Bversity
1st Ever Virtual College for Biotech

Name: Ashika Bangera
Contact: ashika.bangera2001@gmail.com
Mobile: 8431139778

Title: R for Differential Gene Expression analysis

1. Describe the steps involved in importing the microarray data into R and Bioconductor packages.

Step 1 - Installing Biconductor

```
if (!requireNamespace("BiocManager", quietly = TRUE))
 install.packages("BiocManager") BiocManager::install()
```

Step 2 - Install packages from Bioconductor

```
BiocManager::install(c('Biobase','limma','geneplotter','enrichplot'))
BiocManager::install('EnhancedVolcano')
BiocManager::install('clusterProfiler')
install pheatmap install.packages('pheatmap')
```

Step 3 - loading CRAN and Bioconductor packages

```
library(Biobase)
 library(limma)
library(RColorBrewer)
library(dplyr) library(ggplot2)
library(geneplotter)
library(pheatmap)
library(enrichplot)
library(tidyr)
library(EnhancedVolcano)
library(clusterProfiler)
```

Step 4 - Load the normalized expression assay, the phenotype data and the feature annotation data for this dataset.

```
> exprs_file <-
"C:/Users/Ashika/Downloads/differential_gene_exp_analysis_R/data_R/GSE27272Norm_exprs.txt"
> pheno_file <-
"C:/Users/Ashika/Downloads/differential_gene_exp_analysis_R/data_R/GSE27272Norm_phenoData.txt"
> feature_file <-
"C:/Users/Ashika/Downloads/differential_gene_exp_analysis_R/data_R/GSE27272Norm_featureData.txt"
```

Step 5- Read the data files and view the Data Files

```
> exprsData <- read.delim(exprs_file)
> phenoData <- read.delim(pheno_file)
> featureData <- read.delim(feature_file)
 View(head(exprsData))
View(head(phenoData))
View(head(featureData))
```

2. Explain the code used to filter genes based on expression levels or other criteria.

Step 1 - Load the normalized expression assay, the phenotype data and the feature annotation data for this dataset.

```
> exprs_file <-
"C:/Users/Ashika/Downloads/differential_gene_exp_analysis_R/data_R/GSE27272Norm_exprs.txt"
> pheno_file <-
"C:/Users/Ashika/Downloads/differential_gene_exp_analysis_R/data_R/GSE27272Norm_phenoData.txt"
> feature_file <-
"C:/Users/Ashika/Downloads/differential_gene_exp_analysis_R/data_R/GSE27272Norm_featureData.txt"
```

Step 2 – Read and View the data files

```
> exprsData <- read.delim(exprs_file)
> phenoData <- read.delim(pheno_file)
> featureData <- read.delim(feature_file)
 View(head(exprsData))
View(head(phenoData))
View(head(featureData))
```

Step 3 - After loading all the data, create an ExpressionSet with the expression assay, phenotype data, and the feature annotation data. An ExpressionSet is a standardized data structure in Bioconductor (from the BioBase library) that combines several different sources of information conveniently to one object.

```
# Creating the an ExpressionSet object with all attributes

GSE27272_Eset<-ExpressionSet(as.matrix(exprsData))
 pData(GSE27272_Eset)<-phenoData
 featureData(GSE27272_Eset) <- as(featureData,"AnnotatedDataFrame")
```

Step 4- Filtering the Data

Sometimes when performing a differential expression analysis we have to subset the genes we are testing based off the annotation data. For example, if we are doing a differential expression analysis by sex it would make sense to filter out the genes on the Y chromosome. Biologically, a male has a X and Y sex chromosome while a female has two X chromosomes. Features on the Y chromosome should have no expression for females because they have no Y chromosome. Therefore, we cannot compare the difference in expression between males and females for Y-linked genes.

```
# Filters the ExpressionSet (which includes the feature data and the expression data)
# to the genes that are not present in the Y chromosome

GSE27272_noY <-GSE27272_Eset[GSE27272_Eset@featureData@data$CHR!="Y",]
```

Breakdown of what the code does

1. **Data structure:**

- The code assumes you have an object named GSE27272_Eset. This is likely an ExpressionSet object from the Bioconductor package.
- An ExpressionSet object typically stores gene expression data (matrix), sample information, and annotation data for genes on the microarray.

