**Assignment-1**

**Practical Bioinformatics**

**PARAS DHIMAN – 2021482**

The code is a script for performing differential expression analysis on gene expression data using the limma package in R.

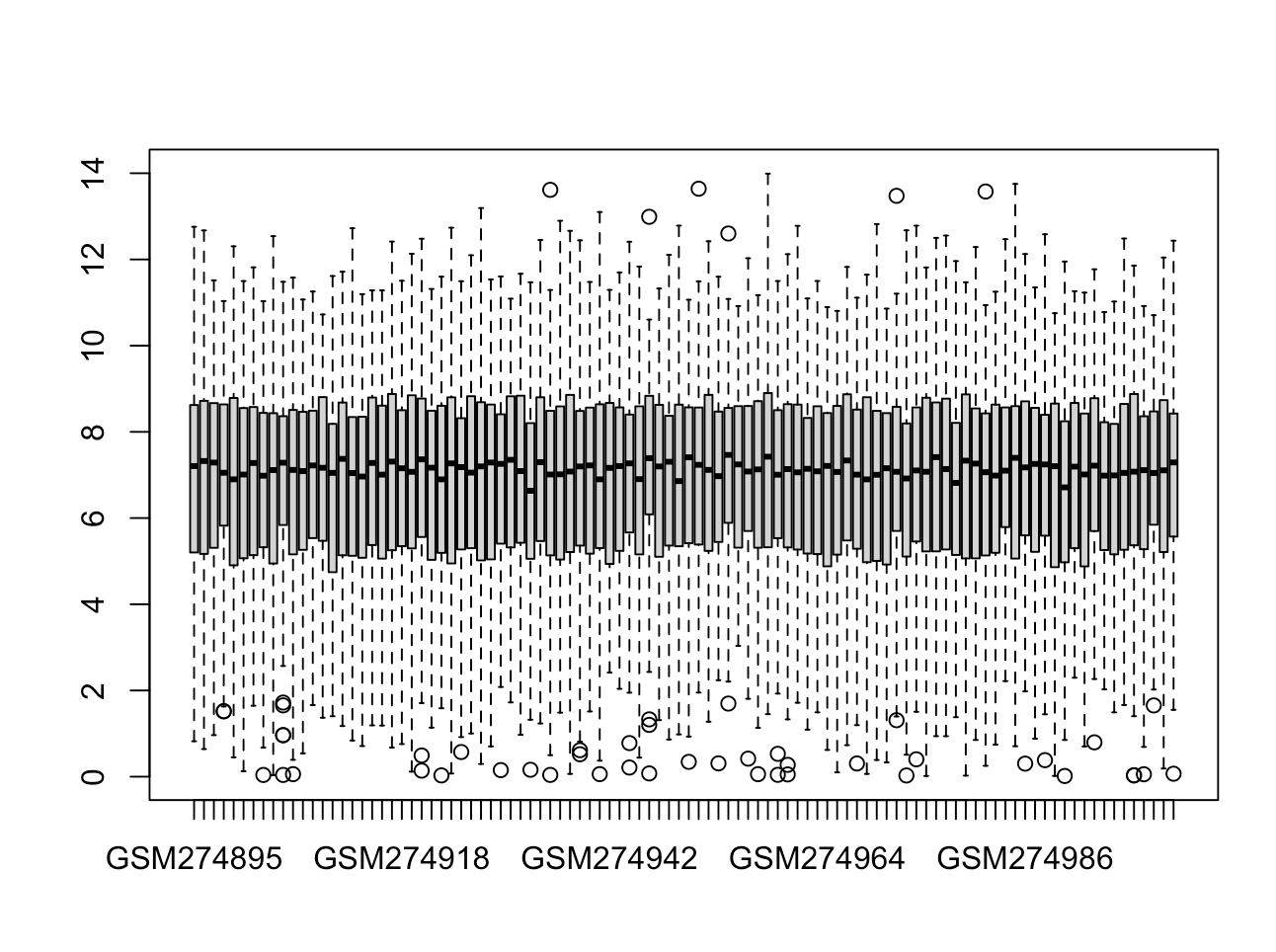
The script first loads the gene expression data from a GEO dataset and applies log2 transformation. It then assigns the samples to different groups and sets up a design matrix for linear modeling. The script then performs contrasts of interest and calculates model coefficients, computes statistics, and generates a table of top significant genes. The script also generates visualizations, including a histogram of P-values, Q-Q plot for t-statistic, volcano plot, and MD plot.

In addition to differential expression analysis, the script also performs general gene expression data analysis, including generating a box-and-whisker plot and expression value distribution plot. The script also uses the UMAP method for dimensionality reduction and generates a UMAP plot.

**1.)** In Code Itself

**2.)** In Code Itself

The Box plot here that we get are:



**3.)** Log transformation is a common preprocessing step in the analysis of microarray data. Here are some of the effects that are observed after completing the log transformation of microarray data:

1. It Reduced the range of expression values: The log transformation compresses the dynamic range of expression values. This means that highly expressed genes are not as far apart from lowly expressed genes as they would be on a linear scale.

2. Made the data more normally distributed: The log transformation can make data more symmetric and normal, which is useful when performing statistical tests that assume a normal distribution of data.

3. It improved the ability to detect small changes: The log transformation can increase the sensitivity of detecting small changes in expression levels. This is because a change in the log-transformed value corresponds to a fold change on the original scale.

