**STAGE 2 TASK - GENE EXPRESSION AND FUNCTIONAL ENRICHMENT ANALYSIS**

**Contributors : -** Keshav Pande ([keshavpande1603@gmail.com](mailto:keshavpande1603@gmail.com)) Flora ([floraoladipupo@gmail.com](mailto:floraoladipupo@gmail.com)) Abdulqudus Oladega ([oladegakenny@gmail.com](mailto:oladegakenny@gmail.com))

**INTRODUCTION**

Glioblastoma is a complex malignancy of the brain with intricate genetic makeup. Gene expression patterns though can provide an understanding of its biological mechanisms. This project seeks to undertake a bioinformatics analysis of the dataset containing of over 500 genes regulated in glioblastoma.

**Objective:**

* This aims at producing heat maps from the gene expression data collected, and ascertaining the patterns of the up and down-regulated genes as well as carrying out functional enrichment analysis. This is important as it helps to provide a biological context of the differentially expressed genes and the pathways they represent.

**Dataset:**

* The dataset consists of the top 500+ conditionally differential expressed genes. Each row corresponds to a separate gene while each column gives the levels of gene expression in the different samples for the column condition (in this case glioblastoma related). The values in the dataset were transformed log2 to minimize the variation.

**Data Preprocessing**

**Data source:** [glioblastoma.csv](https://raw.githubusercontent.com/HackBioInternship/public_datasets/main/Cancer2024/glioblastoma.csv)

* The raw counts from the dataset are first transformed to log2 scale, which stabilizes the variance and reduces the impact of extreme values. This makes the data more suitable for statistical analysis, especially when comparing gene expression levels across different sample conditions.

**Normalization:**

* The gene expression data was normalized using log2 transformation. This transformation reduces the degree of variability and facilitates the discerning of trends in the data.

**# Load the dataset**

genes\_data<read.csv('https://raw.githubusercontent.com/HackBioInternship/public\_datasets/

main/Cancer2024/glioblastoma.csv', row.names = 1)

**# Convert data to matrix and apply log2 transformation to stabilize variance**

genes\_matrix <- log2(as.matrix(genes\_data) + 1)

**Heatmap Visualization**

* **Sequential color palette**: Helps to represent continuous data (e.g., gradual increase in expression).
* **Diverging color palette:** Shows the contrast between upregulated (high expression) and downregulate

(low expression) genes, making it easier to spot differential expression.

