Identification and Interpretation of Differentially Expressed Genes on GSE56481

Name: Kaiyi DENG Student ID: A0251098B
Name: Louis Low Ching Yi Student ID: A0251491H

Introduction

In this assignment, we selected the GSE56481 data set that contains 18 samples. From the study, FACs were used to sort CD4+, CD8+ single-positive and CD4+CD8+ double-positive T-cells derived from 3 GPA(granulomatosis with polyangiitis) patients and 3 healthy controls. Thus, there will be CD4+CD8+ double positive T-cells derived from each of the 6 subjects, likewise the same is done for the CD4+ and CD8+ single-positive T-cells which results in 18 samples in total. The transcript data from these 18 samples were analyzed to identify the differentially expressed genes (DEGs). We used two executable R codes to identify our DEGs in this CA1 assignment. In the first analysis(1st code), DEGs are obtained by comparing the transcript expression levels of T- cells from healthy controls to GPA patients. In the second analysis (2nd code), the transcript expression levels of the different types of T- cells from GPA patients are compared to the corresponding healthy controls to obtain the DEGs.

Overview

Our analysis consists of two parts (two separate R codes): first is the basic comparison of gene expression levels between GPA patients and healthy controls. The 2nd part also involves the comparison of gene expression levels between GPA patients and healthy controls, but it's performed in 3 different T-cell types (CD4+CD8+, CD4+ and CD8+). Most of our result evaluation will be for the second part (2nd code) in this assignment.

Data Extraction

We used getGEO() to extract the Expression set of GSE56481, and we used the supplementary file in the phenodata as reference in read.celfiles() to load all the cel files in the order as the experimental data. We are then able to extract the data of interest by accessing their labels.

Data Processing

We use rma() to do data processing, and we created a contrast between data from GPA patients and healthy controls. We then used functions in limma to construct a linear model.

Annotation and Differentially Gene Expression Identification

The original annotation dbi used in GSE56481 is pd.hugene.2.0.st. However, we found that it cannot be used for annotation. Hence, we found its parental dbi, hugene20sttranscriptcluster.db, to annotate our data [4]. In the R script, we used AnnotationDbi::select() to retrieve our significant up-regulated and down-regulated genes.

Part 1

Differential expression of genes between GPA patients and healthy controls

Fitted Model

```
> summary(decideTests(fitted.ebayes[, "Control"], lfc=1))
       Control
Down       449
NotSig   52970
Up         198
```

```
> contrast_matrix
         Contrasts
Levels    GPA - Control Control - GPA
  GPA                 1            -1
  Control            -1             1
```

```
> summary(decideTests(fit2, lfc=1))
       GPA - Control Control - GPA
Down             198           449
NotSig         52970         52970
Up               449           198
```