4. Stabilized the variance: The log transformation can help stabilize the variance of the data, which is important for statistical analyses that assume equal variances across groups.

5. Facilitated the interpretation of fold changes: The log transformation converts fold changes into additive changes, which can make it easier to interpret the results of differential expression analyses.

It is important to note that the effects of log transformation may depend on the specific microarray platform and experimental design, and that other preprocessing steps (e.g., normalization) may also be necessary to ensure the accuracy and reliability of downstream analyses.

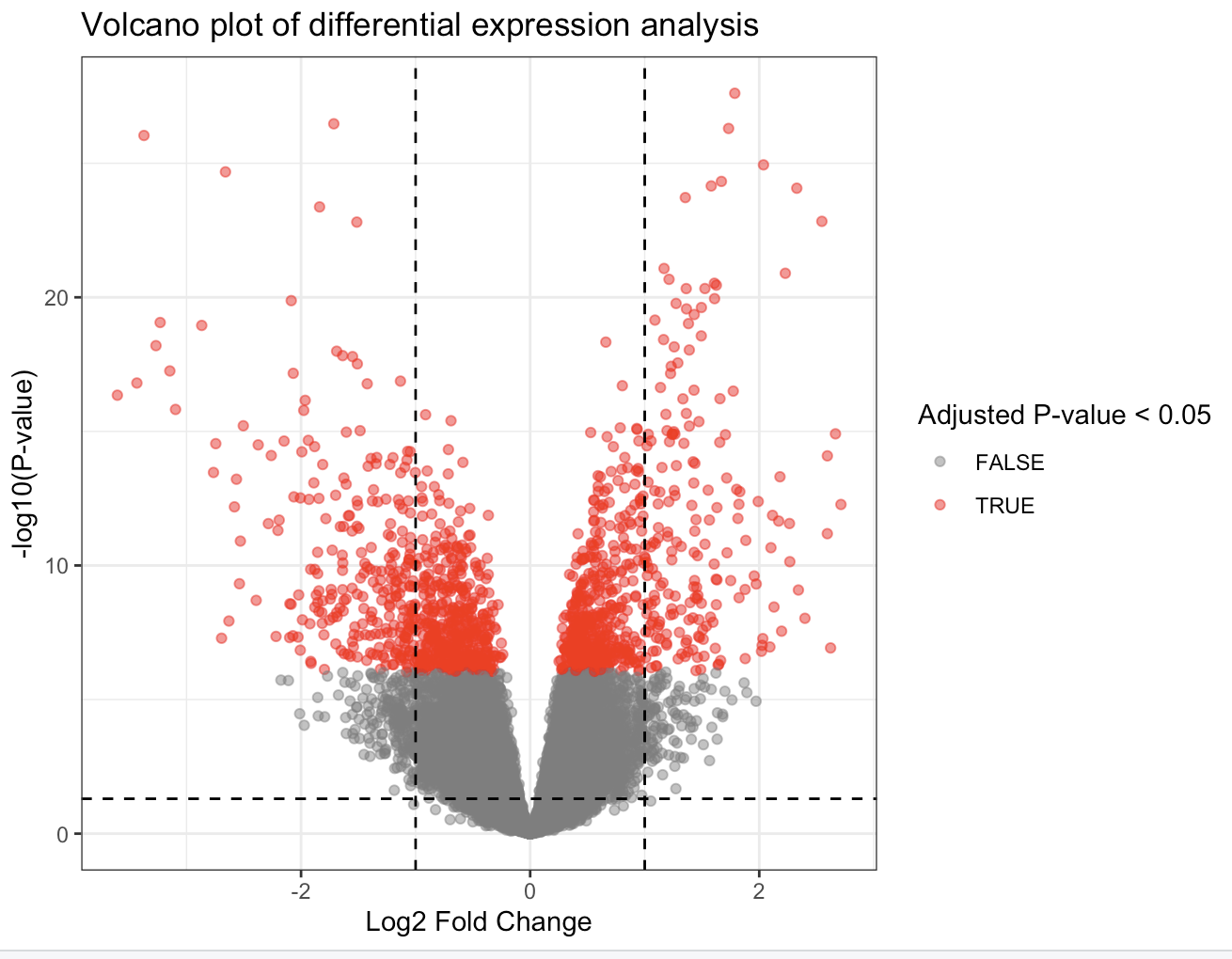
**4.)** Volcano plot that we get is:

The plot uses the ggplot2 library in R and takes in a data frame called t\_results, with columns Log2FC (log2 fold change) and P.Value (p-value).

The plot maps Log2FC to the x-axis and -log10(P.Value) to the y-axis. The color of each point is determined by whether its corresponding Adjusted P.Value (not shown in the code) is less than 0.05, with those meeting the threshold shown in red and those that do not shown in gray.

The plot also includes dashed lines indicating Log2FC of -1 and 1, and a y-intercept of -log10(0.05) indicating a significance threshold of p-value < 0.05.

The title and axis labels are added using the labs() function, and the theme is set to a simple black and white style using the theme\_bw() function.



