Final Project

Microarray Series GSE48558 Differential Expression Analysis

Sharif University of Technology

Student: Hossein Jafarinia(400211557)

Professors: Dr. Somayyeh Koohi, Dr. Ali Sharifi-Zarchi

An Introduction to Microarray

A microarray is a laboratory tool used to detect the expression of thousands of genes at the same time. DNA microarrays are microscope slides that are printed with thousands of tiny spots in defined positions, with each spot containing a known DNA sequence or gene. Often, these slides are referred to as gene chips or DNA chips. The DNA molecules attached to each slide act as probes to detect gene expression, which is also known as the transcriptome or the set of messenger RNA (mRNA) transcripts expressed by a group of genes.

To perform a microarray analysis, mRNA molecules are typically collected from both an experimental sample and a reference sample. For example, the reference sample could be collected from a healthy individual, and the experimental sample could be collected from an individual with a disease like cancer. The two mRNA samples are then converted into complementary DNA (cDNA), and each sample is labeled with a fluorescent probe of a different color. For instance, the experimental cDNA sample may be labeled with a red fluorescent dye, whereas the reference cDNA may be labeled with a green, fluorescent dye. The two samples are then mixed together and allowed to bind to the microarray slide. The process in which the cDNA molecules bind to the DNA probes on the slide is called hybridization. Following hybridization, the microarray is scanned to measure the expression of each gene printed on the slide. If the expression of a particular gene is higher in the experimental sample than in the reference sample, then the corresponding spot on the microarray appears red. In contrast, if the expression in the experimental sample is lower than in the reference sample, then the spot appears green. Finally, if there is equal expression in the two samples, then the spot appears yellow. The data gathered through microarrays can be used to create gene expression profiles, which show simultaneous changes in the expression of many genes in response to a particular condition or treatment. [1]



Figure 1. Microarray filles holes [1]

Introduction to Differential Expression Analysis

There are many databases for microarray datasets and one of the most popular ones is Genome Expression Omnibus (GEO) database [2] and their microarray information can be used and analyzed with different methods and one of the most popular ones is Differential Expression analysis which can be done with different computer programs and one of the most popular ones is with R programming language. The method is generally composed of two parts. First is quality control the second is the analysis itself. More details are as follows:

Quality Control:

Quality control should always be done because there may be cases that the experiment is not done with good care or as precise as it should and many different research can be done based on the wrong results of this experiment. For example, if the samples aren’t cleaned from a certain bacteria called mycoplasma, we essentially will analyze the bacteria’s gene too which gives us wrong results.

In this part we should make some plots like box plot on samples gene data and see if the data are reasonable or not. For example, is there is sample that its box plot is highly different than other it probably had a lot more mRNA it’s better to delete that sample from analysis. And before all these the data should have been normalized (for example with quantile normalizing method) and scaled in in log2 factor so results be more meaningful and realistic.

Dimensionality Reduction:

This part is essentially another part of the quality control. Usually there are many genes being experimented in these kinds of experiments. For example is this microarray experiment there are 32321 genes experimented so our data is going to be 32321 dimension which is impossible for human to be comprehend so we need to do dimensionality reduction with methods like Principal Component Analysis (PCA) which can show us how clustered and differentiated are our data and if they match the overall design section of dataset explanation we can be sure of two things. First, we can be sure that meaningful genes are selected to be experimented and secondly the experiment is done well. This is the most important part of quality control.

Correlation Analysis:

This part can still be considered another part of quality control. In this part we should compute the probabilistic correlation between samples gene data then cluster them see if the clusters are meaningful or not. For example, aside from each sample with itself Normal samples should have the highest correlation with each other than AML samples should have more but less than Normal samples because a tissues cancer cells are bit different from each other.

Differential Expression Analysis:

This is the most important part. In this part based different methods we should analyze if the higher expression of the more expressed genes and lower expression of less expressed genes has meaningful effect on AML or not. We generally must define a test statistic and based on that test statistics calculate the P. Value. In biological studies if P. Value is < 0.05 or < 0.02 we consider it meaningful. Secondly based on biological studies we need to calculate logFC which tells us how many folds the higher expression is. And based on our knowledge in biological studies we want it to be at least > 1 for higher expressed genes and < -1 for lower expressed genes.

