

Bioinformatics Homework 4

Style 2-3: NON - Literature Project

Assignment:

gene expression microarray analysis in R/Bioconductor

More detailed information is available in the presentation entitled 'Gene expression microarray profiling in practice' given at 20/11/2012

Exercise: We will compare the colonic gene expression microarray profiles (U133plus2 arrays) between patients with ulcerative colitis (UC, n=8) and healthy controls (n=6) to identify which genes (probe sets) are differentially expressed between both groups. Next, we will study which biological pathways are overrepresented in the list of significant genes.

Download R - Bioconductor:

- Download the latest version of R (R 2.15.2; <http://cran.r-project.org/>)
- Download the latest version of Bioconductor (version 2.11; (<http://www.bioconductor.org/download>) in R 2.15.2 with following R codes:

R commands:

```
source("http://bioconductor.org/biocLite.R")
biocLite()
```

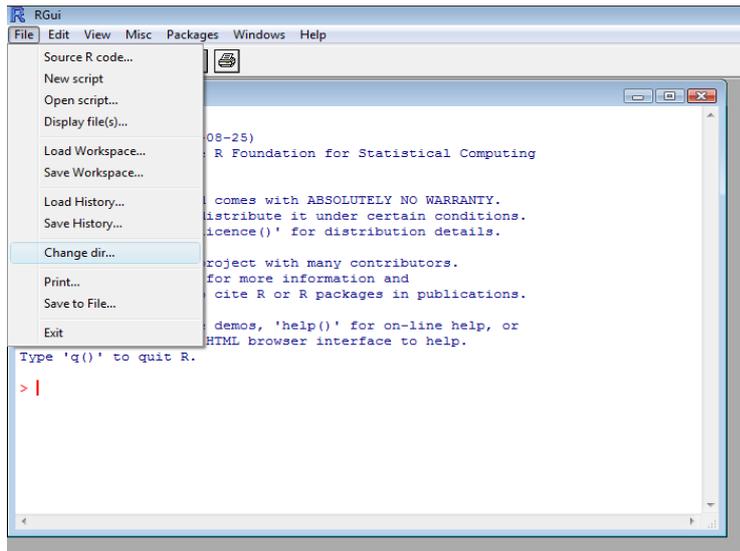
- Download the following packages in R 2.15.2:

R commands:

```
biocLite(c("affy", "hgu133plus2.db", "hgu133plus2cdf", "genefilter", "limma", "gplots", "annaffy"))
```

Load the data in R:

- Make a file on your computer containing all raw data (cel files) and phenoData.txt. There are in total 14 cel files (8 cel files from 8 patients with ulcerative colitis and 8 cel files from normal controls) and a phenodata.txt file containing the phenotype information).
- Go to FILE in R 2.15.2 → Change dir → menu item (the file with the raw data (cel.files) and phenoData.txt)



Data analysis:

- Normalization with RMA method

R commands:

```
library(affy)
library(hgu133plus2.db)
pd<-
read.AnnotatedDataFrame("phenoData.txt",header=TRUE,row
.names=1)
datarma<-
justRMA(filenamees=rownames(pData(pd)),phenoData=pd)
write.exprs(datarma, file="datarma.txt")
```

Question:

How many probe set IDs do we have for further data analysis? (see datarma.txt file that is saved in your working directory)

- Non-specific filtering: To eliminate non-relevant probe sets

R commands:

```
eset<-
datarma[,pData(datarma)[,"Disease"]%in%c("UC","control"
)]
dim(eset)
pData(eset)
library(genefilter)
f1<-pOverA(0.10,log2(100))
f2<-function(x)(IQR(x)>0.5)
```

```

ff<-filterfun(f1,f2)
selected<-genefilter(eset,ff)
sum(selected)
esetSub<-eset[selected,]
table(selected)
esetSub

```

Question:

How many features and samples have 'eset'?

Why are we doing a filtering and which filter criteria did we use in this exercise?

After filtering, how many probe sets are left for further data analysis?

- Unsupervised average-linkage hierarchical clustering on filtered probe sets

R commands:

```

dat<-exprs(esetSub)
dim(dat)
d<-dist(t(dat))
hc<-hclust(d,method="average")
plot(hc,cex=0.8)
labels<-pData(esetSub)[,"Disease"]
plot(hc,cex=0.8,labels)

```

Question:

Explain the figure that you get.

