Candidate Gene Reports from Gene Expression Data

Utah State University – Spring 2014
STAT 5570: Statistical Bioinformatics
Notes 4.2

References

- Chapters 8 & 9 of Bioconductor Monograph (course text)
- PubMed:

  `http://www.ncbi.nlm.nih.gov/pubmed`

What to do after DE test?

- List of "candidate" genes
  - Declared significantly differentially expressed
  - Need to be further validated (RT-PCR, etc.)
- Measure of differential expression
  - Magnitude (Log Fold Change, test statistic, etc.)
  - P-value (adjusted for multiple testing)
- How to effectively represent and communicate these results?

  Use available resources to make a "nice" report

What would make the report "nice"?

- List of "candidate" genes
- Additional information about genes
  - Experimental results - LFC, P-value, etc.
  - Annotation - GO, KEGG, etc.
  - Previous study
- Interactive

PubMed

- Service of U.S. National Library of Medicine
- Archive of over 16 million journal citations (MEDLINE, other life science journals)
  each assigned a PubMed identifier
- Resource for summary of biomedical articles since the 1950s
- Includes links to full text articles and other related resources

How will we use PubMed?

- Quick literature search on “candidate” genes:
  from DE test, e.g.
- Generate report to summarize and investigate previous work
- Effectively communicate results

Summarizing Lists of Candidate Genes

- Bioconductor allows creation of summary tables
  - HTML format – with links to PubMed and others
  - Tab-delimited format – good to use in spreadsheet for sorting
- Good to include:
  - Probe set ID: (gene name in R)
  - Gene information
    (Symbol, PubMed, GO, etc.)
  - Experimental results
    (P-value, test statistic, LFC, etc.)

Summary of Top 25 Genes (limma/eBayes)

Probe	Symbol	Description	PubMed	Gene Ontology	Pathway	eBayes F	FDR-Adj. P-value	Log Fold-Change
38319_at	CD3D	CD3d molecule, delta (CD3-TCR complex)	78	transcription coactivator activity transmembrane signaling receptor activity cytoplasm plasma membrane cell surface receptor signaling pathway integral to membrane T cell differentiation T cell costimulation T cell receptor complex T cell receptor complex alpha-beta T cell receptor complex positive thymic T cell selection positive regulation of transcription from RNA polymerase II promoter protein heterodimerization activity regulation of immune response T cell receptor signaling pathway	Hematopoietic cell lineage T cell receptor signaling pathway Chagas disease (American trypanosomiasis) Primary immunodeficiency	1242.1	2.11013e-64	-4.65504
…								
41165_g_at	IGHM	immunoglobulin heavy constant mu	32	antigen binding extracellular region plasma membrane immune response integral to membrane		197.735	1.51667e-25	2.82841
32649_at	TCF7	transcription factor 7 (T-cell specific, HMG-box)	43	chromatin binding sequence-specific DNA binding transcription factor activity protein binding nucleus transcription factor complex transcription, DNA-dependent regulation of transcription, DNA-dependent regulation of transcription from RNA polymerase II promoter immune response brain development beta-catenin binding Wnt receptor signaling pathway neurogenesis sequence-specific DNA binding transcription regulatory region DNA binding generation of neurons canonical Wnt receptor signaling pathway cellular response to interleukin-4	Wnt signaling pathway Adherens junction Melanogenesis Pathways in cancer Colorectal cancer Endometrial cancer Prostate cancer Thyroid cancer Basal cell carcinoma Acute myeloid leukemia Arrhythmogenic right ventricular cardiomyopathy (ARVC)	181.357	4.07812e-24	-3.47667

color = sign

```
# Load data, filter, and test for DE
# (as on slides 2, 8, and 13 of Notes 3.4)

library(affy); library(ALL); data(ALL)
library(genefilter); e.mat <- 2^exprs(ALL)
ffun <- filterfun(pOverA(0.20,100))
t.fil <- genefilter(e.mat,ffun)
small.eset <- log2(e.mat[t.fil,])
dim(small.eset) # 4305 genes, 128 arrays

library(limma)
Cell <- as.factor(c(rep('B',95),rep('T',33)))
design <- model.matrix(~0+Cell)
colnames(design) <- c('B','T')
fit <- lmFit(small.eset, design)
contrast.Cell <- makeContrasts(B-T, levels=design)
fit.Cell <- contrasts.fit(fit, contrast.Cell)
final.fit.Cell <- eBayes(fit.Cell)

top.Cell <- topTableF(final.fit.Cell, n=nrow(small.eset))
  # see head(top.Cell on slide 15 here)
```



```
# Get gene name, test statistic, adjusted P-value,
# and LFC to include in table (in same order)
gn.25 <- rownames(top.Cell)[1:25]
test.stat <- top.Cell$F[1:25]
adj.P <- top.Cell$adj.P.Val[1:25]
LFC <- top.Cell$B...T[1:25]

# Create report
library(annaffy)
aaf.handler() # Shows available annotation types
# [1] "Probe"            "Symbol"              "Description"
# [4] "Chromosome"       "Chromosome Location" "GenBank"
# [7] "Gene"             "Cytoband"            "UniGene"
#[10] "PubMed"           "Gene Ontology"       "Pathway"
# Choose columns: Probe, Symbol, Description,
#                 PubMed, Gene Ontology, Pathway

anncols <- aaf.handler()[c(1:3,10:12)]
```