Pathway and Gene Ontology Analysis:

After finding the meaningful and valuable genes in the disease we can use the studies and analysis of databases like Enrichr [3]. We usually choose ~250-300 genes and give those gene names to Enrichr [3] and after their analysis we can find gene ontology and pathways related to our gene list.

Dataset

The dataset is Series GSE48558 [4]

Status: Public on Jul 06, 2013

Title: Expression data from normal and Malignant hematopoietic cells

Organism: Homo sapiens

Experiment type: Expression profiling by array

Summary: This data was used to determine levels of BRCA1 and BRCA2 in primary human leukemia samples. Samples were determined to be high BRCA1 and/or BRCA2 or low BRCA1 and/or BRAC2.

This data was used to determine levels of BRCA1 and BRCA2 in primary human leukemia samples. Samples were determined to be high BRCA1 and/or BRCA2 or low BRCA1 and/or BRAC2.

Overall design: AML cell lines and patient samples, B ALL cell lines and Patient samples, T ALL cell lines and patient samples, normal B cells, normal granulocytes, normal monocytes, normal T cells and normal CD34+ cells were used for RNA extraction and hybridization on Affymetrix microarrays. All the AML, B ALL, T ALL cell lines were cultured in vitro under appropriate culture conditions and harvested in their log phase growth for RNA extraction. AML, B ALL, and T ALL patient samples were collected...(I assume these are the PBMCs from either peripheral blood or bone marrow from patients, please confirm). Normal B cells, granulocytes, monocytes, and T cells were purified from human peripheral blood of normal healthy donors.

Contributor: Civin CI, Civin CI

Submission date: Jul 05, 2013

Last update date: Jul 26, 2018

E-mails: ysun@som.umaryland.edu

Organization name: U. of Maryland, Baltimore

[4]

Method

Quality Control:

Based on the GSE48558Analysis.r which is available both in the end of this report and in GSE48558Analysis.r file I have checked max and min values for gene expression values which are respectively 13.76154 and 1.611473 tells us data is in log2 scale and the resulted boxplot in Figure 2 shows that the data are normalized and experimented well and there are no unusual outliers on samples.

Chart

Description automatically generated

Figure 2. Box plot for each samples genes expression

Dimensionality Reduction:

For dimensionality reduction I used PCA method and since the best thing we can show on 2D monitor I just used PC1 and PC2. The code for these two parts can be found in GSE48558Analysis.r.

Figure 3 shows the plot for genes. In this one an extra job of expression value – mean is also done so we get more meaningful result of difference in gene expression and not just expression itself.

Chart, scatter chart

Description automatically generated

Figure 3. Genes PCA plot. Each Circle is a gene

Figure 4 shows the plot for samples. As we can see there are generally five clusters of Normal samples and that matches the experiment description’s overall design which says there are five groups of normal samples.

Chart, scatter chart

Description automatically generated

Figure 4. Samples PCA plot. Each circle is a sample

Correlation Analysis:

I calculated the correlation between each sample and clustered more close ones and plotted the values in a heat map as we can see in GSE48558Analysis.r and the result is in Figure 5. As we can see Normal samples have the highest correlation, Normal samples have the lowest correlation with AML samples and AML samples have some amounts of correlation as we know from biological studies that even cancer cells of a specific tissue can have some levels of difference with each other.

Chart

Description automatically generated

Figure 5. Correlation Heat Map

Differential Expression Analysis:

In this part I used library limma which uses a really complex line fitting method and uses a test statistic named B to calculate P. Values and for the final gene list only took the ones with lower than 0.05 P. Value and > 1 or < -1 logFC(AML / Normal). The code for it can be found in GSE48558Analysis.r and most important genes which are adj.P.Value < 0.05 and logFC > 1 -> Up Genes adj.P.Value < 0.05 and logFC < -1 -> Down Genes can be found in Results folder named AMLDownGenes.txt and AMLUpGenes.txt respectively. Note that some genes are different names of the same gene so we can find in any gene database with ease.

Top 10 of these some of these genes can be seen in table 1.