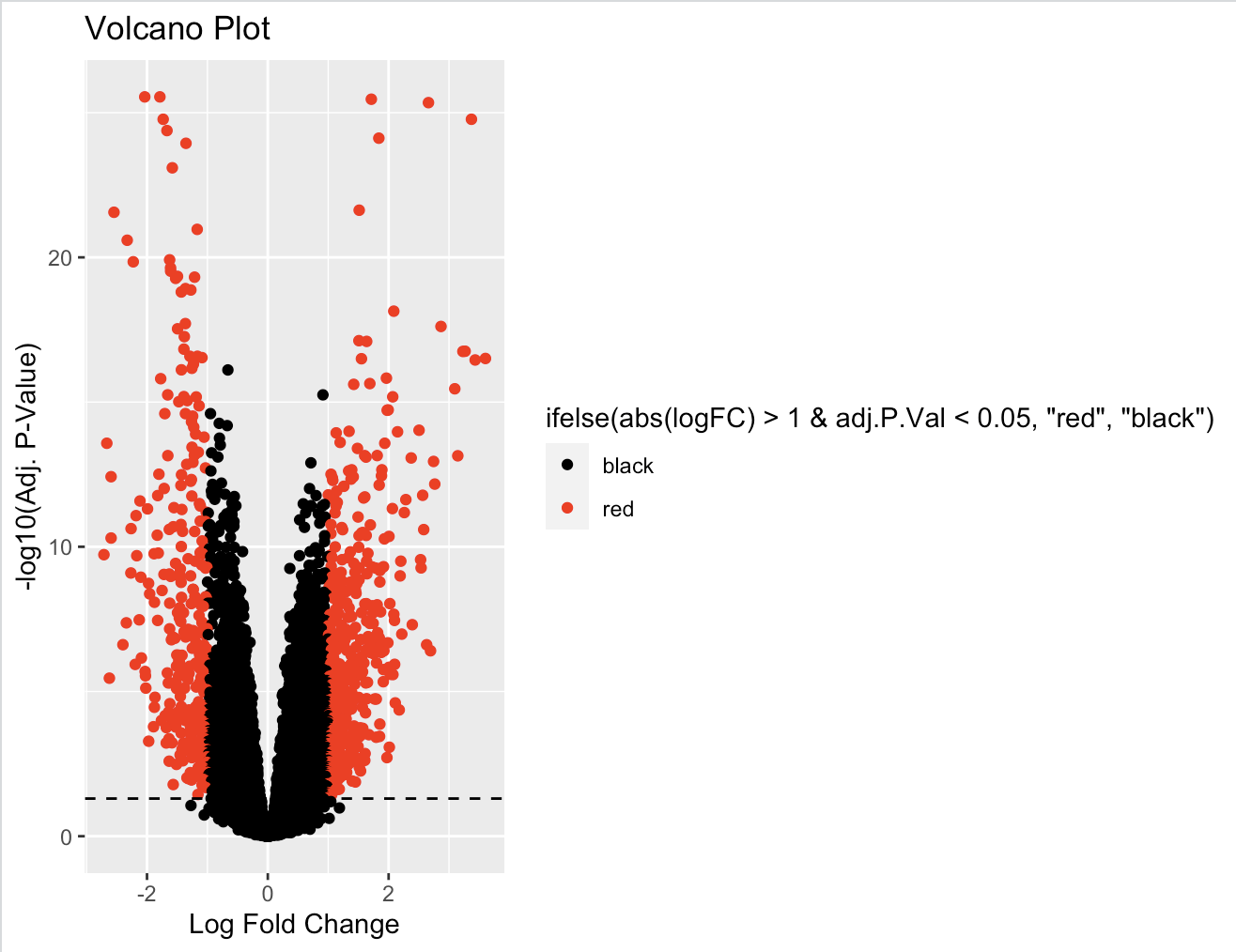
**5.)**

The plot uses the ggplot2 library in R and takes in a data frame called results, with columns logFC (log fold change) and adj.P.Val (adjusted p-value).

The plot maps logFC to the x-axis and -log10(adj.P.Val) to the y-axis. The color of each point is determined by whether its corresponding logFC value is greater than 1 or less than -1 and its corresponding adj.P.Val value is less than 0.05, with those meeting the criteria shown in red and those that do not shown in black.

The plot also includes a dashed line indicating a significance threshold of adj.P.Val < 0.05.

The axis labels are added using the labs() function, and the title is set to "Volcano Plot". The color scale is set to black and red using the scale\_color\_manual() function.



**6.)** 778 genes were found to be significantly differentially expressed with a log fold change cutoff of 1 and an adjusted p-value cutoff of 0.05.

the choice of cutoffs for log fold change (log(FC)) and adjusted p-value depends on various factors such as the experimental design, sample size, and statistical power.

In the given scenario, 778 genes were found to be significantly differentially expressed with a log fold change cutoff of 1 and an adjusted p-value cutoff of 0.05. A log fold change cutoff of 1 means that the expression level of a gene is at least 2-fold higher or lower in one group compared to another group. An adjusted p-value cutoff of 0.05 means that the probability of observing a false positive result is less than 5%.

To choose a significant cutoff for log(FC) and p-values, we had balance the trade-off between sensitivity and specificity. A stringent cutoff will result in a lower false positive rate but may miss some true positive genes, while a lenient cutoff will increase the false positive rate but may include some false positive genes. Therefore, we need to balance the false positive and false negative rates based on the goals of the study.

