M2-BI : Approches méthodologiques en bioinformatiques
Karine Audouze

Exercice 3 : Normalization ans statistical analysis of microarray data exercise

In this exercise we will analyze data from a microarray experiment - comparing two cell types from both HIV infected patients and healthy donors.

1) We will look at the difference between not-normalized and normalized data
2) We will identify genes that are expressed differentially between the HIV and controls by using the common *t*-test and a two-way ANalysis of VAriance (ANOVA).

It takes 2 hours to complete this exercise. Please, take your time and read the instructions care- fully.

HIV virus particles

Introduction

From the blood of each of 10 individuals (5 healthy donors and 5 HIV-infected patients) two cell types have been extracted to measure the gene expression in:

T-helper cells (CD4+ T-cells or Th-cells) &

Cytotoxic T lymphocytes (CD8+ T-cells or CTLs)

The four different categories that we end up with are: Th-cells and CTLs from both control and HIV-infected (5 samples in each category).

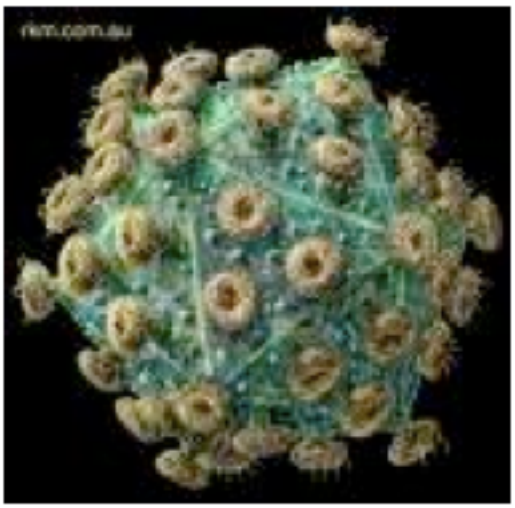
	Healthy	**HIV**
Th	5	5
CTL	5	5

For further background, please, check out this paper: Hyrcsa *et. al*, 2007

The main questions that we would typically seek to address using the data above are:

1. Which genes are expressed differently in HIV infected patients ?- the disease effect –
2. Do the cell types respond in different ways to the infection, i.e. there is any difference between which genes that we find differentially expressed in the two cell types ? (healthy versus infected T-helper cells and healthy versus infected CTLs) - the interaction effect
3. Are they any significant differential gene expression between the two cell types ? (T-helper cells versus CTLs)- the cell type effect

When we do a two-way ANOVA we actually get the answer to all these questions doing only one test, without the fuzz of comparing lists - we even get a p-value with additional statistical power, and hence a more trustworthy result.



Exercise

Please, download and move the files CELL to your local system.

Loading and normalizing:

[In R] First, load the *affy* library and then specify the file names of your data files: `library(affy) #source("http://bioconductor.org/biocLite.R")` to do first if not installed

Now, load the data (creating an *affy batch*):

```
HIV.affyb <- ReadAffy(filenames=sort(list.celfiles("celfiles",
full=TRUE)))# Here, you may need to specify the path to your cel files
```

Use the Robust Multi-array Average (RMA) method to background correct, normalize and sum- marize your probe values into expression index values. The expression index value represents the expression intensity for each probe-set in each sample. Extract these from the resulting 'Expres- sionSet' by the use of the *exprs()* function:

```
ei <- exprs( rma(HIV.affyb) )
```

Now you have a matrix containing the corrected and summarized data at probe-set level:

5 donor CTL cells (CTL) - the columns: 1-5

5 HIV CTL cells (HIV-CTL) - the columns: 6-10

5 HIV T-helper cells (HIV-Th) - the columns: 11-15

5 donor T-helper cells (Th) - the columns: 16-20

This is also indicated by the column names of the matrix 'ei'

```
colnames(ei)
```

one may write a data.frame describing the design like this:

```
design<-data.frame(HIV=rep(c("ctrl","HIV", "HIV",
"ctrl"),each=5), cell=
rep(c("CTL","Th"),each=10), row.names=colnames(ei))
design
```