1) **Sequential Color Palette Heatmap:**

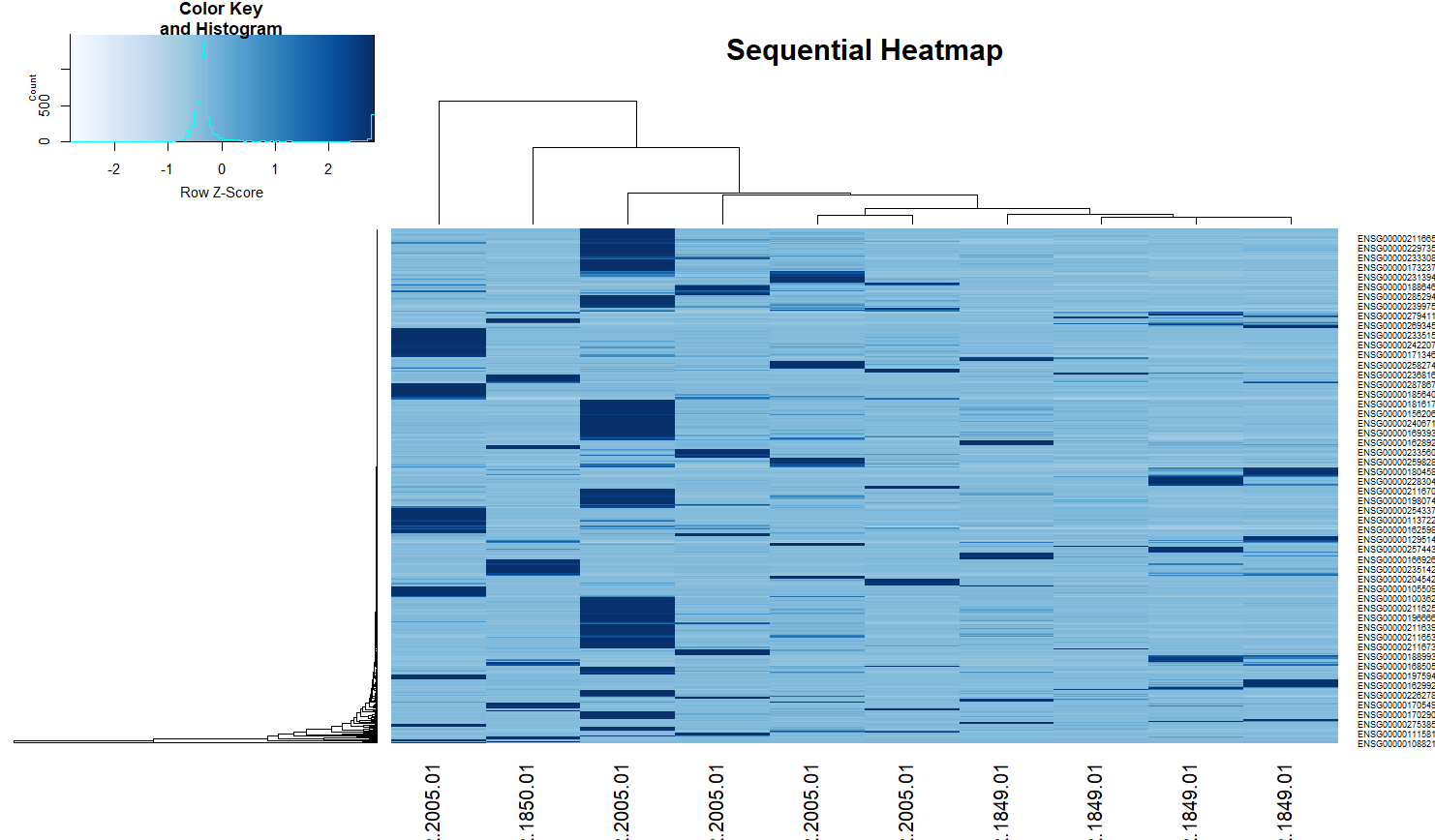
A heatmap was generated using a sequential color palette, which helps display a range of gene expression levels using shades of the same color.

**# Sequential color palette**

**seq\_pal <- colorRampPalette(brewer.pal(n=9, name="Blues"))(100)**

**# Generate Sequential heatmap**

**heatmap.2(genes\_matrix,co l=seq\_pal, scale="row", trace="none", main="Sequential Heatmap")**



**Figure 1 :** *Heatmap of Sequential Color Palette*

2 ) **Diverging Color Palette Heatmap:**

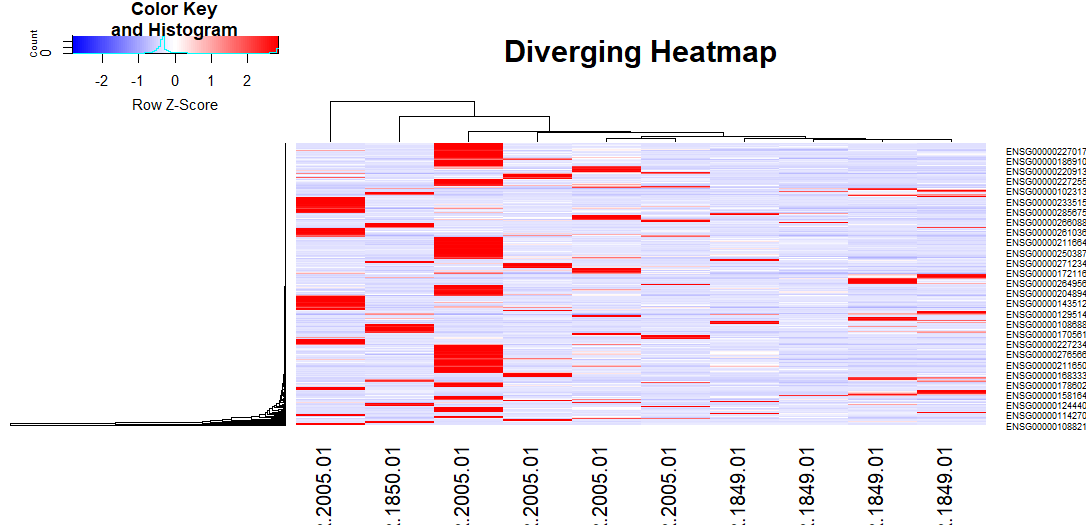
The diverging color palette heatmap helps highlight differences in gene expression between the samples by using contrasting colors.

