Name : Snehal Ashok Kanase

Contact:snehalkanase156@gmail.com

Mobile:7397804023

Title: R for Differential Gene Expression analysis

horizontal line

Answer the following questions below and upload them in the google form.

1. Describe the steps involved in importing the microarray data into R and Bioconductor packages.
2. Explain the code used to filter genes based on expression levels or other criteria.
3. Outline the statistical test used for the differential expression analysis and explain its purpose and limitations.
4. Visualize the differential expression results using heatmaps.
5. Analyze the results of the differential expression analysis. How many genes are significantly differentially expressed by fetal sex? Which genes have the highest fold change?
6. Based on the findings, discuss potential biological implications of the differentially expressed genes in the context of fetal sex and tobacco smoke exposure.

**Answers**

**Que 1.**

**-** Importing microarray data into R and Bioconductor involves several steps.

step-by-step description is following:

**Step 1:**

**Install and Load Bioconductor Manager:**

**-** Before installing Bioconductor packages, first install "BiocManager" package. This package acts as a manager for Bioconductor resources and facilitates the installation of other packages.

**if (!requireNamespace("BiocManager", quietly = TRUE))**

**install.packages("BiocManager")**

**Step 2:**

**Install Bioconductor Packages:**

**-** after installing"BiocManager" now install essential Bioconductor packages for microarray analysis. For Installing following packages "Biobase," "limma," "geneplotter," and "enrichplot." this code is used:

**BiocManager::install(c('Biobase','limma','geneplotter','enrichplot'))**

**Step 3:**

**Install Additional Bioconductor Packages:**

**-** now install some other additional Bioconductor packages which are required. This additional Bioconductor packages "EnhancedVolcano" and "clusterProfiler" are installed using following code:

**BiocManager::install('EnhancedVolcano') BiocManager::install('clusterProfiler')**

**Step 4:**

**Install pheatmap from CRAN:**

**-** In addition to Bioconductor packages, the code installs "pheatmap" from CRAN. This package is commonly used for creating heatmaps**.**

**install.packages('pheatmap')**

**Step 5.**

**Load Required Packages:**

**-** After installing the necessary packages, now I loaded required packages into the R environment. This is doe using the `library` function.

library(Biobase)

library(limma)

library(RColorBrewer)

library(dplyr)

library(ggplot2)

library(geneplotter)

library(pheatmap)

library(enrichplot)

library(tidyr)

library(EnhancedVolcano)

library(clusterProfiler)

after complete all these steps ,then we can proceed with reading our microarray data, performing quality control, normalizing expression values, and conducting differential expression analysis using the tools provided by the loaded Bioconductor packages.

**Question 2:**

The code used to filter genes based on chromosome in the context of gene expression analysis.

Following code is used :

**GSE27272\_noY<GSE27272\_Eset[GSE27272\_Eset@featureData@data$CHR="Y",]**

**Explanation:**

**1. Data Selection:**

**- `**GSE27272\_Eset`: This is represents an ExpressionSet object or a similar data structure containing gene expression data.

**2. Filtering Based on Chromosome:**

**- `**GSE27272\_Eset@featureData@data$CHR != "Y"`: This part of the code filters genes based on the chromosome information stored in the featureData of the ExpressionSet. It selects genes where the chromosome information is not equal to "Y," indicating that genes located on the Y chromosome are excluded from the analysis.

**3. Creation of Filtered ExpressionSet:**

- `GSE27272\_noY <- ...`: The filtered expression set (`GSE27272\_noY`) is created, containing only genes located on autosomes and the X chromosome, as genes on the Y chromosome have been filtered out.

This filtering step or code ensures that the analysis focuses on autosomal and X chromosome genes.

**Que 3:**

Name of the Test:The statistical test used is the moderated t-test as implemented in the `limma` package in R.-Name of the Test: The statistical test used is the moderated t-test as implemented in the `limma` package in R**.**

* **Purpose of the Statistical Test:**

1. dentification of Differentially Expressed Genes:

- The purpose of the moderated t-test is to identify genes that show significant differences in expression levels between different experimental conditions, like according to this data,between females and males.

2. Quantifying Significance:

- The test provides a measure of statistical significance for each gene, in the form of moderated t-statistics and associated p-values.

3. Empirical Bayes Method for Moderation:

- The use of the `eBayes` function applies the empirical Bayes method to squeeze gene-wise residual variance towards a pooled estimate. This helps stabilize variance estimates for genes with low counts, increasing the statistical power of the analysis.