This may also be done by reading a table from a text file eg.

```
# read.table("my_design_file.txt", sep="\t", header=TRUE)
```

Now lets look at the result of the normalization, you can make a density plot of the

normalized data like this:

```
plot ( density ( ei[,1], na.rm=TRUE ), xlim = c(0,16), ylim =
c(0,0.3), main
= "Normalized", xlab = "Intensity" )
for(i in 2:ncol( ei ) ) {
points( density( ei[,i], na.rm = TRUE ), type = "l", col =
rainbow(20)[i] )
}
```

If you wish you can make a pdf file, which will end up in your working directory:

```
dev.copy2pdf(file = "DensityPlots.pdf")
```

As a reference you may plot the un-normalized data like this:

```
hist(HIV.affyb, main="Un-normalized")
```

Questions

Q1: What do you see in this plot and does the data look normalized?

Q2: Could the data be more comparable?

Singular Value Decomposition and outlier assessment

In order to evaluate the quality of microarray data there are many things that can be done. One such thing is the density plots, that you've just made ... If a chip is really bad this may show up as a weird distribution of intensities. Another good way of evaluating your data is to do a Singular Value Decomposition (SVD), which is kind of the same as a Principal Component Analysis (PCA), just designed to take matrices that have many more rows than colomns (or the other way around). An SVD analysis kind of breaks your data down into X components (where X = number of samples). The first component contains the most variation, and the last the least (none).

Do the calculations:

```
m <-  ncol(ei)
X <- ei - rowMeans(ei)
svd <- svd (X)
V <- svd$v
S <-  matrix(0 , m, m)
diag(S)<-svd$d
```

Lets make some plot-able sample names and category colors

```
pnames <- paste(design$HIV, design$cell, rep(1:5, 4), sep="-")

cat_col <- rep(c("blue","darkgreen","red","orange"), each=5)
```

Plot the heatmap of all components and a scatter plot with the first two components:

```
par(mfrow = c(1,2))
image( S%*%t(V), main="Singular Value Decomposition",
ylab = "Samples ( 1-20 )", xlab = "The 20 Components" )

plot(V[,1], V[,2], col = "white", main = "Singular Value
Decomposition",
xlab = "First Component", ylab = "Second Component")
text(V[,1], V[,2], pnames, col = cat_col )

dev.print ( device = pdf, file = "SVD.pdf" )
```

Q3: Do you see any systematic patterns? And what do you think this means?

Q4: Which two components contain the most information in regard to your categories?

The numbers in, *V[,1]* and *V[,2]*, in the plot function (and in the text function) are the compo- nents, that we plot (the first and the second). You could try to plot the first component against another component. There may be one that you think would tell us more about the 4 categories than the second component does.

The *t*-tests:

We will now try to see if we can find any genes that are significantly differentially expressed be- tween control and HIV-infected cells. In order to do this we will do a *t*-test, therefore you will need to load the genefilter package

`require(genefilter)` # if you don t have it , you need to install it :

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

Do the *t*-test for each cell type like this:

```
t.pval.CTL <-
rowttests(ei[,design$cell=="CTL"],design$HIV[design$cell=="CTL"])
$p.value

t.pval.Th  <-
rowttests(ei[,design$cell=="Th"],design$HIV[design$cell=="Th"])$p
.value
```

```
names(t.pval.Th)<-names(t.pval.CTL)<-rownames(ei)
```
to get the gene names

The above function uses factors to group the data. An alternative way to do this (the CTL t-test) could be:

rowttests(ei[,1:10], factor(c(1,1,1,1,1,2,2,2,2,2))) # 1 being the controls

Look at the top 10 p-values from each of the two t-tests:

```
sort(t.pval.CTL)[1:10]
sort(t.pval.Th)[1:10]
```

As you can see all our genes have names like "214057_at" - these are *affy* identifiers. In order to get the gene names you should do the commands below, and try to take at look at the top genes again (as above).