**# Sequential color palette**

**div\_pal <- bluered(100)**

**# Generate Sequential heatmap**

**heatmap.2(genes\_matrix,co l=div\_pal, scale="row", trace="none", main="Diverging Heatmap")**

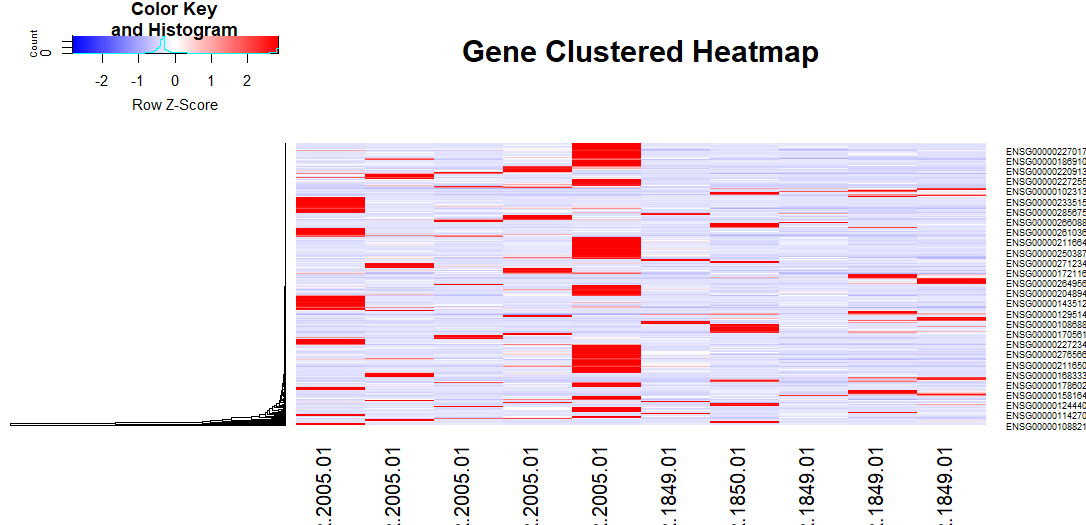


**Figure 2 :** *Heatmap of Diverging Color Palette*

**3) Clustering in Heatmaps:**

The dataset was further analyzed by clustering both genes (rows) and samples (columns) to identify groups of genes with similar expression patterns.

* **Gene Clustering**: Clusters genes based on similar expression patterns.
* **Sample Clustering**: Clusters samples to identify similar conditions.
* **Both Genes and Samples Clustering**: Clustering both allows for a better understanding of relationships between samples and genes.



**# Clustering heatmaps**

**#Row clustering only**

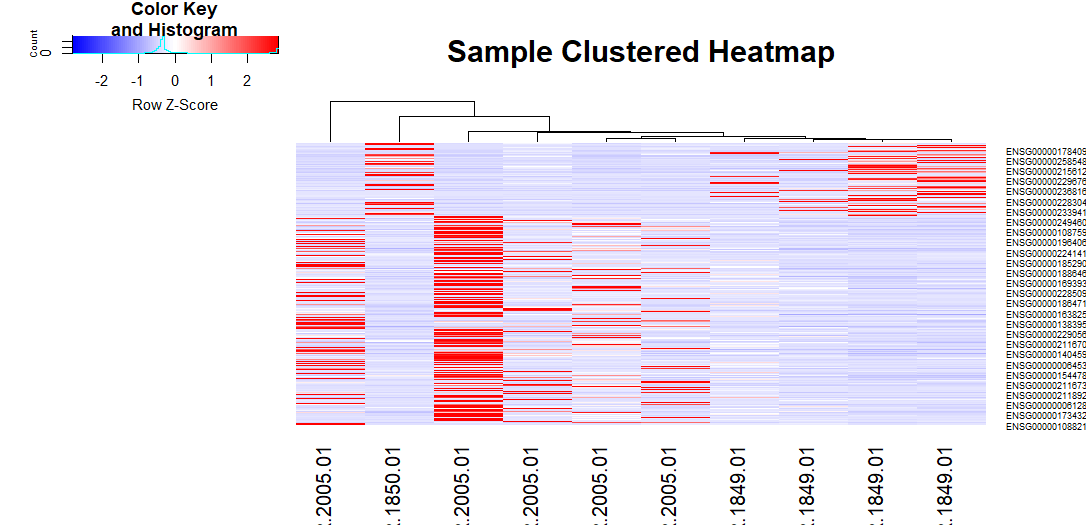
**heatmap.2(genes\_matrix, Rowv=TRUE, Colv=FALSE, dendrogram="row", col=div\_pal, scale="row", trace="none", main="Gene Clustered Heatmap")**

