

MHEDAS

2024-2025

NEXT GENERATION SEQUENCING

COTM GRADED ACTIVITY 2



Author:

Ihona Maria Correa de Cabo

INTRODUCTION

The Nature Cell Biology study by Fu et al. (2015), titled "*EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival*" , explores the role of Mcl-1, a pro-survival protein, in mammary gland development, focusing on how it is regulated during lactation. The study uses mouse models and advanced molecular biology techniques to understand how mammary epithelial cells survive and function in different developmental stages.

During the study, a dataset was generated to examine the gene expression profiles of mammary gland epithelial cells under different physiological conditions, which are virgin, pregnant, and lactating mice. Moreover, it focuses on two specific cell types: basal stem-cell enriched cells (those involved in structural support and progenitor functions) and committed luminal cells (involved in milk production during lactation).

This mentioned dataset has an accession number GSE60450 and can be downloaded from the Gene Expression Omnibus database (GEO).

The goal is to use the GenewiseCounts matrix for the GSE60450 dataset to perform a differential gene expression analysis and an enrichment analysis between the three groups (virgin, pregnant and lactating mice).

This report provides a detailed overview of the steps followed and the results obtained.

EXERCISE 1

The goal of this exercise is to download the GenewiseCounts matrix for the GSE60450 dataset and prepare the sample information matrix.

The sample information matrix was taken from the graded activity statement and converted into a .txt file. This information matrix contains metadata of the samples of the dataset, mapping the cell type (luminal or basal) and the status of the mice (virgin, pregnant or lactate) to each sample.

On the other hand, the GenewiseCounts matrix or counts matrix contains raw RNA sequencing (RNA-seq) counts of gene expression for different biological samples in the study. Here, each row corresponds to a specific gene, identified by its unique Entrez Gene ID, whereas the columns represent individual RNA-seq samples. The column names are a sample identifier which correspond to a specific biological condition that can be mapped through the sample information matrix. Also, there are two extra columns in the matrix, "EntrezGeneID", which contains the unique identifier for each gene, and "Length", which is the gene length in base pairs.

The values in the matrix represent raw counts of RNA-seq reads mapped to each gene for each sample. Interestingly, these counts are proportional to the expression level of the gene in the corresponding sample.

EXERCISE 2

The main goal of this exercise is to load the information matrix and the counts matrix (both in .txt format) into Rstudio, format them, and perform a correlation analysis and a Principal Component Analysis (PCA).

After loading both matrices, it is checked whether the information matrix (or metadata) aligns with the counts matrix.

Then, a quality control is performed to measure the similarity of the samples with each other. Samples that are supposed to be biological replicates (for example, samples from the same group or condition) should look very similar to each other in terms of gene expression.

This similarity is measured through a correlation analysis and a PCA.

Figure 1 shows a correlation heatmap plot from the counts matrix obtained using the *pheatmap* library. *Figure 2* represents the same correlations but uses information from the metadata file in order to distinguish between the different status of the mice. The dendrogram groups samples based on how similar they are, and samples that are most similar are connected by short branches.

Finally, *Figure 3* uses the *corrplot* package to plot a more appealing pairwise correlation matrix.

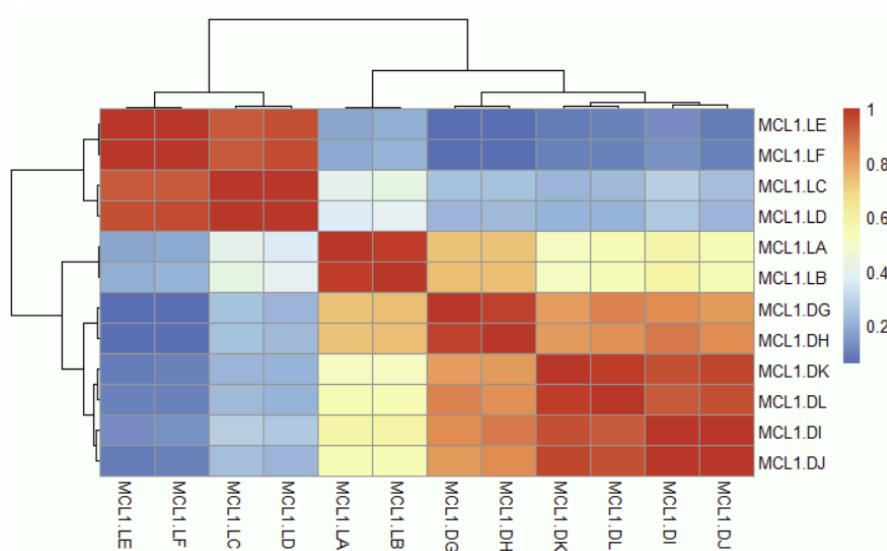


Figure 1.- Correlation heatmap from the counts matrix without “EntrezGeneID” and “Length” columns.