- Identifying differentially expressed genes (probe sets) between UC and controls.

R commands:

```

esetsel<-
esetSub[,pData(esetSub)[,"Disease"]%in%c("UC","control")]
dim(esetsel)
library(limma)
f<-factor(as.character(esetsel$Disease))
f
design<-model.matrix(~f)
design
fit<-lmFit(exprs(esetsel),design)
fit2<-eBayes(fit)
options(digits=2)
topTable(fit2,coef=2,adjust="BH",number=100)

```

```

topTableall<-topTable(fit2,coef=2,adjust="BH",number=9183)
write.table(topTableall,file="topTableall.xls",sep="\t",quote=F)
library(annaffy)
topTableall.descs<-
aafDescription(topTableall$ID,"hgu133plus2.db")
topTableall.descs[1:3,]
gn<-as.character(topTableall$ID)
genesymbols<-
unlist(mget(gn,hgu133plus2SYMBOL,ifnotfound=NA))
topTableallgenesymbols<-
data.frame(topTableall,genesymbols)
write.table(topTableallgenesymbols,file="topTableallgenesymbols.xls",sep="\t",quote=F)
topTablesig<-topTableall[topTableall$adj.P.Val<0.05,]
dim(topTablesig)
topTablesig2FC<-
topTableall[topTableall$adj.P.Val<0.05&(topTableall$logFC>1|topTableall$logFC<(-1)),]
dim(topTablesig2FC)
uptopTablesig2FC<-
topTableall[topTableall$adj.P.Val<0.05&(topTableall$logFC>1),]
downtopTablesig2FC<-
topTableall[topTableall$adj.P.Val<0.05&(topTableall$logFC<(-1)),]
dim(uptopTablesig2FC)
dim(downtopTablesig2FC)

```

Question:

Which package in R did we use to identify differentially expressed genes?

Which method did we use for multiple testing correction?

How many probe sets are significantly (adjusted p-value <5%) differentially expressed between UC and control? In this list of significant probe sets, how many probe sets have >2-fold change? Please note that FC are logFC, e.g. $\log FC 1 = FC 2^1$, $\log FC 2 = FC 2^2$.

Which is the most significantly differentially expressed gene (probe set)? Is this gene up- or downregulated in UC vs. controls? (see also topTableallgenesymbols.xls in your working directory)

How many probe sets (genes) are >2-fold significantly INCREASED in UC vs. controls, and which gene (probe set) is most significant? (see also topTableallgenesymbols.xls in your working directory)

How many probe sets (genes) are >2-fold significantly DECREASED in UC vs. controls, and which gene (probe set) is the most significant one? (see also topTableallgenesymbols.xls in your working directory)

Bio Functional analysis:

- Go to DAVID (<http://david.abcc.ncifcrf.gov/>) → Functional annotation → Paste the significant probe set IDs that are >2-fold increased in UC vs. controls (copy these probe sets IDs from topTableallgenesymbols.xls file that is present in your working directory) in the LEFT panel under STEP 1 (see figure below) → In STEP 2, take as identifier AFFYMETRIX_3PRIME_IVT_ID (because we used U133plus2 arrays) → In STEP 3, take the option 'gene list' → In STEP 4, submit list → Click on chart under Gene Ontology Category Biological pathways (GOTERM_BP_ALL) (see figure below) → Biological pathways which are most predominant to the list of increased genes in UC vs. controls are given.

DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service Why DAVID? About Us

*** Announcing the new DAVID Web Service which allows access to DAVID from various programming languages. [More info...](#) ***

Shortcut to DAVID Tools

Functional Annotation
Gene-annotation enrichment analysis, functional annotation clustering, BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and [more](#)

Gene Functional Classification
Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. [More](#)

Gene ID Conversion
Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. [More](#)

Gene Name Batch Viewer
Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. [More](#)

<http://david.abcc.ncifcrf.gov/summary.jsp>

Recommending: A [paper](#) published in *Nature Protocols* describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.7

2003 - 2012

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 is an update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes not in the list
- List interacting proteins
- Explore gene names in batch
- Link gene-disease associations
- Highlight protein functional domains and motifs
- Redirect to related literatures
- Convert gene identifiers from one type to another.