Graphical user interface

Description automatically generated with low confidence

Table 1. Effective genes with their adj.P.Value and logFC

Pathway and Gene Ontology Analysis:

I took ~250-300 top genes of the Up Genes list and submit that to Enrichr [3] and chose the most meaningful ones with adj.P.Value < 0.05.

here are the top 10 results of Pathway analysis from Wikipathway 2021 Human (the full list can be found at WikiPathway\_2021\_Human\_table.txt (the full list can be found at WikiPathway\_2021\_Human\_table\_up/down.txt):

Table

Description automatically generated

Table 2. Top 10 meaningful UP results from WikiPathway 2021 Human

Table

Description automatically generated

Table 3. Top 10 meaningful Down results from WikiPathway 2021 Human

It is also worth to note results from Kegg 2021 Human (the full list can be found at KEGG\_2021\_Human\_table\_up/down.txt):

Graphical user interface

Description automatically generated with medium confidence

Table 4. Top 10 meaningful UP results from KEGG 2021 Human

Table

Description automatically generated

Table 5. Top 10 meaningful Down results from KEGG 2021 Human

And results from Reactom 2016 (the full result can be found at Reactome\_2016\_table\_up/down.txt):

Table

Description automatically generated

Table 6. Top 10 meaningful UP results from Reactom 2016

Table

Description automatically generated

Table 7. Top 10 meaningful Down results from Reactom 2016

Here are the top 10 results of Ontology Analysis from Go Cellular Component (the full result can be found at GO\_Cellular\_Component\_2021\_table\_up/down.txt):

Table

Description automatically generated

Table 8. Top 10 meaningful UP results from GO Cellular Component 2021

Graphical user interface, text, application

Description automatically generated

Table 9. Top 10 meaningful Down results from GO Cellular Component 2021

It is worth to note results from Go Biological Process 2021(the full result can be found at GO\_Biological\_Process\_2021\_table\_up/down.txt):

Table

Description automatically generated

Table 10. Top 10 meaningful UP results from GO Biological Process 2021

Table

Description automatically generated

Table 11. Top 10 meaningful Down results from GO Biological Process 2021

Also, worth to note results from Go Molecular Function 2021(the full result can be found at GO\_Molecular\_Function\_2021\_table\_up/down.txt):

Table

Description automatically generated

Table 12. Top 10 meaningful UP results from GO Molecular Function 2021

Table

Description automatically generated

Table 13. Top 10 meaningful Down results from GO Molecular Function 2021

And also, worth to note the results form MGI Mammalian Phenotype Level 4 2021 (the full result can be found at MGI\_Mammalian\_Phenotype\_Level\_4\_2021\_table\_up/down.txt):

Table

Description automatically generated

Table 14. Top 10 meaningful UP results from MGI Mammalian Phenotype Level 4 2021

Table

Description automatically generated with medium confidence

Table 15. Top 10 meaningful Down results from MGI Mammalian Phenotype Level 4 2021

Conclusion

Although this data has been analyzed before but in a hypothetical situation, I can confidently say that I did Differential Expression Analysis on a type of leukemia and found highly effective genes in which have significant effects on the disease and this result can be shared with Physicians, Biologists, Pharmacists, … and we may be able to cure this disease with methods like inhibit some of these pathways. Although in real situations I’m actually able to analyze new data too.

Extra Work

I- Explanation of biological processes underneath the results.

For pathways:

A biological pathway is a series of interactions among molecules in a cell that leads to a certain product or a change in a cell. Such a pathway can trigger the assembly of new molecules, such as a fat or protein. Pathways can also turn genes on and off or spur a cell to move [5].

In this case each up genes make a certain kind of protein and that protein can cause increase or decrease in some other genes expression and protein synthesis, and all of this will cause malfunction in a way that the patient becomes an AML patient. For example, as one of the biggest and most important ones Retinoblastoma gene in cancer WP2446 [6] is a tumor suppressor protein that is dysfunctional in several major cancers. One function of pRb is to prevent excessive cell growth by inhibiting cell cycle progression until a cell is ready to divide [7] and this pathway is dysfunctional in our specific cancer. We can see this pathway in Figure 6 and some genes like TOP2A and PCNA in the pathway.

Diagram

Description automatically generated

Figure 6. Retinoblastoma gene in cancer (Homo sapiens)

For down genes there are generally two cases either for random or external reasons for example radiation the expression of some genes have been decreased and/or increase in expression of some genes caused this decrease in expression and they also have pathway.