require(annotate)
```
library( hgu133a.db)
```

if you have it there is no need to install it again

```
source("http://bioconductor.org/biocLite.R")
biocLite("hgu133a.db")
```

```
AffyID <- rownames( ei )
GeneNames <- getSYMBOL( AffyID , "hgu133a" )
### ne marche pas sur les machines de la fac
```

To look up the gene names of the most significant genes:

```
GeneNames[names(sort(t.pval.CTL)[1:10])]
GeneNames[names(sort(t.pval.Th)[1:10])]
```

It is of course nice to know which genes that are on top of the list, but we would also like to know how these genes are affected (up or down)? We therefore calculate the log2 to the fold- change.

```
fc.CTL<- rowMeans(ei[,design$cell=="CTL" & design$HIV=="HIV"]) -
rowMeans(ei[,design$cell=="CTL" & design$HIV=="ctrl"])
fc.Th <- rowMeans(ei[,design$cell=="Th" & design$HIV=="HIV"]) -
rowMeans(ei[,design$cell=="Th" & design$HIV=="ctrl"])
```

Take a look at the fold changes of the 30 most significant genes:

```
fc.CTL[ order(t.pval.CTL) ][1:30]
fc.Th[ order(t.pval.Th) ][1:30]
```

Lets visualize

The top 30 in a heatmap (yellow: High, red: Low - gene expression):

```
Th_Top<-ei[ order(t.pval.Th)[1:30],  design$cell=="Th"]
rownames(Th_Top)<-GeneNames[order(t.pval.Th)[1:30]]
x11()
heatmap(Th_Top , main = "Th-cells", scale="none", mar=c(8,5))

CTL_Top<-ei[ order(t.pval.CTL)[1:30],  design$cell=="CTL"]
rownames(CTL_Top)<-GeneNames[order(t.pval.CTL)[1:30]]
x11()
heatmap(CTL_Top, main = "CTL-cells", scale= "none", mar=c(8,5))
```

t-test

Questions

Q5: What are we testing in these experiments?

Do you find the same genes on the two lists (the top 30)? hint: `intersect(x,y)` # where x and y could be two lists of gene names And if so: Are the genes behaving in the same way (up or down)? hint: Find them in the heat- map.

To compare you could also plot p-values against p-values using the plot function. An example could be:

```
#plot(p.val.x, p.val.y, log="xy")
```

Q6: How many genes have we tested? hint: `length(ei[,1])`

Q7: At which p-value would you apply the cutoff if you were to use the Bonferroni correction at a 0.05 level?

Q8: How many genes would we trust in each of the two experiments using this cutoff? hint: `min(t.pval.Th <= the bonferoni cutoff )`

Q9: Which correction for multiple testing is most strict, Bonferroni or Benjamini-Hochberg? (hint BH: `P = i*0.05/N`)

Q10: If you choose to use the Benjamini-Hochberg cutoff in the "Th" test, how many genes will you then find to be significant?

Volcano Plot

A Volcano plot shows the connection between the p-values and the log2 of the fold change, com- pared to the same analysis of the same data ... randomized. The permutation should be

a balanced randomization of the columns of the experiment. After the permutation we run the *t*-test again and calculate a new log2 fold change (M) value. Finally, we plot the two outputs on top of each other in order to compare.

So, in order to make a Volcano Plot we need to obtain the random or permuted p-values for each data set. To get those you need to perform a *t*-test on the same data, but with permuted labels. Similarly for the "permuted" fold-change. Finally we will plot the "permuted" p-values vs per- miuted fold changes (red dots) on top of the "real" data.

```
random_lables<-sample(design$HIV[design$cell=="CTL"])
```

Take a look at *random_lables* to see how the data was randomized. Compare to the original de- sign. If your random selection is not approximately balanced try again (**Q** can it be perfectly bal- anced?).

