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()
```

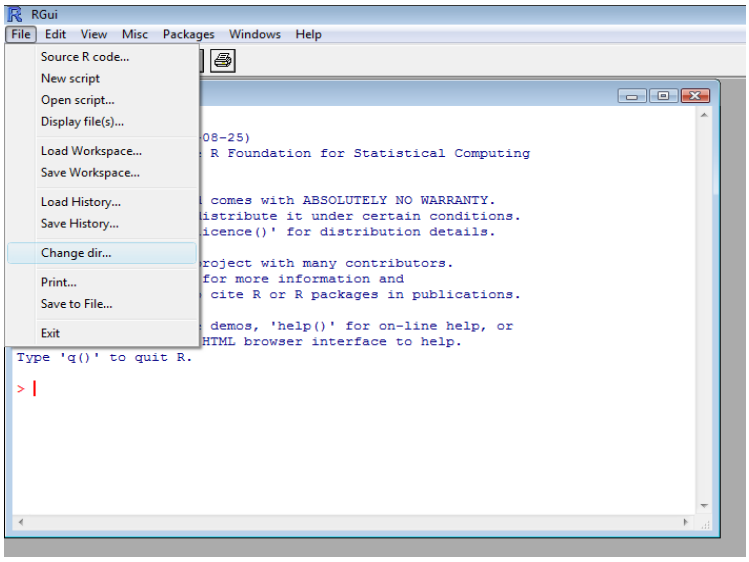
- Download the following packages in R 2.15.2:

  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 celfiles 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(filenames=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",qu
ote=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="topTableallgenesy
mbols.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. logFC 1 = FC 2^1, logFC2, FC 2^2.

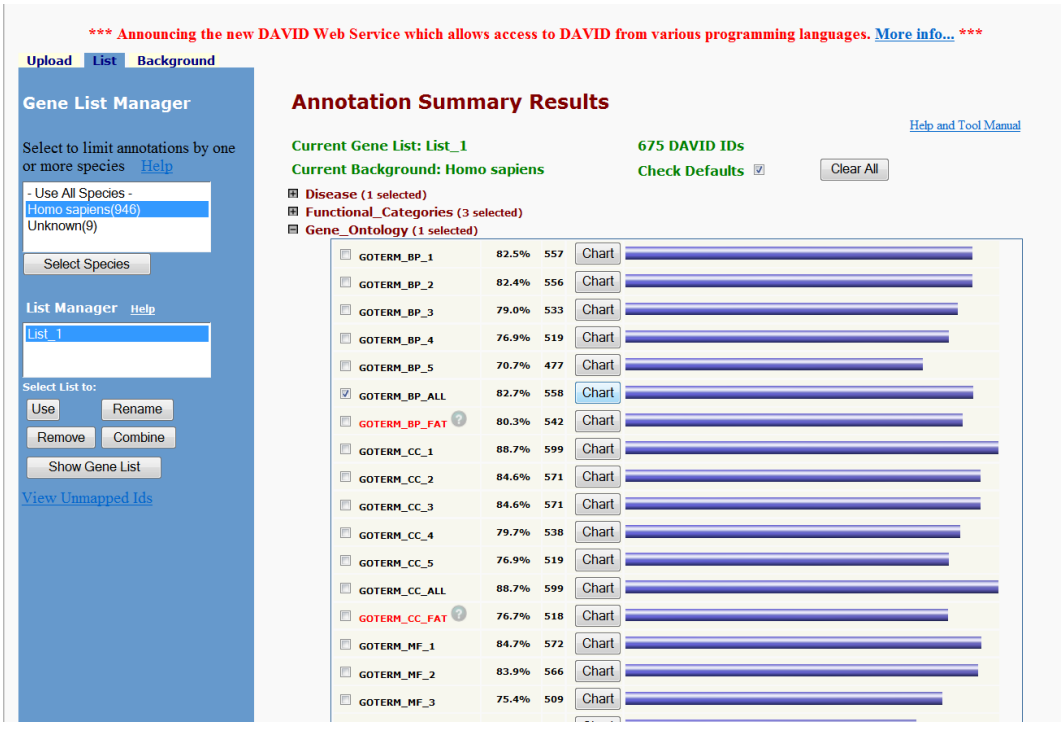
Which is the most significantly differentially expressed gene (probe set)? Is this gene up- or downregulated in UC vs. controls? (see also topTableallgenesymbols.xls in your working directory)

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.



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)