* **Limitations of the Statistical Test:**

1. Sensitivity to Outliers:

- The t-test is sensitive to outliers, and extreme values in the data could influence the results. It's important to preprocess the data correctly.

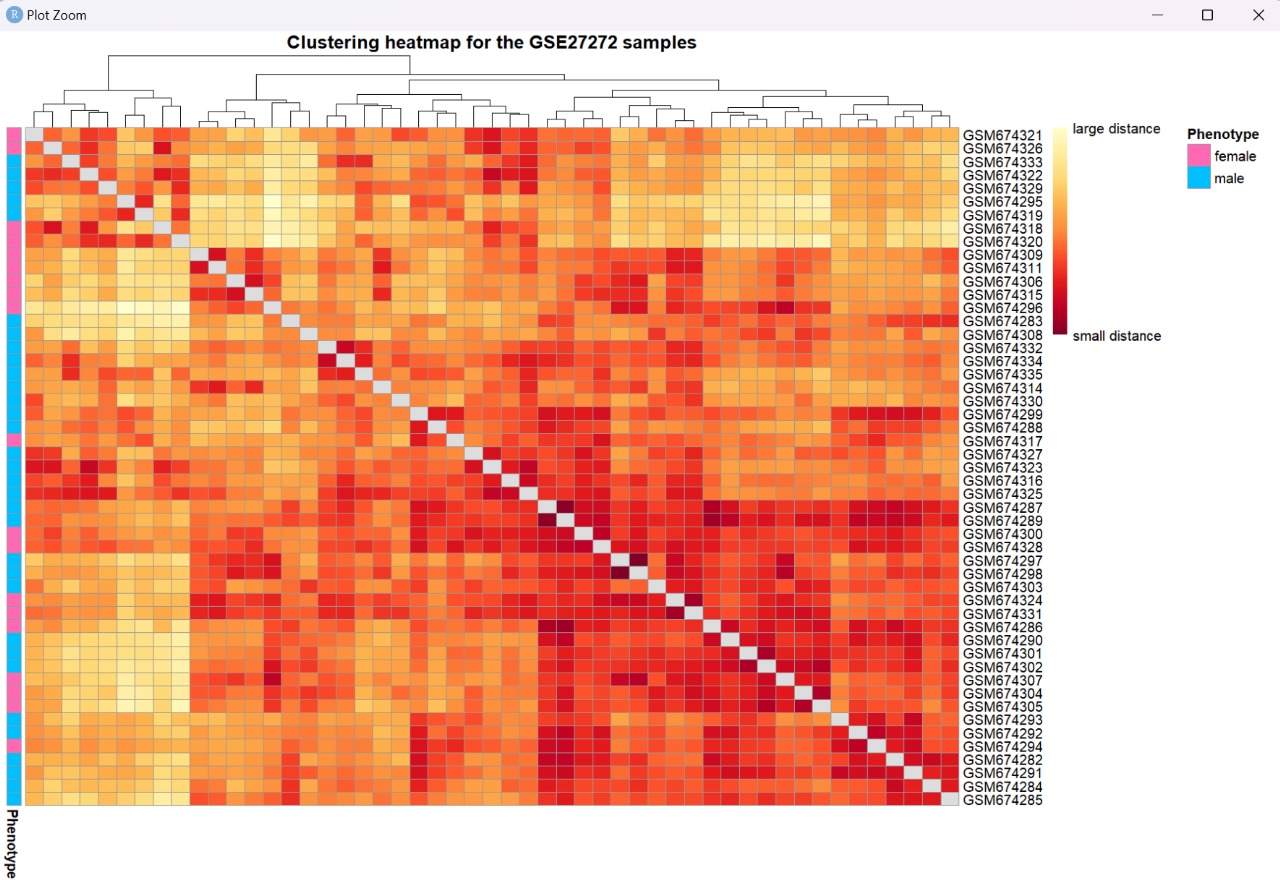
2. Multiple Testing Issues:

- Here in this data approximately 10 thousand genes are present . So When dealing with a large number of genes, there is an increased risk of obtaining false positives. The use of adjusted p-values control for multiple testing issues.