**# Column clustering only**

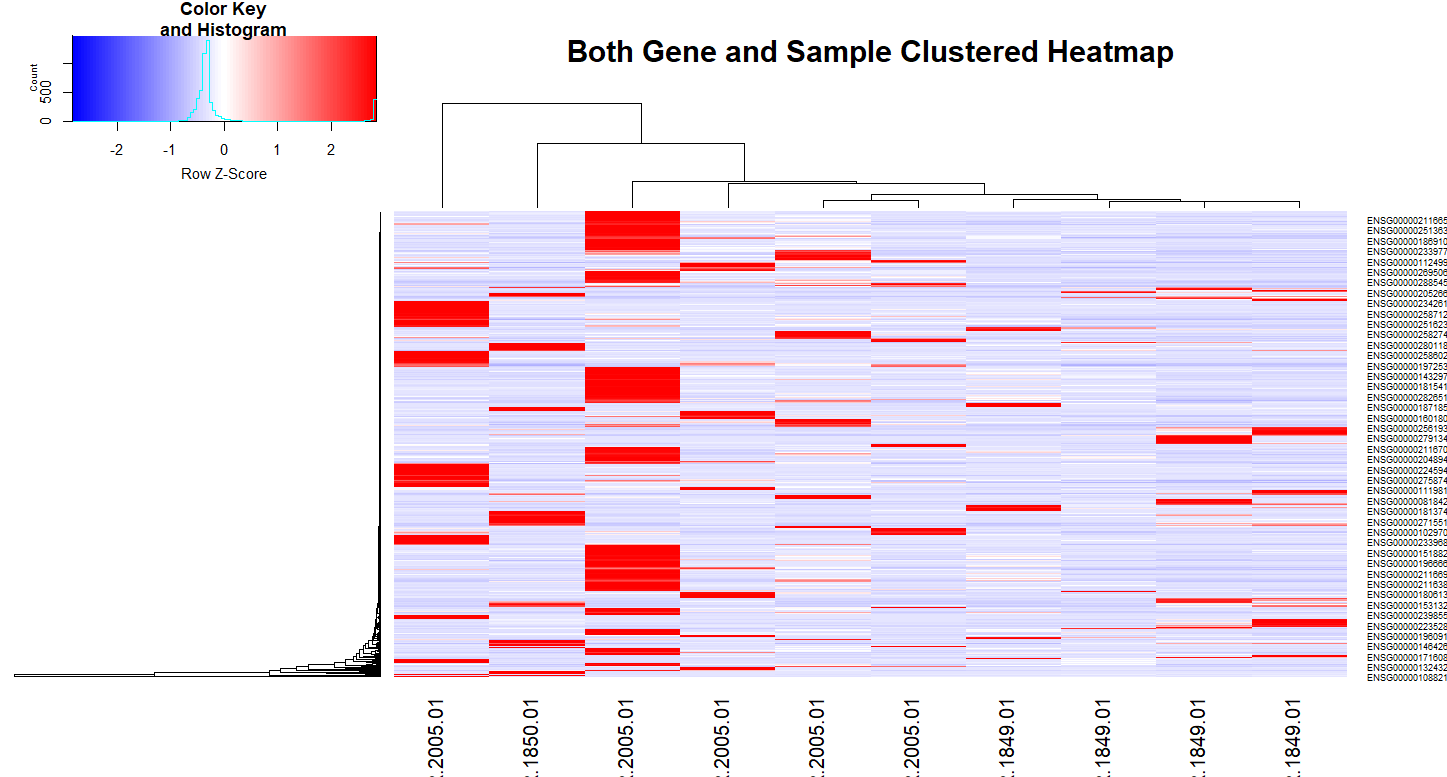
**heatmap.2(genes\_matrix, Rowv=FALSE, Colv=TRUE, dendrogram="col", col=div\_pal, scale="row", trace="none", main="Sample Clustered Heatmap")**