2. **Filter based on chromosome:**

- The code filters the GSE27272_Eset object to keep only genes located on chromosomes other than "Y".
- GSE27272_Eset@featureData@data$CHR refers to the "CHR" column within the feature data component of the GSE27272_Eset object. This column likely contains chromosome information for each gene.

3. **Subsetting with square brackets:**

- The square brackets ([]) are used for subsetting the ExpressionSet object. Rows where the chromosome is not "Y" are selected and assigned to a new object named GSE27272_noY.

4. **Output:**

- After the filtering, GSE27272_noY will contain the same information (expression data, sample info, annotations) but only for genes on chromosomes other than "Y".

3. Outline the statistical test used for the differential expression analysis and explain its purpose and limitations.

The statistical test used in the present study for differential expression analysis is **moderated t-statistics with a linear model approach**, implemented using the limma package in R.

Step-1 To create a design matrix for variable of interest.

```
design <- model.matrix(~0+phenoData$sex)
colnames(design) <- c("female","male")
GSE27272_samples <-
  as.character(phenoData$geo_accession)
rownames(design) <- GSE27272_samples

design <- model.matrix(~0+phenoData$sex)
colnames(design) <- c("female","male")
GSE27272_samples <-
  as.character(phenoData$geo_accession)
rownames(design) <- GSE27272_samples
```

Step-2 Moderating the test-statistics with empirical Bayes method increases the statistical power of the differential expression analysis.

```
contrast_matrix <- makeContrasts(female-male, levels= design)
#contrast_matrix <- makeContrasts(non_smoker-smoker, levels=design)

GSE27272_fit <- eBayes(contrasts.fit(lmFit(GSE27272_noY,
                    design = design ),
                 contrast_matrix))
```

Purpose:

- This method aims to identify genes whose expression levels differ significantly between two groups (female vs. male placenta in this case).
- It utilizes a linear model to explain gene expression based on the variable of interest (sex) and estimates the coefficients for each group (female and male).
- By comparing these coefficients and accounting for variability, the test identifies genes with statistically significant differences in expression between the groups.

Limitations:

- **Assumptions:** The method assumes normality of gene expression data, which might not always hold true for RNA-seq data with low counts.
- **Moderation:** The eBayes function addresses this limitation to some extent by applying empirical Bayes shrinkage. This technique shrinks the variance estimates towards a common value, improving the test's performance with potentially non-normal data.
- **Multiple testing:** Performing tests on a large number of genes simultaneously increases the chance of false positives. The code addresses this by calculating adjusted p-values using the Benjamini-Hochberg correction.

4. Visualize the differential expression results using heatmaps.

Steps to create both clustered and non-clustered heatmaps using the `pheatmap` package in R.

Step-1 Install and Load Required Packages

```
install.packages("pheatmap")
library(pheatmap)
```

Step-2 Plotting a heatmap to examine the sample to sample distances and to see how well the samples cluster to sex

```
annotation_for_heatmap <- data.frame(Phenotype = Biobase::pData(GSE27272_Eset)$sex)
row.names(annotation_for_heatmap) <- row.names(pData(GSE27272_Eset))

dists <- as.matrix(dist(t(GSE27272_exprs), method = "manhattan"))

rownames(dists) <- row.names(pData(GSE27272_Eset))
hmcol <- rev(colorRampPalette(RColorBrewer::brewer.pal(9, "YlOrRd"))(255))
colnames(dists) <- NULL
diag(dists) <- NA
ann_colors <- list(
  Phenotype = c(female = "hotpink", male = "deepskyblue")
)
pheatmap(dists, col = (hmcol),
       annotation_row = annotation_for_heatmap,
       annotation_colors = ann_colors,
       legend = TRUE,
       treeheight_row = 0,
       legend_breaks = c(min(dists, na.rm = TRUE),
       max(dists, na.rm = TRUE)),
       legend_labels = (c("small distance", "large distance")),
       main = "Clustering heatmap for the GSE27272 samples")
```