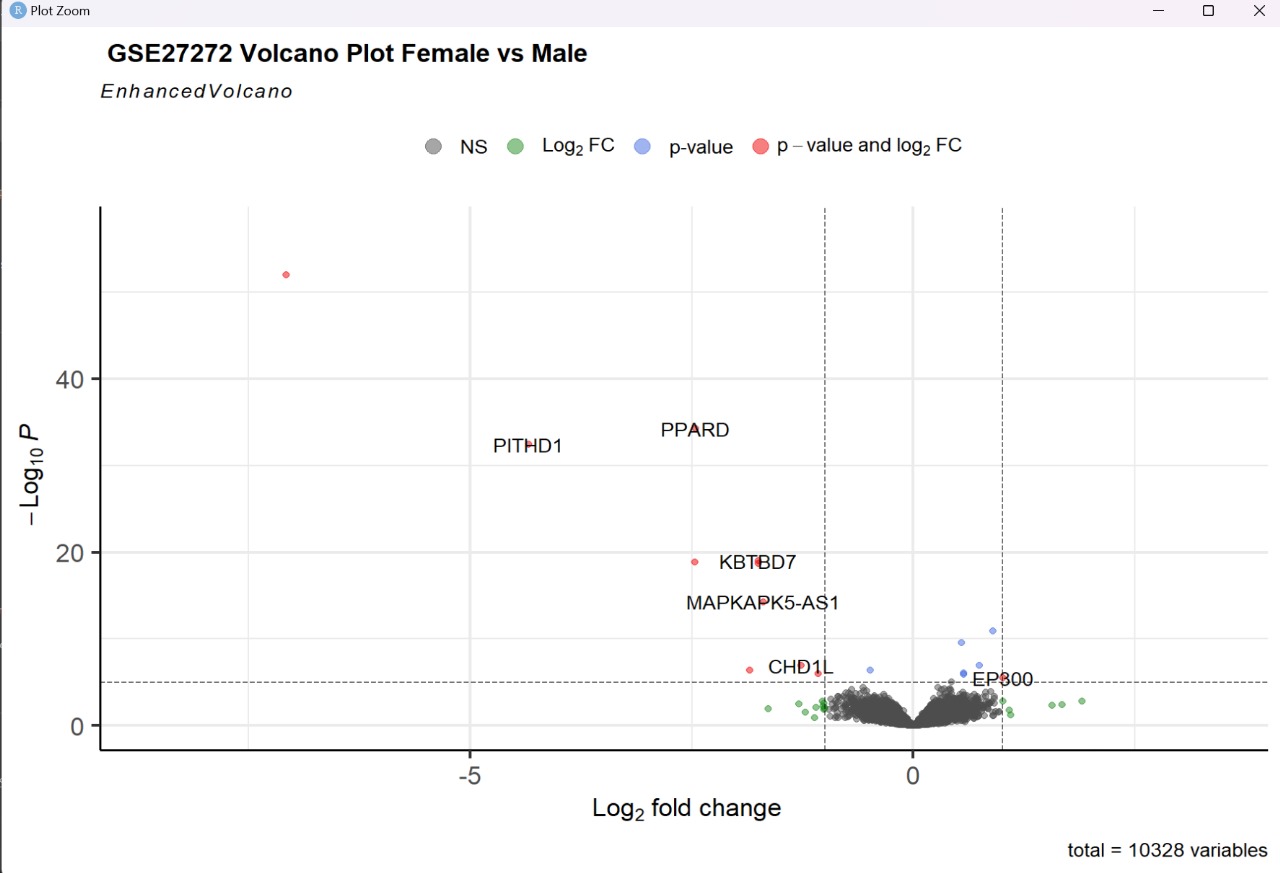
So,Always validate findings biologically to ensure that statistical significance translates to biological significance. And also use correction methods for multiple testing, such as adjusting p-values.