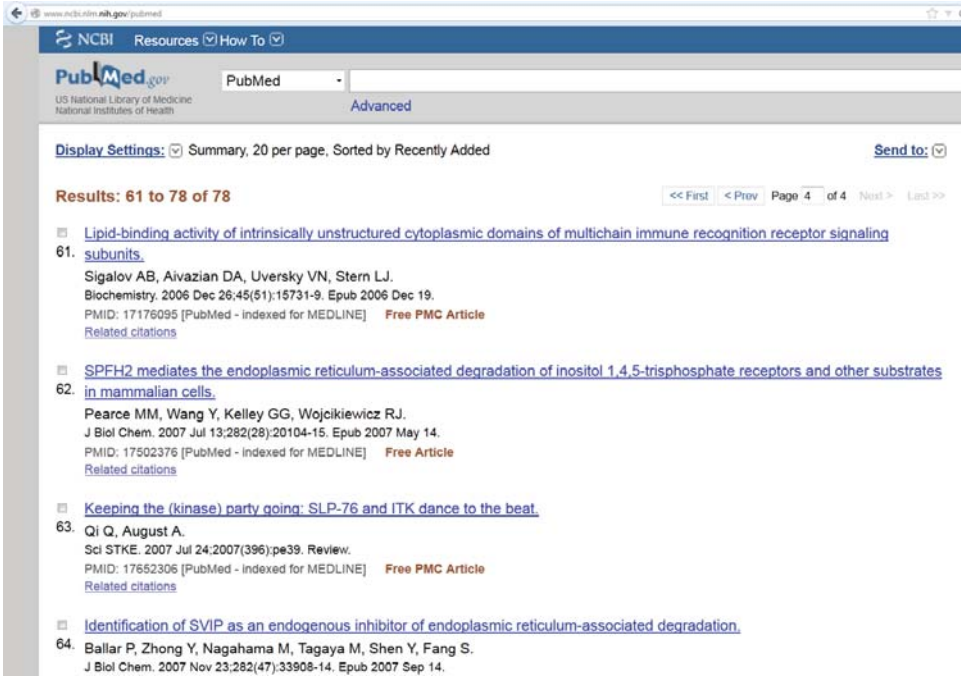
```
# Construct table with desired information
anntable <- aafTableAnn(gn.25,"hgu95av2.db",anncols)
add.table <- aafTable("eBayes F"=test.stat,
  "FDR-Adj. P-value"=adj.P, "Log Fold-Change"=LFC, signed=T)
new.table <- merge(anntable,add.table)

# Look at HTML format
fname <- "C:\\folder\\limma.LFC.html"
saveHTML(new.table,fname,
         title="Summary of Top 25 Genes (limma/eBayes)")
browseURL(fname)
```


Summary of Top 25 Genes (limma/eBayes)

Probe	Symbol	Description	PubMed	Gene Ontology	Pathway	eBayes F	FDR-Adj. P-value	Log Fold-Change
38319_at	CD3D	CD3d molecule, delta (CD3-TCR complex)	78	transcription coactivator activity transmembrane signaling receptor activity cytoplasm plasma membrane cell surface receptor signaling pathway integral to membrane T cell differentiation T cell costimulation T cell receptor complex T cell receptor complex alpha-beta T cell receptor complex positive thymic T cell selection positive regulation of transcription from RNA polymerase II promoter protein heterodimerization activity regulation of immune response T cell receptor signaling pathway	Hematopoietic cell lineage T cell receptor signaling pathway Chagas disease (American trypanosomiasis) Primary immunodeficiency	1242.1	2.11013e-64	-4.65504
…								
41165_g_at	IGHM	immunoglobulin heavy constant mu	32	antigen binding extracellular region plasma membrane immune response integral to membrane		197.735	1.51667e-25	2.82841
32649_at	TCF7	transcription factor 7 (T-cell specific, HMG-box)	43	chromatin binding sequence-specific DNA binding transcription factor activity protein binding nucleus transcription factor complex transcription, DNA-dependent regulation of transcription, DNA-dependent regulation of transcription from RNA polymerase II promoter immune response brain development beta-catenin binding Wnt receptor signaling pathway neurogenesis sequence-specific DNA binding transcription regulatory region DNA binding generation of neurons canonical Wnt receptor signaling pathway cellular response to interleukin-4	Wnt signaling pathway Adherens junction Melanogenesis Pathways in cancer Colorectal cancer Endometrial cancer Prostate cancer Thyroid cancer Basal cell carcinoma Acute myeloid leukemia Arrhythmogenic right ventricular cardiomyopathy (ARVC)	181.357	4.07812e-24	-3.47667

color = sign



Summary

- PubMed is a good resource to search previous work on a list of “candidate” genes
- Use PubMed and summary tables to create nice-looking reports – to effectively communicate results
- Other Bioconductor interfaces to online resources (see Ch. 8 of course text):
  - KEGGSOAP – look at pathways and sequence motifs
  - Biostrings – look at gene sequence information