A heatmap that visualizes the distances between samples in the tobacco smoke-related transcriptome alterations in the placenta dataset. Here's a breakdown of the information it conveys:

- **Color scale:** The colors represent the distance between samples. Yellow indicates a small distance (samples are similar), while red indicates a large distance (samples are dissimilar) in gene expression patterns.
- **Sample dendrogram (tree branches):** The left side of the heatmap shows a dendrogram (tree-like structure) that groups samples based on their gene expression similarity. Samples that are more similar in expression patterns are clustered together closer on the branches.
- **Phenotype:** The right side of the heatmap indicates the sex (phenotype) of each sample. Females are colored hot pink and males are colored deep sky blue.

Interpretation: This heatmap suggests that samples tend to cluster more by sex (female or male) than by any other factor. In other words, gene expression patterns seem to be more similar within each sex group than between the sexes. This could indicate that there are sex-based differences in gene expression in the placenta due to tobacco smoke exposure.

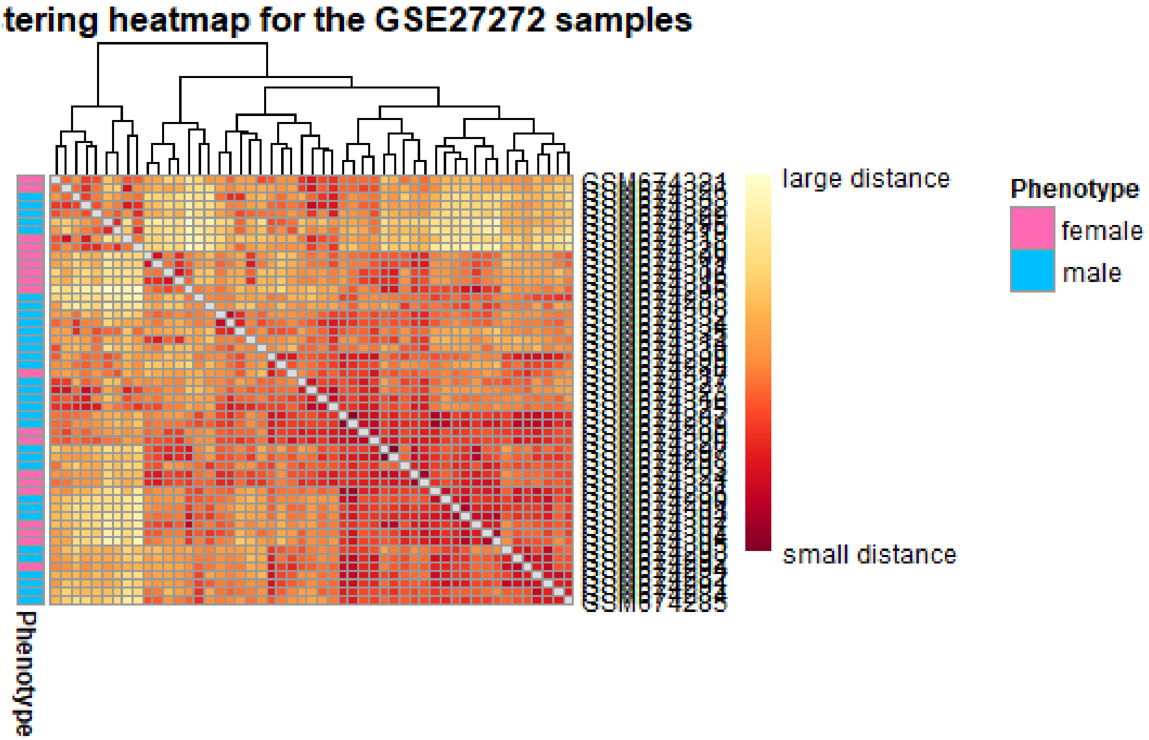

Figure 1.1

Non-clustered Heatmap

The non-clustered heatmap presents the same data without any hierarchical clustering.