Alternatively you may write a balanced "random" design like this:

```
random_lables<-factor(c("ctrl","HIV","ctrl","HIV","ctrl",
"HIV","c-
trl","HIV","ctrl","HIV"))
```

Calculate the log2 foldchange and the p-value using the random ordering:

```
perm_t.pval.CTL <- rowttests(ei[,design$cell=="CTL"],
random_lables)$p.value
perm_fc.CTL<- rowMeans(ei[,design$cell=="CTL" &
random_lables=="HIV"])  -
rowMeans(ei[,design$cell=="CTL" & random_lables=="ctrl"])
```

The Volcano Plots:

```
plot(fc.CTL, t.pval.CTL, main = "Volcano Plot\nCTL", log = "y",
xlab = "M
(log2 fold change)", ylab = "p-value", pch = 20, col = "blue")
points(perm_fc.CTL, perm_t.pval.CTL, type = "p", pch = 20, col =
"red")
legend("topleft", col=c("red","blue"), legend=c("perm", "real"),
pch=20,
bg="white")

dev.print( device = pdf, file = "VolcanoPlots.pdf" )
```

Q11: Do you think that your cutoff from Q5 and Q7 were reasonable?

Normally one should make a couple of Volcano plots using balanced randomization's in order to decide where to put your cutoff/ or estimate FDR, but for now one is fine.

Q12: Where do you think the best cutoff would be?

Q13: Do you think that these plots/calculations could somehow be used to estimate the False Discovery Rate? How would you do that?

Q14: Are we now able to answer the first biological question?

1. Which genes are expressed differently between HIV-infected and control cells?

The Two-Way ANOVA:

In order to answer the second biological question we will proceed to the next the Two-Way ANOVA.

The two-way ANOVA uses all the data in a single test. This enables us to not only find the genes that are expressed differently in infected compared to healthy cells, but also the genes that are expressed differently between the two cell types. On top of this we will also get a list of genes that exhibit an interaction effect between the two factors. Herein genes that are found to behave different exclusively in one of the two cell types. We therefore get three p-values for each gene - which each contain the possibility of rejecting the null hypothesis of one of the three hypotheses (disease, cell type, interaction), which again enable us to answer the three questions that we asked in the beginning ... in a single analysis!

Run the two-way ANOVA by the use of the following R-commands:

```
anova.2F <- function(vector, F1, F2, design.df){
y<-anova(aov(vector ~ design.df[,F1]*design.df[,F2]))$Pr[1:3]
return(y)
}

ANOVA.pval <- apply(ei, 1, anova.2F, F1="HIV", F2="cell", design)
# this command may take too long ( 20-30 min) ... perhaps you should start to write answer from the previous questions
```

Extract the three lists of p-values:

```
HIV.pval <- ANOVA.pval[1,]
CellType.pval <- ANOVA.pval[2,]
Interaction.pval <- ANOVA.pval[3,]
```

Check out the top 20's:

```
GeneNames[ order(HIV.pval)[1:20] ]
GeneNames[ order(CellType.pval)[1:20] ]
GeneNames[ order(Interaction.pval)[1:20] ]
```

Make a heatmap of the top 30 from the *HIV* result:

```
HIV_Top<-ei[ order(HIV.pval)[1:30],]
rownames(HIV_Top)<-GeneNames[order(HIV.pval)[1:30]]
heatmap(HIV_Top, scale="none")
dev.print ( device = pdf, file = "Heatmap.pdf")
```

Questions

Q15: Are the top p-values higher or lower than top p-values from the *t*-tests?

(Hint: min(HIV.pval))

Q16: How many genes are significantly affected when using the Bonferroni correction at 0.05?

Please, answer Q14 again ...

Q17: Would you think that the two-way ANOVA is more powerful than the *t*-tests?

The Biological Interpretation

Pick genes and check out what they do (hint: Do a google or look it up in OMIM).

Q18: Do you find any significantly affected genes, that you would suspect may play a role during the acute HIV infection?

Q19: Which genes would you choose to investigate further?

Q20: How would you verify these findings? ... crazy-wild ideas are accepted at this stage ...