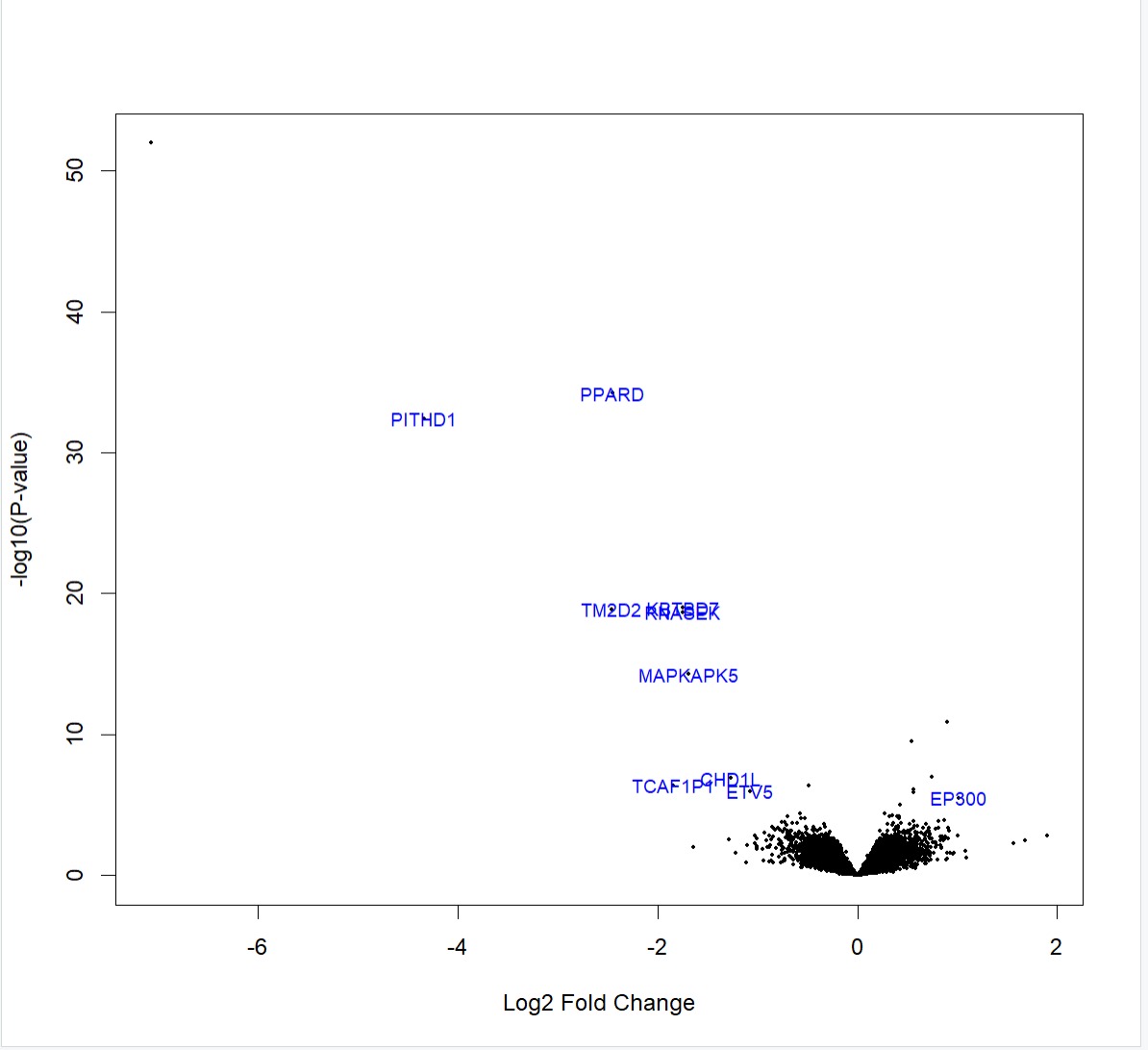
**Que.4**

**Heatmap visualization**

****

**Que.5**

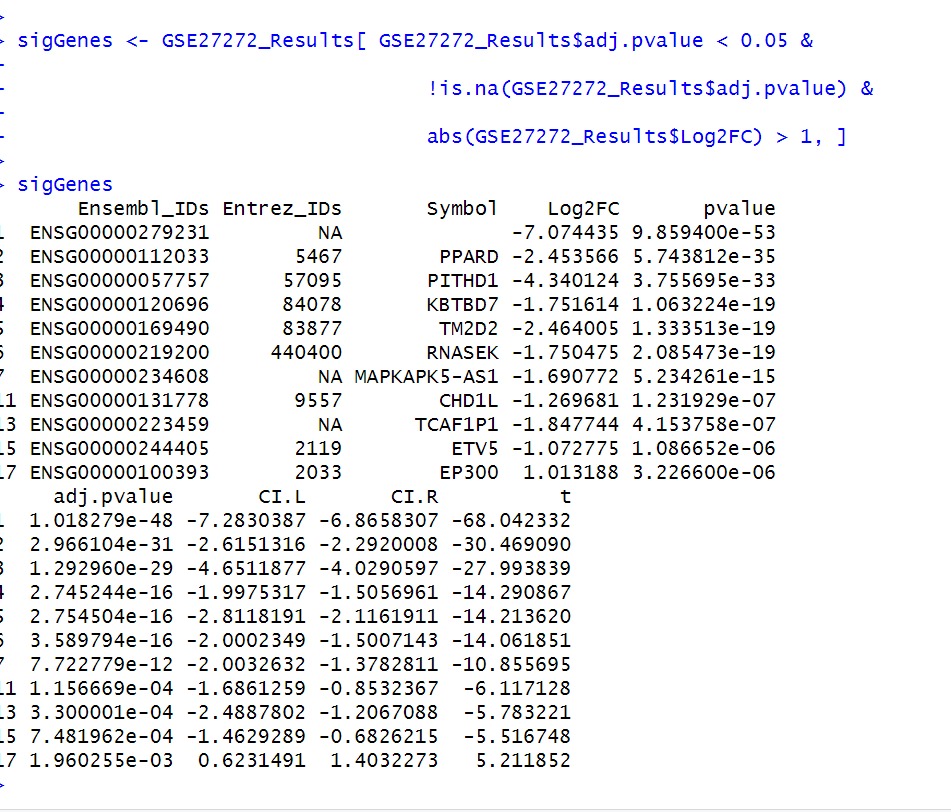


****

**Differential Expression Analysis Results - Fetal Sex**

In the analysis of differential gene expression by fetal sex, we utilized a volcano plot to visually inspect the relationship between log2 fold change and the negative logarithm (base 10) of the p-value. We set our significance criteria at -log10(p-value) > 1.3 and an absolute log2 fold change > 1.

By analysing the volcano plot, we observed an elevated cluster of points higher on the y-axis and positioned to the right on the x-axis. Using our specified criteria, we identified a total of 5 genes that are significantly differentially expressed by fetal sex. Notable genes in this set include PITHD1, PPARD, TM2D2, MAPKAPK5, and TCAF1PGHP15.

By analyzing above data,he genes with the highest absolute log2 fold change values, indicating the magnitude of the differential expression, are as follows:

### PARD = -7.074435

* PITHD = -2.453566
* KBTBD7 = -4.340124
* RNASEK = -2.464005
* MAPKAPK50-AS1 = -1.750475
* KBTBD7 = -4.340124

These values represent the log2 fold change, and the negative sign indicates downregulation.

**Que 6.**

When discussing the potential biological implications of differentially expressed genes in the context of fetal sex and tobacco smoke exposure, you should consider the specific genes identified, their known functions, and how these functions might relate to fetal development and the effects of tobacco smoke.

-**biological Implications of Differentially Expressed Genes in Fetal Sex and Tobacco Smoke Exposure**:

The analysis of differentially expressed genes (DEGs) in the context of fetal sex and tobacco smoke exposure has unveiled a set of genes that may play crucial roles in fetal development and response to environmental factors. The identified genes, such as PPARD, PITHD1, KBTBD7, and MAPKAPK5-AS1 and exhibit significant changes in expression levels in response to these factors.

**1**. **PPARD (Peroxisome Proliferator-Activated Receptor Delta):**

- PPARD, with a substantial downregulation, is known to be involved in various cellular processes. Its altered expression which shows potential impacts on fetal lipid metabolism and inflammatory responses in the context of tobacco smoke exposure.

**2.** **PITHD1 (PITH Domain-Containing 1):**

- PITHD1, downregulated in this analysis, has limited known functions. Further research shows its role in fetal development, especially in response to tobacco smoke.

**3. KBTBD7 (Kelch Repeat and BTB Domain-Containing 7):**

- KBTBD7, significantly downregulated, is associated with muscle development. Its altered expression may indicate potential effects on fetal muscle development and function, possibly influenced by tobacco smoke exposure.

4. MAPKAPK5-AS1 (MAPKAPK5 Antisense RNA 1):

- MAPKAPK5-AS1, downregulated, is an antisense RNA, and its role in fetal development is not well-identified. So it is difficult to explain role of tobacco smoke exposure during fetal development.

-**Biological Implications in the Context of Fetal Sex:**

- Examining these DEGs in the context of fetal sex reveals potential sex-specific differences in response to tobacco smoke exposure. For understanding the role between fetal sex and gene expression patterns is crucial for deciphering the complex mechanisms underlying developmental processes.