Based on the information, we have chosen a log fold change cutoff of 1 and an adjusted p-value cutoff of 0.05. A log fold change cutoff of 1 means that the expression level of a gene is at least 2 fold higher or lower in one group compared to another group. This cutoff is less stringent than others, but still biologically meaningful. An adjusted p-value cutoff of 0.05 means that the probability of observing a false positive result is less than 5%. This cutoff is more stringent than others and also reasonable for many studies.

However, it is important to note that the choice of cutoff is subjective and may depend on the specific research question and the experimental design. It is recommended to perform sensitivity analysis by varying the cutoffs and evaluating the robustness of the results.

**7.)** Done In code itself

**8.)** The code snippet provided continues the visualization of the results of a gene set enrichment analysis performed using the ego function in R. Here is a brief explanation of the code:

* The lattice package is loaded, which provides additional plotting functions.
* The ego\_df data frame is created again, which contains the GO term names and p-values from the ego object.
* A dot plot of the p-values is created using the dotplot function, as shown in the previous code snippet.
* A bar chart of the p-values is created using the barplot function. The horiz parameter makes the chart horizontal, and las = 1 rotates the y-axis labels for better visibility.
* Another bar chart is created, this time showing the number of genes associated with each enriched GO term. The xlim parameter sets the x-axis limits to include all counts, and the main, xlab, and ylab parameters provide informative labels for the chart.
* The ggplot2 package is loaded, which provides additional plotting functions.
* The enrichResult object generated by the ego function is converted into a data frame called go\_terms, which contains columns for GO term names, gene ratios, p-values, and q-values.
* The dplyr package is loaded, which provides a handy top\_n function for filtering data frames.
* The go\_terms\_top data frame is created by selecting the top 10 enriched GO terms based on their q-values.

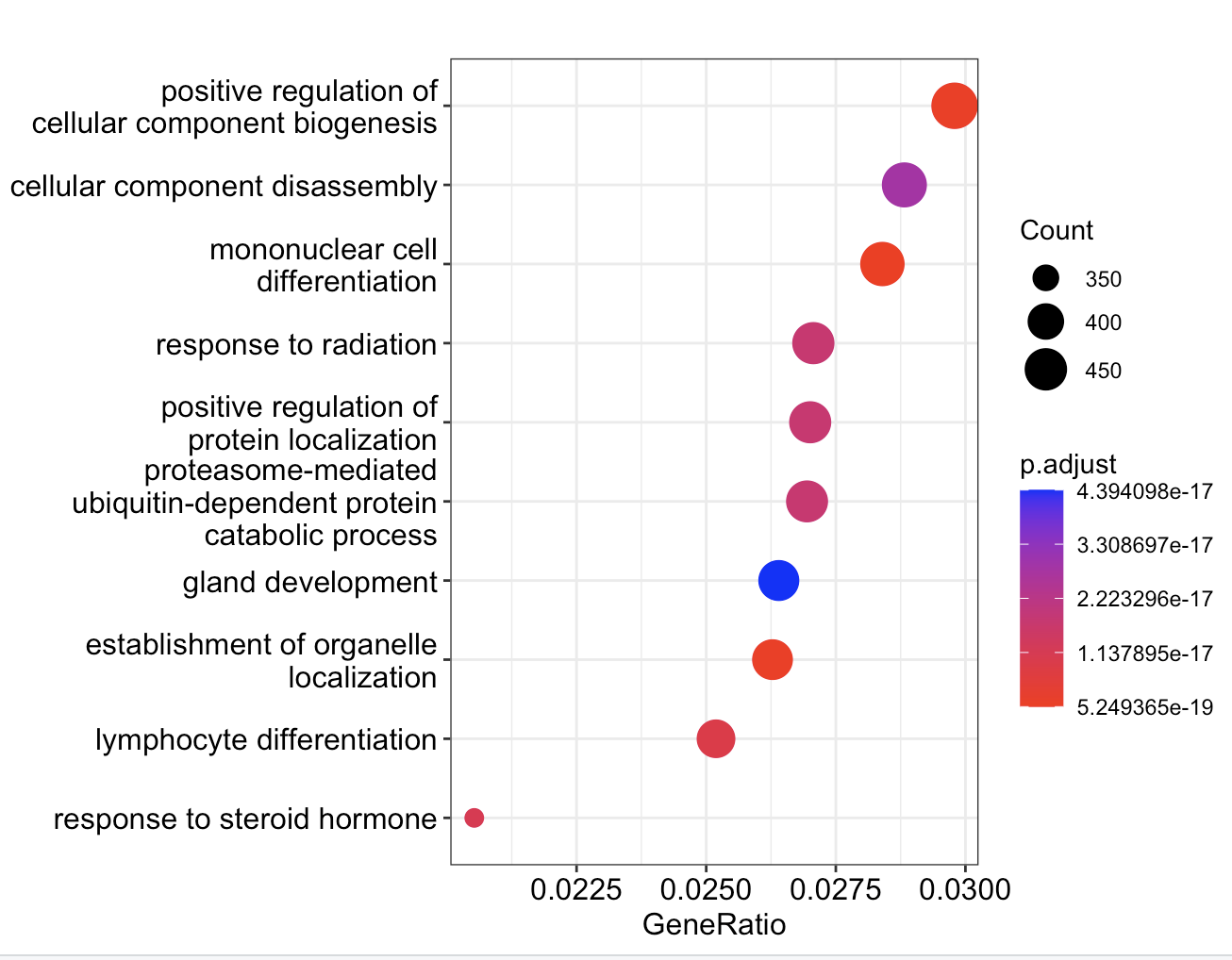
Observations:

* The dot plot and bar chart of the p-values show a similar pattern of enrichment, where a few GO terms have much lower p-values than the others.
* The bar chart of gene counts shows that some GO terms are associated with many genes, while others have only a few.
* The ggplot2 code could be used to create additional plots, such as a heatmap showing the relationships between enriched GO terms and the genes associated with them. Overall, exploring different types of plots can help reveal different aspects of the enrichment analysis results.