For gene ontologies:

The Gene Ontology is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species [8]. It has three major sections which I’m going to explain them and their relation to our study

GO Cellular Component:

In this part we are concerned about the cellular component that are involves in this disease. For example, as a very significant case CMG complex (GO:0071162) is A protein complex that contains the GINS complex, Cdc45p, and the heterohexameric MCM complex, and that is involved in unwinding DNA during replication [9]. And we all know that cancer is highly related to cell division and DNA unwinding.

GO Biological Process:

In this part we are concerned about the biological processes that are involved in this disease. For example, as a very significant case we have mitotic spindle elongation (GO:0000022) is the cell cycle process in which the distance is lengthened between poles of the mitotic spindle. Mitotic spindle elongation begins during mitotic prophase and ends during mitotic anaphase B [10]. We all know that cancer is highly related to cell division and DNA unwinding.

GO Molecular Function:

In this part we are concerned about the molecular function that are involved in this disease. For example, as a very significant case as a very significant case DNA replication origin binding (GO:0003688) is binding to a DNA replication origin, a unique DNA sequence of a replicon at which DNA replication is initiated and proceeds bidirectionally or unidirectionally [11].

II- Transcription Analysis

Here I analyzed the results of transcription factors related the found genes (the full result can be found at TRANSFAC\_and\_JASPAR\_PWMs\_table\_up/down.txt)

Table

Description automatically generated



Table 16. Top 10 meaningful UP results from TRANSFAC and JASPAR PWMs

Text, table

Description automatically generated with medium confidence

Table 17. Top 10 meaningful down results from TRANSFAC and JASPAR PWMs

III- Diseases/Drugs Analysis

Here I analyzed the results of diseases/drugs related the found genes (the full result can be found at LINCS\_L1000\_Ligand\_Perturbations\_up\_table\_up/down.txt)

IV- Cell Types Analysis

Here I analyzed the results of cell types related the found genes (the full result can be found at Human\_Gene\_Atlas\_table\_up/down.txt)

III- Literature Review

To check and see how valuable our results are we

Code

This is the code (GSE48558Analysis.r):

# Version info: R 3.2.3, Biobase 2.30.0, GEOquery 2.40.0, limma 3.26.8

################################################################

#   Differential expression analysis with limma

#+ Hossein Jafarinia modifications on R4.1.1 and Rstudio

#necessary libraries

library(GEOquery)

library(limma)

library(umap)

library(pheatmap)

library(ggplot2)

library(gplots)

library(reshape2)

library(plyr)

library(Biobase)

library(ggplot2)

library(reshape2)

library(plyr)

library(dplyr)

#Changing default VROOM\_CONNECTION\_SIZE so we can capture data

Sys.setenv("VROOM\_CONNECTION\_SIZE" = 131072 \* 3)

#Making filing easy

curD <- dirname(rstudioapi::getActiveDocumentContext()$path)

setwd(sub(paste0("/", sub("(.+)/", "", curD)), "", curD))

#load series and platform data from GEO

series = "GSE48558"

platform = "GPL6244"

gset <- getGEO(series, GSEMatrix =TRUE, AnnotGPL=TRUE, destdir = "Data/")

if (length(gset) > 1) idx <- grep(platform, attr(gset, "names")) else idx <- 1

gset <- gset[[idx]]

#make proper column names to match toptable

fvarLabels(gset) <- make.names(fvarLabels(gset))

# group membership for all samples

# gsms <- paste0("1111111111111XXXXXXXXXXXXXXXXXXXXXXXXXX10XX10XX1X1",

               # "1X11X11XX1XX1XX1XX0X01XX0X0000X01X001X0010X010X010",

               # "XX10XX10XX1XXXXXXXXXXXXXXXXXXXXXXXXXX0000000110111",

               # "00000000000000000000")

gsms <- paste0("1111111111111XXXXXXXXXXXXXXXXXXXXXXXXXXX0XXX0XXXXX",

               "XXXXXXXXXXXXXXXXXX0X0XXX0X0000X0XX00XX00X0X0X0X0X0",

               "XXX0XXX0XXXXXXXXXXXXXXXXXXXXXXXXXXXXX0000000110111",

               "00000000000000000000")

sml <- strsplit(gsms, split="")[[1]]