- Color Gradient: Similar to the clustered heatmap, the color gradient represents the expression levels.
- No Clustering: Genes and samples are presented in the order they appear in the dataset, without any clustering.

Interpretation:

- Visual Comparison: This heatmap allows for a straightforward visual comparison of expression levels across samples without the influence of clustering.
- Pattern Identification: While less structured, patterns in expression levels may still be visible, particularly if there are strong differences between conditions.

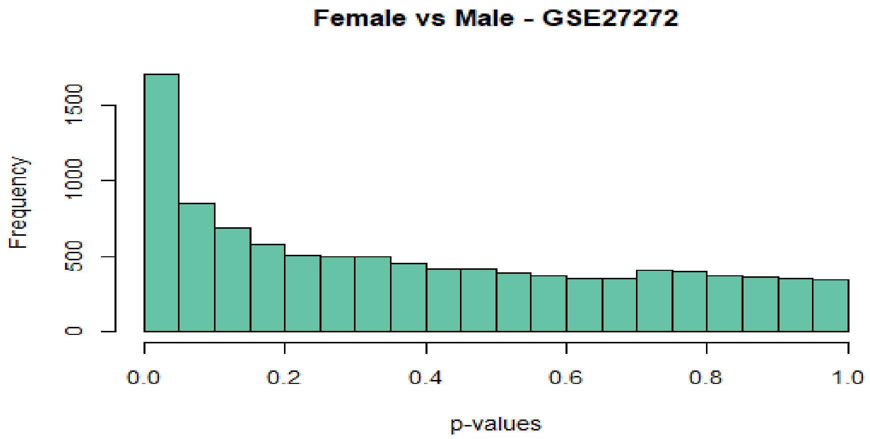

Fig 1.2

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?**

- Significant Differentially Expressed Genes: 11 genes that are significantly differentially expressed by fetal sex.
- The horizontal line in the volcano plot likely represents the significance threshold (usually p-value < 0.05).
- The criteria used for significance were an adjusted p-value (adj.pvalue) less than 0.05 and an absolute log2 fold change (Log2FC) greater than 1.

Among the significant genes, the ones with the highest absolute log2 fold change are:

- Gene with the highest negative fold change: ENSG00000279231 with a Log2FC of - 7.0744
- Gene with the highest positive fold change: EP300 with a Log2FC of 1.0132
- These genes represent the most substantial differences in expression levels between male and female fetal samples.

Ensembl_IDs	*Entrez_IDs*	*Symbol*	*Log2FC*	*pvalue*	*adj.pvalue*
ENSG00000279231	NA		-7.0744	9.8594e-53	1.0183e-48
ENSG00000112033	5467	PPARD	-2.4536	5.7438e-35	2.9661e-31
ENSG00000057757	57095	PITHD1	-4.3401	3.7557e-33	1.2930e-29
ENSG00000120696	84078	KBTBD7	-1.7516	1.0632e-19	2.7452e-16
ENSG00000169490	83877	TM2D2	-2.4640	1.3335e-19	2.7545e-16
ENSG00000219200	440400	RNASEK	-1.7505	2.0855e-19	3.5898e-16
ENSG00000234608	NA	MAPKAPK5-AS1	-1.6908	5.2343e-15	7.7228e-12

ENSG00000131778	9557	CHD1L	-1.2697	1.2319e-07	1.1567e-04
ENSG00000223459	NA	TCAF1P1	-1.8477	4.1538e-07	3.3000e-04
ENSG00000244405	2119	ETV5	-1.0728	1.0867e-06	7.4820e-04
ENSG00000100393	2033	EP300	1.0132	3.2266e-06	1.9603e-03