**# Both row and column clustering**

**heatmap.2(genes\_matrix, Rowv=TRUE, Colv=TRUE, dendrogram="both", col=div\_pal, scale="row", trace="none", main="Both Gene and Sample Clustered Heatmap")**

**Fig.3 Gene** Clustered Heatmap

**Fig.4** Sample Clustered Heatmap

 **Fig.4** Both Gene and Sample Clustered Heatmap

**4. Identification of Differentially Expressed Genes**

4. Distribution of Fold Change

group\_1 <- genes\_matrix[, 1:5] # Select the first 5 columns as Group 1

group\_2 <- genes\_matrix[, 6:10] # Select the next 5 columns as Group 2

group\_1\_mean <- apply(group\_1, 1, mean)

group\_2\_mean <- apply(group\_2, 1, mean)

fold\_change <- log2(group\_1\_mean) - log2(group\_2\_mean)

hist(fold\_change, xlab = "log2 Fold Change (Group 1 vs Group 2)",

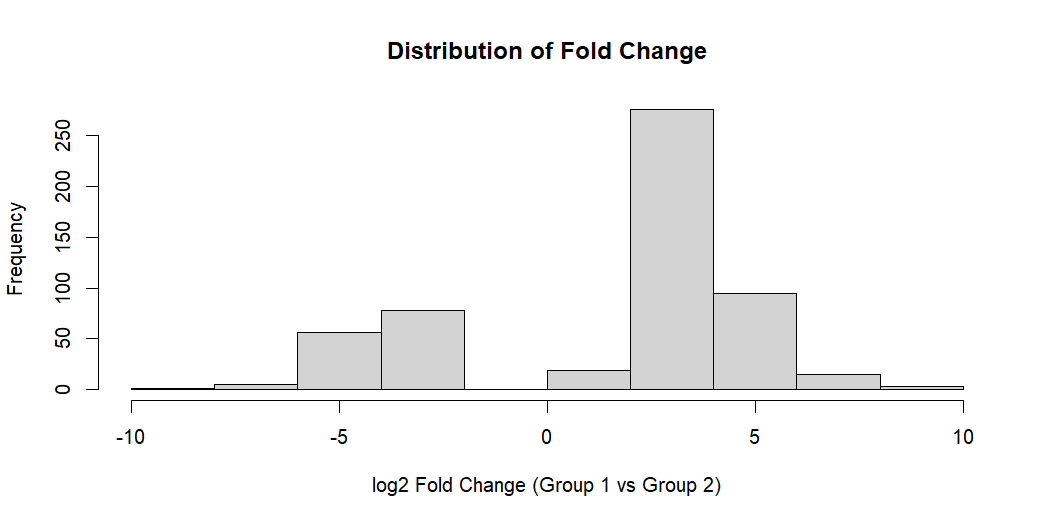
main = "Distribution of Fold Change")

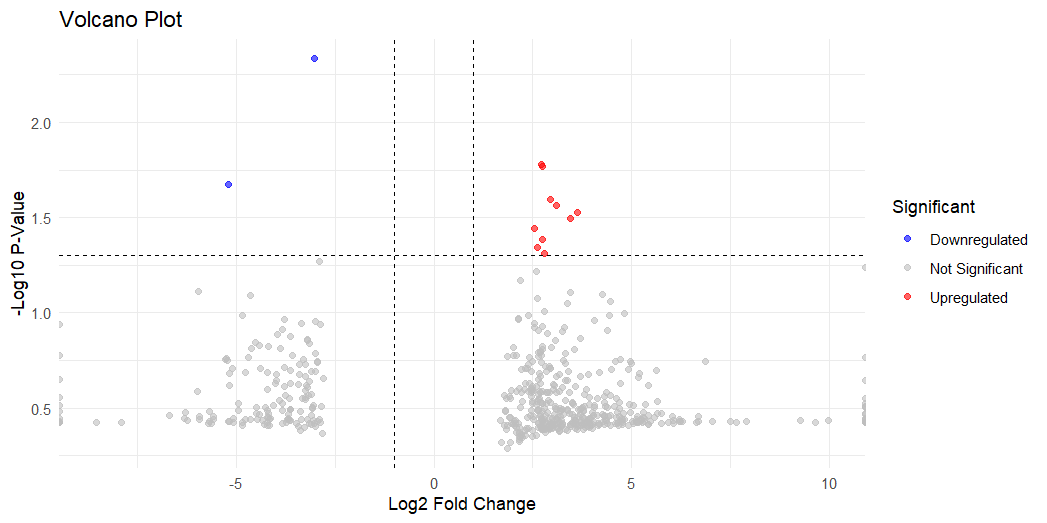
To further investigate the data, the differentially expressed genes were identified. The primary metrics used are:

* **Log2 Fold Change (log2FC)**: Used to measure changes in gene expression between two groups of samples.
* **p-values**: Measure the statistical significance of these changes.

Genes with:

* **log2FC > 1 and p-value < 0.05** were considered **upregulated**.
* **log2FC < -1 and p-value < 0.05** were considered **downregulated**.



**5 . Volcanao Plot**

Volcano Plot

ggplot(significant\_genes, aes(x = log2FoldChange, y = -log10(pValue), color = Significant)) +

geom\_point(alpha = 0.6, size = 1.5) +

scale\_color\_manual(values = c("Upregulated" = "red", "Downregulated" = "blue", "Not Significant" = "grey")) + theme\_minimal() +

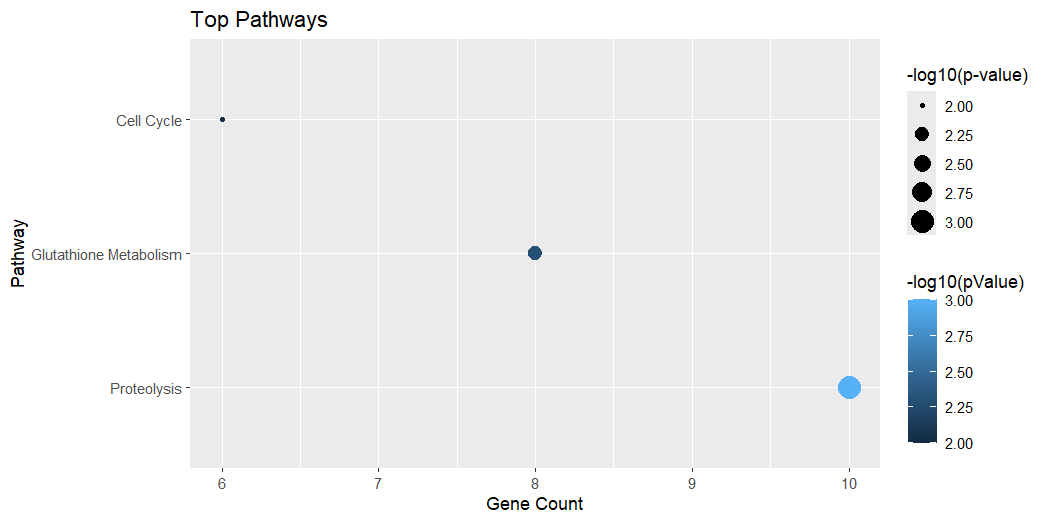
labs(title = "Volcano Plot",

x = "Log2 Fold Change",

y = "-Log10 P-Value") +

geom\_hline(yintercept = -log10(p\_value\_threshold), linetype = "dashed", color = "black") +

geom\_vline(xintercept = c(-fold\_change\_threshold, fold\_change\_threshold), linetype = "dashed", color = "black")ggsave("volcano\_plot.png", width = 10, height = 6)

**6. lollipop chart**

top\_pathways <- data.frame(Pathway=c("Proteolysis", "Glutathione Metabolism", "Cell Cycle"), nGenes=c(10, 8, 6), pValue=c(0.001, 0.005, 0.01))

ggplot(top\_pathways, aes(x=reorder(Pathway, -nGenes), y=nGenes)) +

geom\_point(aes(size=-log10(pValue), color=-log10(pValue))) +

coord\_flip() +

labs(title="Top Pathways", x="Pathway", y="Gene Count", size="-log10(p-value)")