# filter out excluded samples (marked as "X")

sel <- which(sml != "X")

sml <- sml[sel]

smlWithName <- recode(sml, "1" = "AML", "0" = "Normal")

gset <- gset[ ,sel]

ex <- exprs(gset)

max(ex)

min(ex)

#As we can see data or in log2

#Quality Control:

#Drawing Boxplot for quality control

gs <- factor(smlWithName)

ord <- order(gs)  # order samples by group

palette(c("#1B9E77", "#7570B3", "#E7298A", "#E6AB02", "#D95F02",

          "#66A61E", "#A6761D", "#B32424", "#B324B3", "#666666"))

par(mar=c(7,4,2,1))

title <- paste ("GSE48558", "/", annotation(gset), sep ="")

boxplot(ex[,ord], boxwex=0.6, notch=T, main=title, outline=FALSE, las=2, col=gs[ord])

legend("topleft", groups, fill=palette(), legend = c("AML", "Normal"), bty="n")

corex = cor(ex)

pheatmap(corex, labels\_row = smlWithName, labels\_col = smlWithName, border\_color = NA, main = "Smaples Correlation Heat Map") #As we can see Normal samples of different groups have highest correlation with each other and the lowest with AML samples and although AMS samples have some level of correlation with each other its not as much as normal samples due to the fact that cancer cells have some levels of difference between each other

#Dimensionality reduction

#PCA for genes

ex.scale <- t(scale(t(ex), scale = F)) #ex - mean(ex)

pc <- prcomp(ex.scale) #Finding PCAs

plot(pc, main = "PCA", xlab = "PCs")

#only PC1 and PC2 are sufficient and thats generally what our 2 dimensional page can show best

plot(pc$x[,1:2], main = "Genes PCA") #As we can see genes have a relatively meaningful distribution

#PCA for samples

pcr <- data.frame(pc$r[, 1:3], Group = smlWithName) #r means rotation we at least mean 3 pcs

ggplot(pcr, aes(PC1, PC2, color = Group)) + geom\_point(size = 3) + theme\_bw() #As we can wee AML and normal cells are generally seperated well and we know from the examination summary says there are 5 types of Normal cells and we can see 5 clusters of normal cells in our plot which makes us conclude that the quality is good

#Differential Expression Analysis:

#assign samples to groups and set up design matrix

gset$group <- smlWithName

design <- model.matrix(~group + 0, gset) #finds pset

colnames(design) <- levels(gs) #design

#differential analysis main part by limma

fit <- lmFit(gset, design)  # fit linear model its a linear  model that fits a line to each model and based on the difference between line it say how much different they are

#set up contrasts of interest and recalculate model coefficients

cont.matrix <- makeContrasts(contrasts="AML-Normal", levels=design)

#cont.matrix and we see it wants to compare tumor and normal

fit2 <- contrasts.fit(fit, cont.matrix) #put the output in fit2

#compute statistics and table of top significant genes

fit2 <- eBayes(fit2, 0.01) #bayesian prior of 1% cancer related genes from biological studios

tT <- topTable(fit2, adjust="fdr", sort.by="B", number=Inf) # the function that calculated p.values and logFC "fdr" is our method for adjutment "B" is limmas test statistics and Inf for all genes

tT = subset(tT, select=c("Gene.symbol", "Gene.ID", "adj.P.Val", "logFC")) # a table of the columns we actually care about

write.csv(tT,"Results/tT.csv", row.names = FALSE)

#the genes that are expressed more in cancer samples(at least 2 times) their effect is meaningful

tT.Up.Gene <- subset(tT, adj.P.Val < 0.05 & logFC > 1)

upGenes = unique(tT.Up.Gene$Gene.symbol)

AML.Up.Genes.AllNames <- unique(as.character(strsplit2(upGenes, "///")))

write.table(AML.Up.Genes.AllNames, "Results/AMLUpGenes.txt", row.names = F, col.names = F, quote = F)

#the genes that are expressed less in cancer samples(at least 2 times) their effect is meaningful

tT.Down.Gene <- subset(tT, adj.P.Val < 0.05 & logFC < -1)

downGenes = unique(tT.Down.Gene$Gene.symbol)

AML.Down.Denes.AllNames <- unique(as.character(strsplit2(downGenes, "///")))

write.table(AML.Down.Denes.AllNames, "Results/AMLDownGenes.txt", row.names = F, col.names = F, quote = F)