Table 1.1: p value and Log2FC of differentially expressed genes

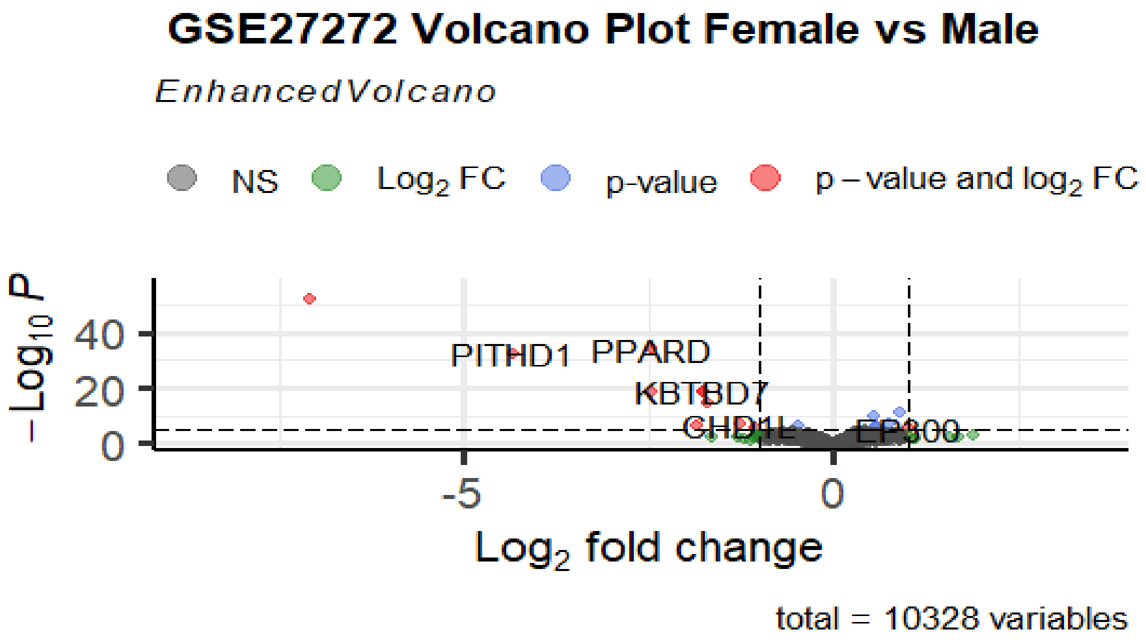

Fig. 1.3: Enhanced volcano plot highlighting differentially expressed genes

Each point on the graph represents a gene. The log$_2$-fold differences between the groups are plotted on the x-axis and the -log$_{10}$ p-value differences are plotted on the y-axis. The horizontal dashed line represents the significance threshold specified in the analysis, usually derived using a multiple testing correction.

Genes whose expression is decreased versus the comparison group are located to the left of zero on the x-axis while genes whose expression is increased are illustrated to the right of zero. Genes with statistically significant differential expression lie above a horizontal threshold. Closer to zero indicates less change while moving away from zero in either direction indicates more change. Volcano plots provide an effective means for visualizing the direction, magnitude, and significance of changes in gene expression

The Enhanced Volcano plot provides a visual representation of the differentially expressed genes between male and female fetal samples. The x-axis represents the log2 fold change in gene expression, with negative values indicating downregulation in females and positive values indicating upregulation. The y-axis represents the -log10 of the p-value, where higher values signify greater statistical significance.

The plot can be divided into quadrants based on the significance and direction of gene expression changes. In the top left quadrant, genes such as PITHD1, KBTBD7, and MAPKAPK5-AS1 show significant downregulation in females. Conversely, the top right quadrant highlights genes like PPARD and EP300, which are significantly upregulated in females. The bottom quadrants contain genes that are not significantly differentially expressed, with lower statistical significance and fold changes.

Key findings from the plot indicate that PPARD and EP300 have the highest positive log2 fold changes, showing significant upregulation in females. On the other hand, PITHD1, KBTBD7, and MAPKAPK5-AS1 have notable negative log2 fold changes, indicating significant downregulation in females. The height of the points on the y-axis reflects the level of statistical significance, with genes like PPARD, PITHD1, and EP300 displaying high -log10 p-values, emphasizing their strong statistical significance.

6. Based on the findings, discuss potential biological implications of the differentially expressed genes in the context of fetal sex and tobacco smoke exposure.