**Conclusion**

**1)GO:0006508 - Proteolysis**

Proteolysis is a fundamental biological process involved in the breakdown of proteins into smaller peptides or amino acids, often mediated by proteases. In glioblastoma, this pathway is crucial for tumor progression due to the role of proteolytic enzymes in:

* Degrading the extracellular matrix (ECM): Glioblastoma cells overexpress matrix metalloproteinases (MMPs), which degrade ECM components, facilitating the invasion of cancer cells into surrounding tissues. Increased MMP expression is correlated with poor prognosis and aggressive tumor behavior .
* Activating growth factors: Proteolysis also contributes to activating pro-tumorigenic growth factors such as transforming growth factor-beta (TGF-β), further driving glioblastoma cell proliferation and invasion .
* Dysregulated apoptosis: Caspase-dependent proteolysis, essential in normal apoptosis, is often dysregulated in glioblastoma, leading to evasion of cell death, which contributes to therapy resistance .

**2 ) GO:1901685 - Glutathione Derivative Metabolic Process**

This pathway includes the metabolism of the glutathione (GSH) derivatives and its metabolic processes, which are important for cell detoxification and the redox balance in cells of glioblastoma tumors. Some of them are:

* **Redox regulation:** Bacterial cells possess the ability to generate fast-proliferating reactive oxygen species bacteria (ROS). As a result of this oxidative stress, glutathione system including GSH and its derivatives gets upregulated and this upregulation provides a selective advantage for survival .
* **Modeling Chemotherapy Resistance in Human Glioblastomas**: One reason might be scavenging of the drugs by elevated glutathione metabolism in glioblastoma, as GSH is known to protect against oxidative stresses caused by drugs. This is one of the reasons why temozolomide and radiotherapy fail .
* **Glutathione S-transferases (GSTs):** Along with this, gliobatoma cells also overexpress other glutathione detoxifying enzymes, whose specific relevance is the ability to conjugate harmful electrophilic compounds, which improves further the detoxification and resistance to **chemotherapy.**

**3 ) 3. GO:1901687 - Glutathione Derivative Biosynthesis**

This pathway encompasses the synthesis processes of glutathione derivatives which is important in the regulation of redox balance and oxidative stress in gbm:

* GSH conjugates: Glioblastoma cells promote the synthesis of GSH conjugates to detoxify the metabolites and chemotherapeutic drugs. This overexpression is an important strategy employed by glioblastoma to evade treatment.
* NADPH dependence: The NADPH-producing biosynthetic machinery is responsible for these cellular reductions and this cofactor nadph comes from sources such as the pentose phosphate pathway ppp. NADPH is often produced in excess in glioblastoma cells to meet the increased requirement of GSH biosynthesis.
* Metabolic reprogramming: Glucose metabolic reprogramming is characteristic of glioblastoma as well as other processes such as the enhancement of the synthesis of glutathione derivatives. These processes are aimed at supporting the antioxidant capacity of the tumor which is another aspect of its aggressiveness.

**References:**

1. Aaberg-Jessen, C., et al. (2009). "Invasion of glioma cells is regulated by multiple matrix-degrading enzymes and is directly correlated to MMP-2 and MMP-9 activity." *J Neuropathol Exp Neurol*.
2. Hagemann, C., & Blank, J. L. (2013). "The role of MMPs in glioblastoma." *Cell Adhesion & Migration*.
3. Wick, W., et al. (2001). "TGF-β: A key oncogenic factor in glioblastoma." *Cancer Research*.
4. Bredesen, D. E. (2007). "Apoptosis: Mechanisms relevant to cancer and neurodegeneration." *Nature*.
5. Cairns, R. A., et al. (2011). "Regulation of cancer cell metabolism." *Nature Reviews Cancer*.
6. Singh, N., & Dhindsa, R. S. (2020). "Glutathione and its role in cancer drug resistance." *Cancer Metabolism*.
7. Townsend, D. M., et al. (2003). "The role of glutathione-S-transferase in cancer drug resistance." *Oncogene*.
8. Traverso, N., et al. (2013). "Role of glutathione in cancer progression and resistance." *Oxidative Medicine and Cellular Longevity*.
9. DeBerardinis, R. J., et al. (2008). "The role of metabolic reprogramming in cancer." *Oncogene*.
10. Vander Heiden, M. G., et al. (2009). "Understanding the Warburg effect: The metabolic requirements of cell proliferation." *Science*.