What's Important in DAVID?

- [Current \(v 6.7\) release note](#)
- [New requirement to cite DAVID](#)
- [IDs of Affy Exon and Gene arrays supported](#)
- [Novel Classification Algorithms](#)
- [Pre-built Affymetrix and Illumina backgrounds](#)
- [User's customized gene background](#)
- [Enhanced calculating speed](#)

Statistics of DAVID

DAVID Bioinformatic Resources Citations

Year	Citations
2003	~500
2004	~700
2005	~1000
2006	~1500
2007	~1800

***** Announcing the new DAVID Web Service which allows access to DAVID from various programming languages. [More info...](#) *****

Upload | List | Background

Upload Gene List

[Demolist 1](#) [Demolist 2](#)
[Upload Help](#)

Step 1: Enter Gene List

A: Paste a list

208850_s_at

204673_at

224796_at

200788_s_at

Or

B: Choose From a File

Multi-List File ?

Step 2: Select Identifier

AFFYMETRIX_3PRIME_IVT_ID

Step 3: List Type

Gene List

Background

Step 4: Submit List

Functional Annotation Tool

← Submit your gene list to start the tool!

[Tell us how you like the tool](#)
[Read technical notes of the tool](#)
[Contact us for questions](#)

Key Concepts:

The DAVID Gene Concept

DAVID 6.7 is designed around the "DAVID Gene Concept", a graph theory evidence-based method to agglomerate species-specific gene/protein identifiers from a variety of public genomic resources including NCBI, PIR and Uniprot/SwissProt. The DAVID Gene Concept method groups tens of million of identifiers from over 65,000 species into 1.5 million unique protein/gene records. [More](#)

Term/Gene Co-Occurrence Probability

Ranking functional categories based on co-occurrence with sets of genes in a gene list can rapidly aid in unraveling new biological processes associated with cellular functions and pathways. DAVID 6.7 allows investigators to sort gene categories from dozens of annotation systems. Sorting can be based either the number of genes within each category or by the EASE-score. [More](#)

Gene Similarity Search

Any given gene is associating with a set of annotation terms. If genes share similar set of those terms, they are most likely involved in similar biological mechanisms. The algorithm tries to group those related genes based on the agreement of sharing similar annotation terms by Kappa statistics. [More](#)

Term Similarity Search

Typically, a biological process/term is done by a corporation of a set of genes. If two or more biological processes are done by similar set of genes, the processes might be related in the biological network somehow. This search function is to identify the related biological processes/terms by quantitatively measuring the degree of the agreement how terms share the similar participating genes. [More](#)

***** Announcing the new DAVID Web Service which allows access to DAVID from various programming languages. [More info...](#) *****

Upload | **List** | Background

Gene List Manager

Select to limit annotations by one or more species [Help](#)

- Use All Species -

Homo sapiens(946)

Unknown(9)

List Manager [Help](#)

List_1

Select List to:

[View Unmapped Ids](#)

Annotation Summary Results

[Help and Tool Manual](#)

Current Gene List: List_1 **675 DAVID IDs**

Current Background: Homo sapiens **Check Defaults**

Disease (1 selected)

Functional_Categories (3 selected)

Gene_Ontology (1 selected)

Category	Percentage	Count	Chart
<input type="checkbox"/> GOTERM_BP_1	82.5%	557	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_BP_2	82.4%	556	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_BP_3	79.0%	533	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_BP_4	76.9%	519	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_BP_5	70.7%	477	<input type="button" value="Chart"/>
<input checked="" type="checkbox"/> GOTERM_BP_ALL	82.7%	558	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_BP_FAT ?	80.3%	542	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_CC_1	88.7%	599	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_CC_2	84.6%	571	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_CC_3	84.6%	571	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_CC_4	79.7%	538	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_CC_5	76.9%	519	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_CC_ALL	88.7%	599	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_CC_FAT ?	76.7%	518	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_MF_1	84.7%	572	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_MF_2	83.9%	566	<input type="button" value="Chart"/>
<input type="checkbox"/> GOTERM_MF_3	75.4%	509	<input type="button" value="Chart"/>

Question:

Give the top 5 biological pathways that are most significant to the dataset of increased (>2-fold) genes in UC vs. controls? Are these pathways relevant for UC?

Due date of report: 15 January (before midnight)