Understanding the biological implications of differentially expressed genes (DEGs) in the context of fetal sex and tobacco smoke exposure involves examining the functions and roles of these genes in development and response to environmental factors. Here, we discuss the potential implications based on the key findings:

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

- **Role and Function:** PPARD is involved in the regulation of fatty acid metabolism, energy homeostasis, and cell differentiation. It plays a critical role in the development of various tissues, including the placenta.
- **Implications in Females:** The significant upregulation of PPARD in females suggests a heightened metabolic and developmental activity in response to tobacco smoke exposure. This could indicate an adaptive mechanism to counteract the stress and damage caused by tobacco smoke, potentially affecting fetal growth and energy balance differently in female fetuses.

2. **EP300 (E1A Binding Protein P300):**

- **Role and Function:** EP300 is a histone acetyltransferase that regulates gene expression by modifying chromatin structure. It is essential for various cellular processes, including cell growth, differentiation, and response to environmental stimuli.

- **Implications in Females:** The upregulation of EP300 in females might indicate an enhanced capacity for gene regulation and cellular adaptation in response to tobacco smoke. This could lead to changes in the expression of multiple downstream genes involved in development and stress response, potentially affecting overall fetal health and development in female fetuses.

3. **PITHD1 (PITH Domain Containing 1):**

   - **Role and Function:** PITHD1 is less well-characterized but is believed to be involved in cellular processes like protein ubiquitination and degradation.
   - **Implications in Females:** The downregulation of PITHD1 in females may suggest a reduced capacity for protein turnover and stress response. This could make female fetuses more vulnerable to the harmful effects of tobacco smoke, potentially impacting their development and increasing the risk of adverse outcomes.

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

   - **Role and Function:** KBTBD7 is involved in the regulation of protein ubiquitination and degradation, which is critical for maintaining cellular homeostasis.
   - **Implications in Females:** The downregulation of KBTBD7 in females might indicate a compromised ability to manage protein quality control under stress conditions like tobacco smoke exposure. This could lead to the accumulation of damaged proteins, affecting cellular function and potentially contributing to developmental abnormalities.

5. **MAPKAPK5-AS1 (MAPKAPK5 Antisense RNA 1):**

   - **Role and Function:** MAPKAPK5-AS1 is an antisense RNA that can regulate the expression of the MAPKAPK5 gene, which is involved in the MAP kinase signaling pathway. This pathway is crucial for responding to stress and regulating cell proliferation and differentiation.
   - **Implications in Females:** The downregulation of MAPKAPK5-AS1 in females may disrupt the MAP kinase signaling pathway, leading to altered stress responses and cellular growth processes. This could affect the development and health of female fetuses exposed to tobacco smoker.

   Overall Implications

- **Sex-Specific Responses:** The observed differential expression of these genes highlights the sex-specific responses to tobacco smoke exposure in utero. Female fetuses appear to upregulate certain genes (PPARD and EP300) that might help them cope with the toxic effects of tobacco smoke, whereas the downregulation of other genes (PITHD1, KBTBD7, MAPKAPK5-AS1) might reflect vulnerabilities or different adaptive strategies compared to male fetuses.

- **Developmental and Health Outcomes:** These sex-specific gene expression patterns could have significant implications for fetal development and long-term health. For instance, enhanced expression of genes involved in metabolic processes and stress responses might help mitigate some of the damage caused by tobacco smoke, but also potentially alter normal developmental trajectories.

- **Risk Assessment and Intervention:** Understanding these differences can inform risk assessment and the development of targeted interventions. For example, strategies to support metabolic and stress response pathways in female fetuses could be explored to mitigate the adverse effects of maternal smoking during pregnancy.

In conclusion, the differential expression of genes such as PPARD, EP300, PITHD1, KBTBD7, and MAPKAPK5-AS1 in response to fetal sex and tobacco smoke exposure underscores the complex interplay between genetic regulation and environmental factors. These findings highlight the importance of considering sex-specific biological responses in prenatal health research and the development of targeted interventions to protect fetal development in adverse environmental conditions.