Annotation

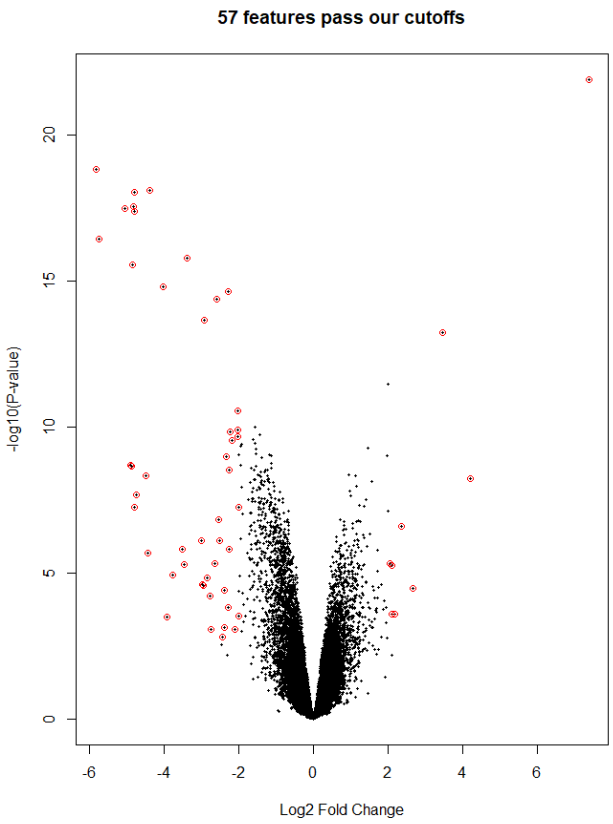
```
> head(keys(hugene20sttranscriptcluster.db, keytype="PROBEID"))
[1] "16650001" "16650003" "16650005" "16650007" "16650009" "16650011"
> AnnotationDbi::select(hugene20sttranscriptcluster.db, ps,
+                       c("SYMBOL", "ENTREZID", "GENENAME"), keytype="PROBEID")
'select()' returned 1:1 mapping between keys and columns
    PROBEID SYMBOL ENTREZID                                                    GENENAME
1  17112149   XIST     7503                                  X inactive specific transcript
2  17116977    UTY     7404 ubiquitously transcribed tetratricopeptide repeat containing, Y-linked
3  17117126  KDM5D     8284                                                lysine demethylase 5D
4  17116384 TXLNGY   246126                                 taxilin gamma pseudogene, Y-linked
5  17118451 TXLNGY   246126                                 taxilin gamma pseudogene, Y-linked
6  17116194 TTTY15    64595                              testis-specific transcript, Y-linked 15
7  17116200  USP9Y     8287                              ubiquitin specific peptidase 9 Y-linked
8  17116251  DDX3Y     8653                                        DEAD-box helicase 3 Y-linked
9  17116050   PRKY     5616                           protein kinase Y-linked (pseudogene)
10 17115971   <NA>     <NA>                                                            <NA>
```

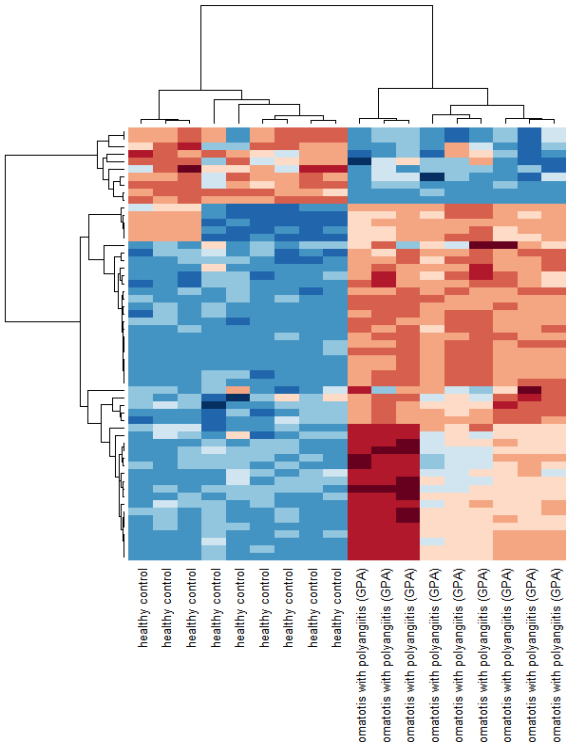
The select() interface is used to select objects from the hugene20sttranscriptcluster.db annotation package. We used it to extract the symbols and gene names to annotate our DEGs [4].

Up-regulated and Down-regulated genes (The whole list can be accessed in the R script.)

```
> head(dplyr::mutate(df_up,GENENAME=stringr::str_trunc(GENENAME,30)))
   PROBEID   SYMBOL  ENTREZID                       GENENAME
1 17112149     XIST      7503 X inactive specific transcript
2 17104924     <NA>      <NA>                           <NA>
3 16655389     <NA>      <NA>                           <NA>
4 16691879  PPIAL4G    644591 peptidylprolyl isomerase A ...
5 17104920     TSIX      9383 TSIX transcript, XIST antis...
6 16670359 RNVU1-20 101954268 RNA, variant U1 small nucle...
> head(dplyr::mutate(df_down,GENENAME=stringr::str_trunc(GENENAME,30)))
   PROBEID SYMBOL ENTREZID                      GENENAME
1 17116977    UTY     7404 ubiquitously transcribed te...
2 17117126  KDM5D     8284          lysine demethylase 5D
3 17116384 TXLNGY   246126 taxilin gamma pseudogene, Y...
4 17118451 TXLNGY   246126 taxilin gamma pseudogene, Y...
5 17116194 TTTY15    64595 testis-specific transcript,...
6 17116200  USP9Y     8287 ubiquitin specific peptidas...
```