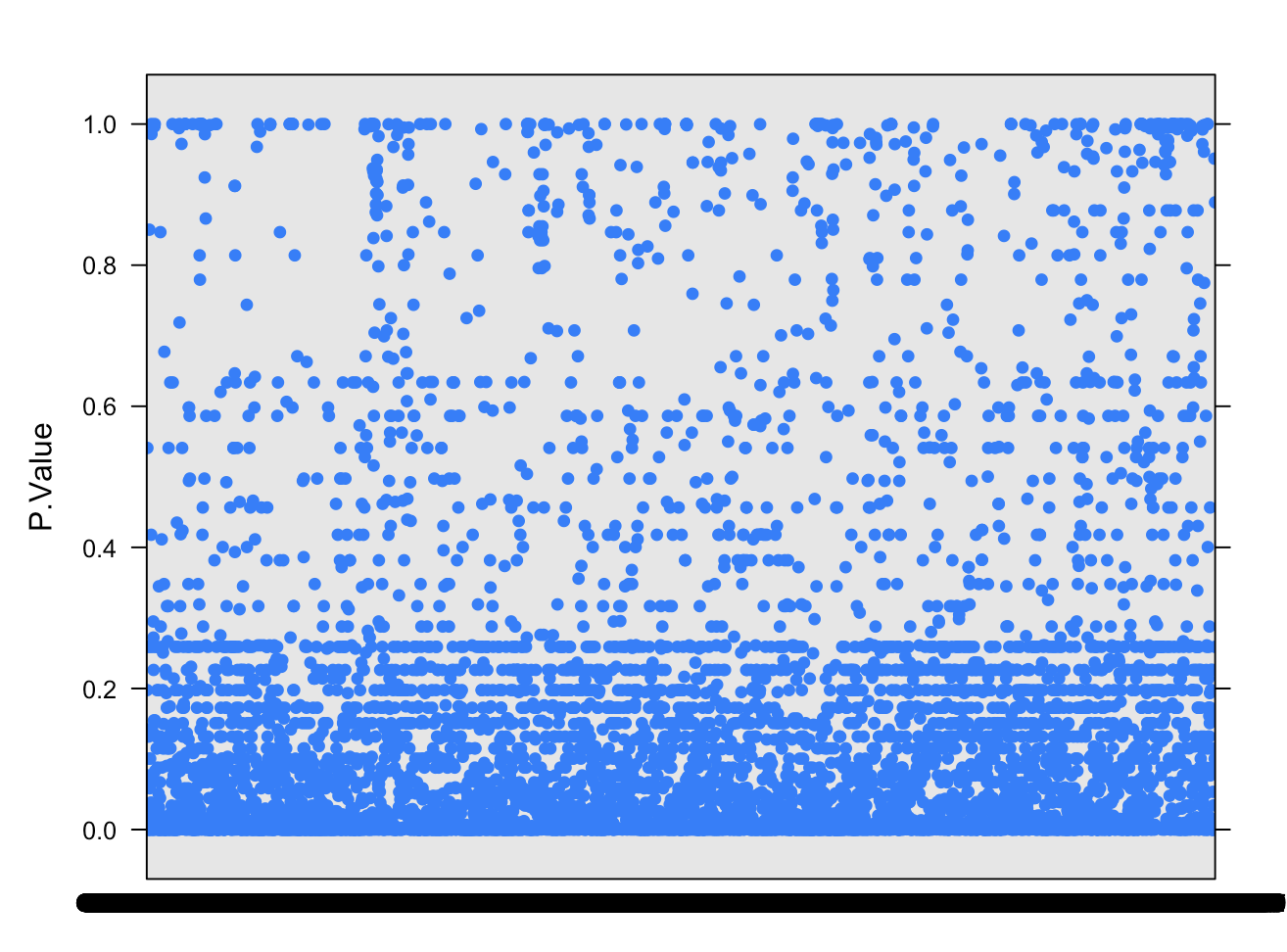
dot plot of the p-values is created using the dotplot function:

There are 2 dotplots:

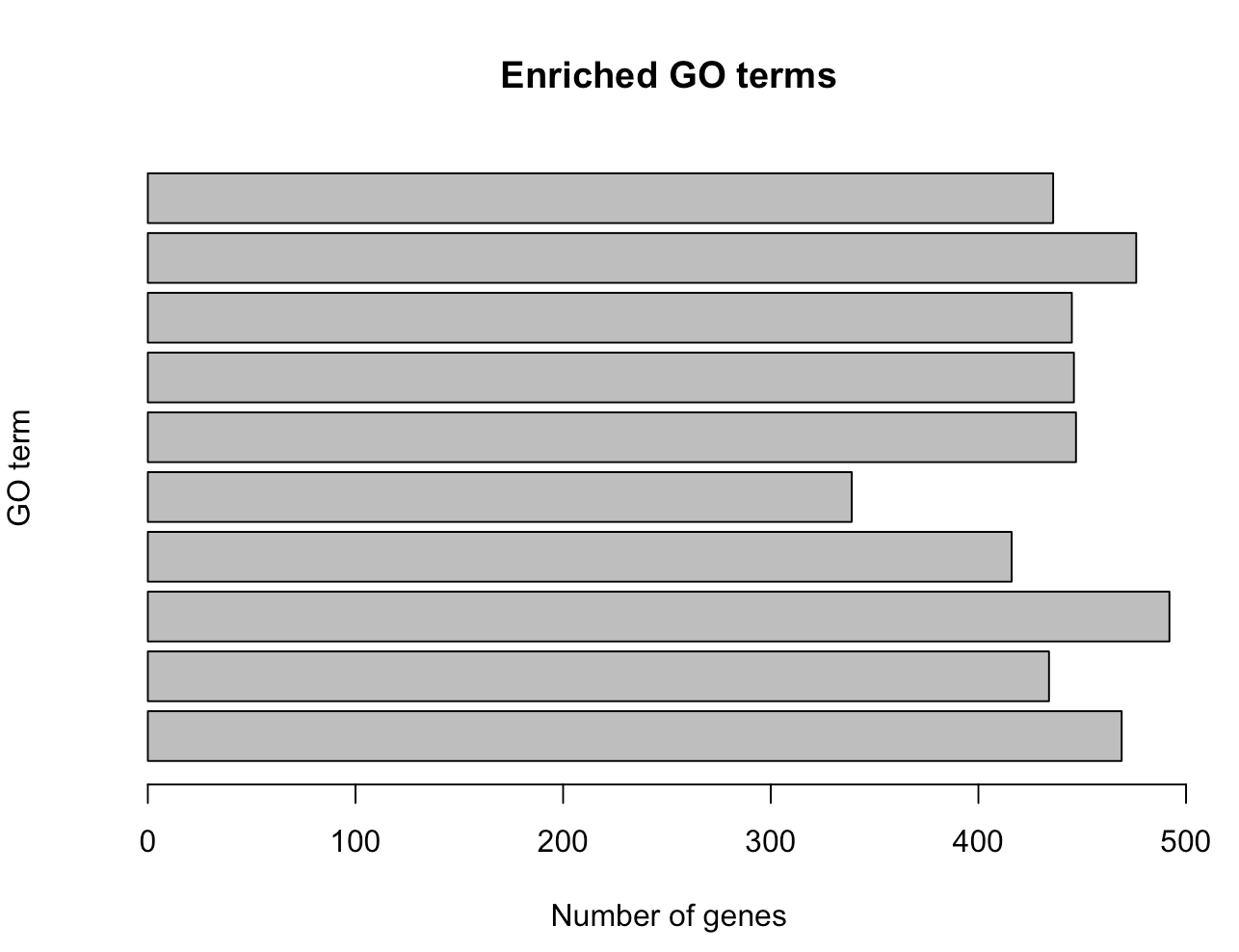
Dotplot1 is givem as(this has also taken into consideration the size-representing counts and colors representing p.adjust):

