</td>
<td>-117710476</td>
<td>33</td>
<td>transmembrane receptor activity, protein binding</td>
<td>4.65504</td>
<td>35.302</td>
<td>4.6283e-64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>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</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>regulation of progression through cell cycle, regulation of progression through cell cycle, nucleotide binding</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **“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**

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Array ID’s at Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>0</td>
</tr>
<tr>
<td>Kindred &amp; water</td>
<td>158, 189</td>
</tr>
<tr>
<td>Kindred &amp; <em>Septoria passerinii</em></td>
<td>159, 193</td>
</tr>
<tr>
<td>Kindred &amp; <em>Septoria tritici</em></td>
<td>97, 197</td>
</tr>
<tr>
<td>Peruvian &amp; water</td>
<td>161, 201</td>
</tr>
<tr>
<td>Peruvian &amp; <em>Septoria passerinii</em></td>
<td>162, 205</td>
</tr>
</tbody>
</table>
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)