Data Visualization





From these two graphs, we can see that some genes show very large fold changes with relatively small p-values. In the volcano plot, a cluster of genes appear on the top-left space, which indicates that they are down-regulated in healthy controls (which means up-regulated in GPA patients) and they are statistically significant. This also corresponds to the heatmap.

Part 2

In the second R code/analysis, most of the results describe how the DEGs are obtained from GPA vs healthy controls in CD4+CD8+ double positive T- cells, but the differential expression is performed on all 3 types of T-cells. The R code can also be easily changed to obtain the annotations and diagrams of the DEGs from the other 2 cell types, CD4+ and CD8+.

To identify DEGs in CD4+CD8+ T-cells derived from GPA patients and CD4+CD8+ T-cells derived from healthy controls, we used the makeContrasts function in this 2nd code. The transcript expression data of CD4+CD8+ T-cells from healthy controls is subtracted by the expression data of CD4+CD8+ T-cells from GPA patients (GPA.CD4_CD8positive - Control.CD4_CD8positive). Therefore, a negative log Fold Change value means that the expression of that particular transcript is lower in GPA patients compared to the Control, which means that it is downregulated when compared to the control. Likewise, a positive Fold Change value means that the particular gene expression is higher in GPA patients compared to the Control, which means that it is upregulated when compared to the control. The same Argument is

also performed for the expression data of CD4+ in GPA patients vs CD4+ in healthy controls and CD8+ in GPA patients vs CD8+ in healthy controls using the makeContrasts function.

Next, lmFit is used to fit a linear model for each of the 53617 genes and we assigned it to "gse56481_fit"[1]. After that, we input the "gse56481_fit" into the contrasts.fit function, to re-orientate the fitted models to a contrast design previously set under "contrasts_matrix". An example is that if a given expression of a particular gene is grouped under "GPA.CD4_CD8positive ", it will only be compared to the expression of the same gene grouped under "Control.CD4_CD8positive" for the contrast to be computed [2]. This is followed by using the eBayes function to rank the DEGs using statistical tests [3]. Below is the summary of the results obtained by using the decideTests function after performing the eBayes function. For the decideTests function, we can set the required minimum log2-fold-change to be 1, which means that all DEGs have a minimum of a 2-fold change. The lfc can also be set to 2 to set a higher threshold, and we will therefore obtain fewer significant genes that are upregulated or downregulated [5][8].