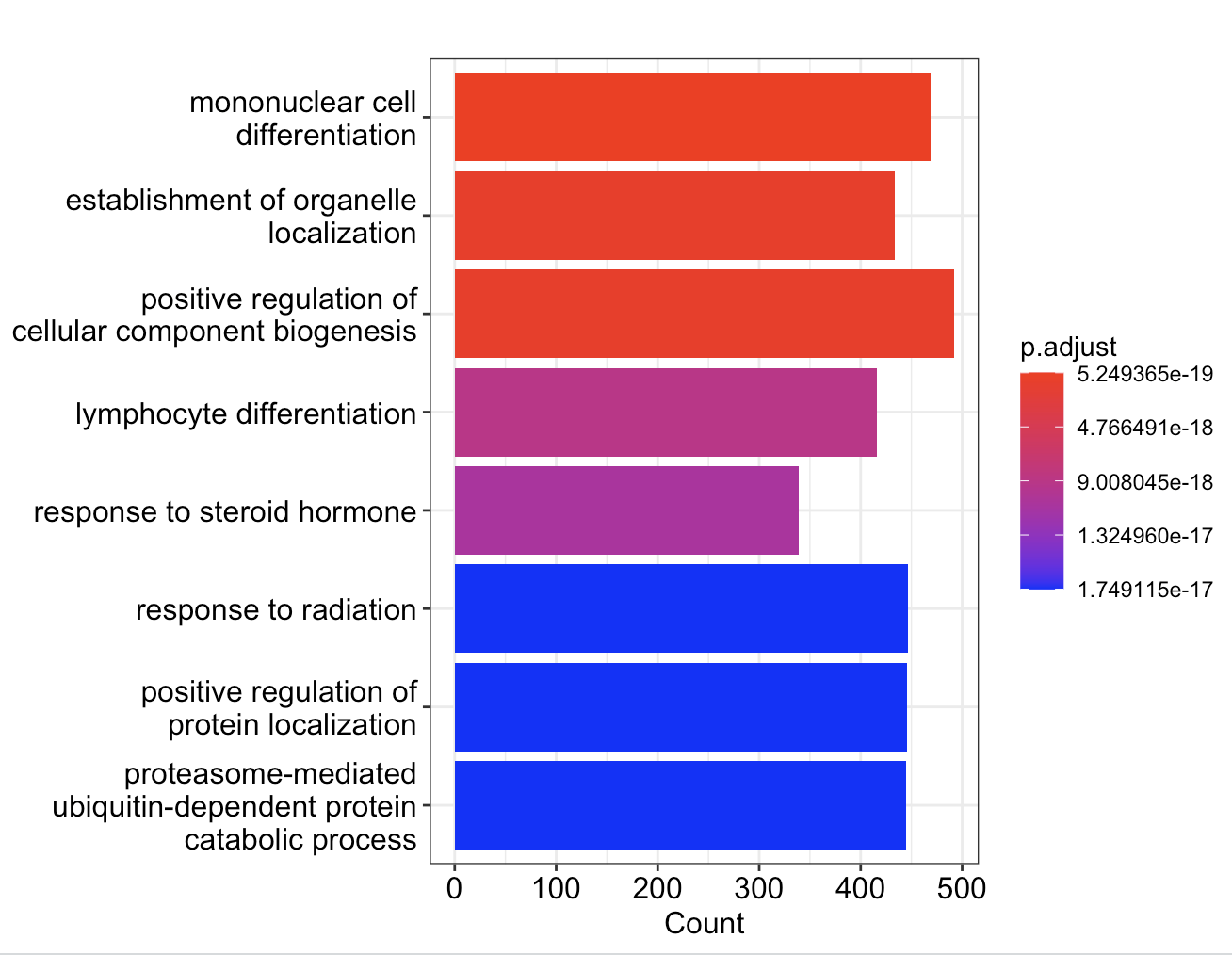


Dot plot 2 is given as:

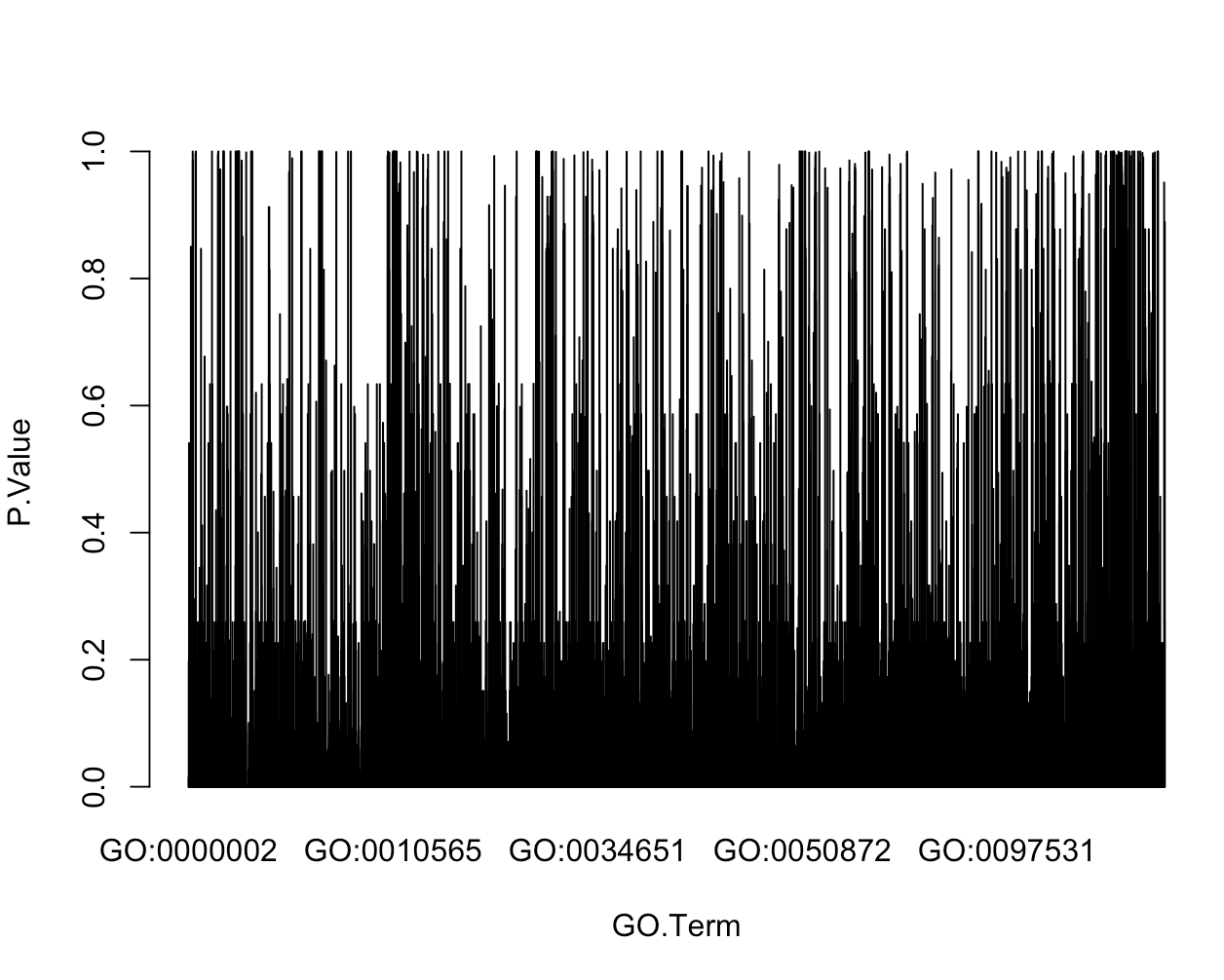


Bar chart of the p-values is created using the barplot function. The horiz parameter makes the chart horizontal, and las = 1 rotates the y-axis labels for better visibility .