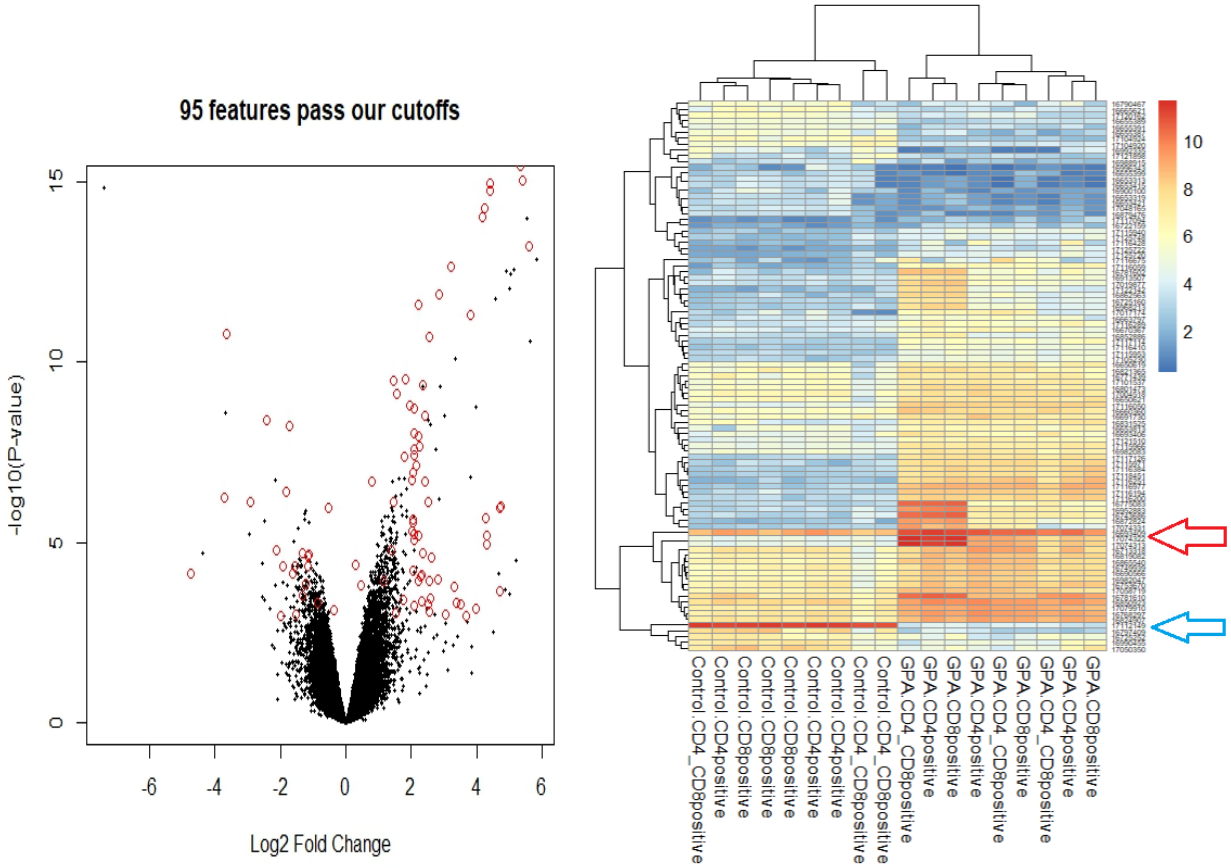
Summary

```
> summary(decideTests(gse56481_fit2, p.value=0.05,lfc=1))
       de_CD4_CD8positive de_CD4positive de_CD8positive
Down                  100             67             42
NotSig              53208          53359          53453
Up                    309            191            122
> summary(decideTests(gse56481_fit2, p.value=0.05,lfc=2))
       de_CD4_CD8positive de_CD4positive de_CD8positive
Down                    8              9              8
NotSig              53565          53581          53584
Up                     44             27             25
```

Furthermore, we used the The topTable function to select our top-ranked genes (lfc= 2), using the data previously generated from the eBayes function. The most significant DEG (smallest P-value) will be placed at the top [6]. This is followed by extracting the data of DEGs only for CD4+CD8+ T-cells. Afterwards, we annotated these top-ranked DEGs using data retrieved from the "hugene20sttranscriptcluster.db" annotation package. The picture below shows some significant DEGs that are either upregulated or downregulated in CD4+CD8+ cells (GPA.CD4_CD8positive - Control.CD4_CD8positive). It shows that PROBEID:17112149 is the most significantly downregulated gene when GPA is compared to control. When we compare the results to Part 1, PROBEID:17112149 is the most significantly upregulated gene because the contrast in Part 1 was set as "Control - GPA". It therefore gave us directly opposite results but it similarly shows that 17112149 expression is higher in healthy control.

```
> head(dplyr::mutate(df_up,GENENAME=stringr::str_trunc(GENENAME,30)))
   PROBEID SYMBOL ENTREZID                        GENENAME
1 17115971   <NA>     <NA>                            <NA>
2 17116977    UTY     7404 ubiquitously transcribed te...
3 17116194 TTTY15    64595 testis-specific transcript,...
4 17116384 TXLNGY   246126 taxilin gamma pseudogene, Y...
5 17118451 TXLNGY   246126 taxilin gamma pseudogene, Y...
6 17116200  USP9Y     8287 ubiquitin specific peptidas...
> head(dplyr::mutate(df_down,GENENAME=stringr::str_trunc(GENENAME,30)))
   PROBEID SYMBOL ENTREZID                        GENENAME
1 17112149   XIST     7503 X inactive specific transcript
2 17104924   <NA>     <NA>                            <NA>
3 17104920   TSIX     9383 TSIX transcript, XIST antis...
4 16655389   <NA>     <NA>                            <NA>
5 16655391   <NA>     <NA>                            <NA>
6 16797409 ZCWPW2   152098 zinc finger CW-type and PWW...
```

Data visualization



The Volcano Plot represents the DEGs of all the 3 T-cell types. Only the red circles represent the significant DEGs (lfc=2) of CD4+CD8+ cells. As expected, most of the red circles fall on the right side of the plot, since the topTable function has previously shown that most of the top-ranked significant genes (lowest adjusted P-values) are upregulated when cells of GPA patients are compared to healthy controls [6]. The Volcano Plot also shows that quite a number of the upregulated DEGs of CD4+CD8+ cells have very significant P-values as compared to all the upregulated DEGs [7].

The Heatmap also shows that more genes are upregulated in GPA samples compared to the control samples (more red/yellow regions on the right side of the heatmap). An example of a gene that is highly upregulated in its GPA samples compared to the controls is PROBEID:17074322 (indicated with the red arrow in the heatmap). The gene name was found to be “defensin alpha”, when we ran the code to show the annotation results in R. Moreover, the heatmap shows that PROBEID:17112149 (indicated with blue arrow) is extremely upregulated in every control sample, thereby similarly showing that ID:17112149 expression is higher in healthy controls compared to GPA.

References

[1]“LmFit: Linear Model for series of arrays,” *RDocumentation*. [Online]. Available: https://www.rdocumentation.org/packages/limma/versions/3.28.14/topics/lmFit. [Accessed: 22-Sep-2022].

[2]“Contrasts.fit: Compute contrasts from linear model fit,” *RDocumentation*. [Online]. Available: https://www.rdocumentation.org/packages/limma/versions/3.28.14/topics/contrasts.fit. [Accessed: 22-Sep-2022].

[3]“Ebayes: Empirical Bayes statistics for differential expression,” *RDocumentation*. [Online]. Available: https://www.rdocumentation.org/packages/limma/versions/3.28.14/topics/ebayes. [Accessed: 22-Sep-2022].

[4]“Hugene20sttranscriptcluster.db,” *Bioconductor*. [Online]. Available: https://bioconductor.org/packages/release/data/annotation/html/hugene20sttranscriptcluster.db.html. [Accessed: 22-Sep-2022].

[5]“DecideTests: Multiple testing across genes and contrasts,” *RDocumentation*. [Online]. Available: https://www.rdocumentation.org/packages/limma/versions/3.28.14/topics/decideTests. [Accessed: 22-Sep-2022].

[6]“Toptable: Table of top genes from linear model fit,” *RDocumentation*. [Online]. Available: https://www.rdocumentation.org/packages/limma/versions/3.28.14/topics/toptable. [Accessed: 22-Sep-2022].

[7]“Volcanoplot: Volcano plot,” *RDocumentation*. [Online]. Available: https://www.rdocumentation.org/packages/limma/versions/3.28.14/topics/volcanoplot. [Accessed: 23-Sep-2022].

[8]“Identifying differentially expressed genes using linear models (part 2, factorial designs),” *Introduction to gene expression microarray analysis in R and Bioconductor: Identifying differentially expressed genes using linear models (part 2, factorial designs)*. [Online]. Available: https://gtk-teaching.github.io/Microarrays-R/07-factorial_designs/index.html. [Accessed: 29-Sep-2022].