BarChart 1 is given as: 

Barchart2 given as:



**9.)**

The top 10 pathways obtained through the gene set enrichment analysis are:

1. mononuclear cell differentiation

2. establishment of organelle localization

3. positive regulation of cellular component biogenesis

4. lymphocyte differentiation

5. response to steroid hormone

6. response to radiation

7. positive regulation of protein localization

8. proteasome-mediated ubiquitin-dependent protein catabolic process

9. cellular component disassembly

10. gland developmentThe top pathways obtained from the analysis provide insights into the biological processes that are enriched in the gene list. Pathway 1, mononuclear cell differentiation, suggests that the gene set is involved in the differentiation of mononuclear cells, a type of white blood cell important in immune function. Pathway 2, establishment of organelle localization, suggests that the gene set is involved in the regulation of cellular organelles, which are important for maintaining cellular function and preventing cellular damage.

Pathways 3 and 4, positive regulation of cellular component biogenesis and lymphocyte differentiation, respectively, suggest that the gene set is involved in the regulation of cellular component biogenesis and the differentiation of lymphocytes, both of which are crucial for the development and maintenance of a healthy immune system.

Pathway 5, response to steroid hormone, suggests that the gene set is involved in the cellular response to steroid hormones, which play a role in regulating growth, development, and metabolism.

Pathway 6, response to radiation, suggests that the gene set is involved in the cellular response to radiation exposure, which can cause DNA damage and other cellular damage.

Pathway 7, positive regulation of protein localization, suggests that the gene set is involved in the regulation of protein localization within cells, which is important for maintaining cellular function and preventing cellular damage.

Pathways 8 and 9, proteasome-mediated ubiquitin-dependent protein catabolic process and cellular component disassembly, respectively, suggest that the gene set is involved in the breakdown and removal of damaged or unwanted cellular structures.

Finally, pathway 10, gland development, suggests that the gene set is involved in the development of specialized organs that produce and secrete substances such as hormones or digestive enzymes, which are important for regulating growth, development, and metabolism.

Overall, the pathways identified in this analysis suggest that the given gene set is involved in various biological processes, including immune system regulation and function, cellular regulation and repair, and gland development. By understanding the molecular mechanisms underlying these processes, researchers may gain insights into potential therapeutic targets for various diseases.

Each of these processes represents a collection of genes that are involved in specific cellular functions or activities. Here's a brief description of each pathway and its potential biological significance:

1. Mononuclear cell differentiation: This pathway involves the differentiation of mononuclear cells, which are a type of white blood cell that plays a role in immune function. This process may be important for the development and maintenance of a healthy immune system.
2. Establishment of organelle localization: This pathway involves the localization of cellular organelles, such as the mitochondria or endoplasmic reticulum, to specific regions within the cell. This process is important for maintaining cellular function and preventing cellular damage.
3. Positive regulation of cellular component biogenesis: This pathway involves the regulation of cellular component biogenesis, which includes the synthesis and assembly of proteins, organelles, and other cellular structures. This process may be important for maintaining cellular function and responding to changes in the environment.
4. Lymphocyte differentiation: This pathway involves the differentiation of lymphocytes, which are a type of white blood cell that plays a role in immune function. This process may be important for the development and maintenance of a healthy immune system.
5. Response to steroid hormone: This pathway involves the response of cells to steroid hormones, which are important signaling molecules that regulate many cellular processes. This process may be important for regulating growth, development, and metabolism.
6. Response to radiation: This pathway involves the cellular response to radiation exposure, which can cause damage to DNA and other cellular structures. This process may be important for protecting cells from damage and preventing the development of cancer.
7. Positive regulation of protein localization: This pathway involves the regulation of protein localization within cells. This process is important for maintaining cellular function and preventing cellular damage.
8. Proteasome-mediated ubiquitin-dependent protein catabolic process: This pathway involves the breakdown of proteins by the proteasome, a large protein complex that is responsible for degrading unwanted or damaged proteins. This process is important for maintaining cellular function and preventing the accumulation of toxic proteins.
9. Cellular component disassembly: This pathway involves the disassembly of cellular components, such as organelles or protein complexes. This process may be important for responding to changes in the environment or for the removal of damaged or unwanted cellular structures.
10. Gland development: This pathway involves the development of glands, which are specialized organs that produce and secrete substances such as hormones or digestive enzymes. This process may be important for regulating growth, development, and metabolism.

Overall, these pathways represent a diverse range of cellular processes that may be important for maintaining cellular function, responding to changes in the environment, and regulating growth and development. By analyzing the genes that are involved in these processes, researchers can gain a better understanding of the underlying molecular mechanisms and potential therapeutic targets for various diseases.