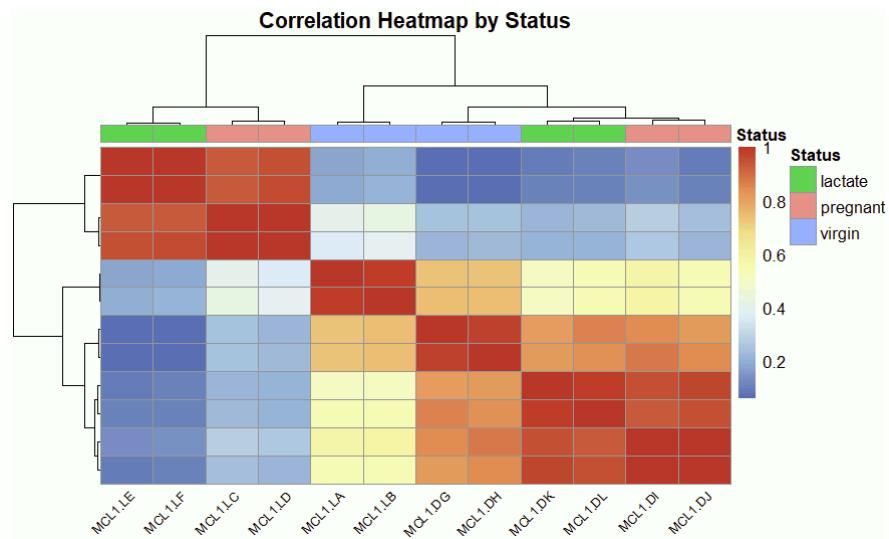


Figure 2.- Correlation heatmap from the counts matrix without “EntrezGeneID” and “Length” columns and divided by status.

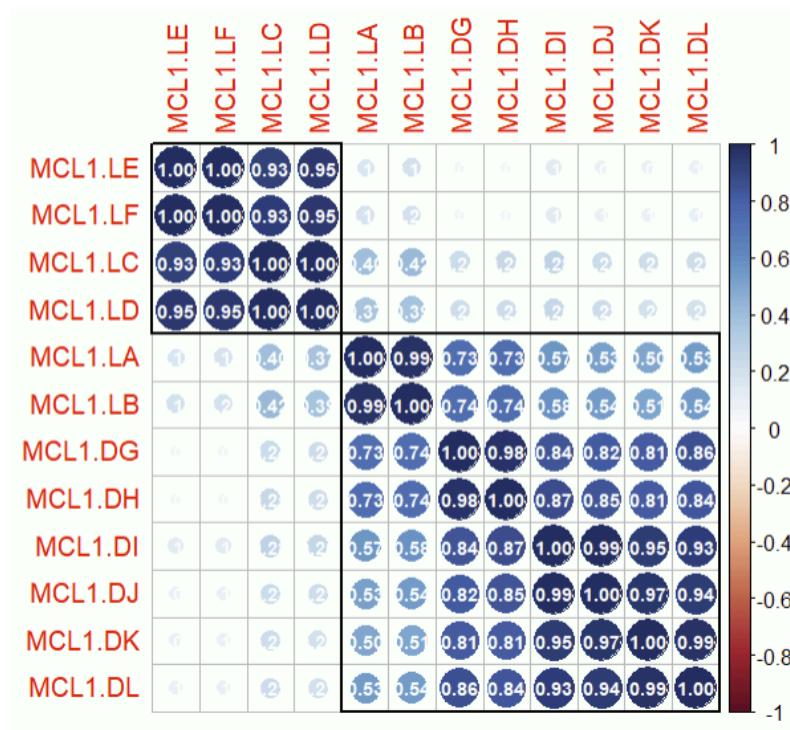


Figure 3.- Pairwise correlation matrix from the counts matrix . Correlation coefficients appear as numbers in white color.

From the different correlation plots and heatmaps, it is possible to see that samples from the same status (virgin, pregnant, or lactating) tend to have higher correlation.

The high correlation between samples suggests a strong similarity in gene expression profiles, while lower correlation highlights the biological differences in gene expression between these conditions (maybe due to developmental changes in the mammary gland).

The fact that biological replicates cluster together was expected and supports the quality of data.

Taking a deeper overview on the correlation between the different status, it is observed a moderate correlation between lactating and pregnant mice, which may indicate some possible biological overlap in these stages. Nevertheless, pregnant samples versus virgin ones showed reduced correlation, and lactating versus virgin even lower correlation. This shows significant biological differences in the mammary gland as it transitions from a virgin state to a lactation state.

Another technique to assess similarity is the PCA, as mentioned before, which is a dimensionality reduction technique that can be used to visualize patterns.

Before building the PCA, the counts matrix was transposed in order to have the genes (our variables) as columns and the samples (our observations) as rows.

This transposed matrix was then converted to the log2 scale to make the data less skewed and easier for PCA to handle.

Figure 3 represents a two dimension scatter plot of the first two principal components.

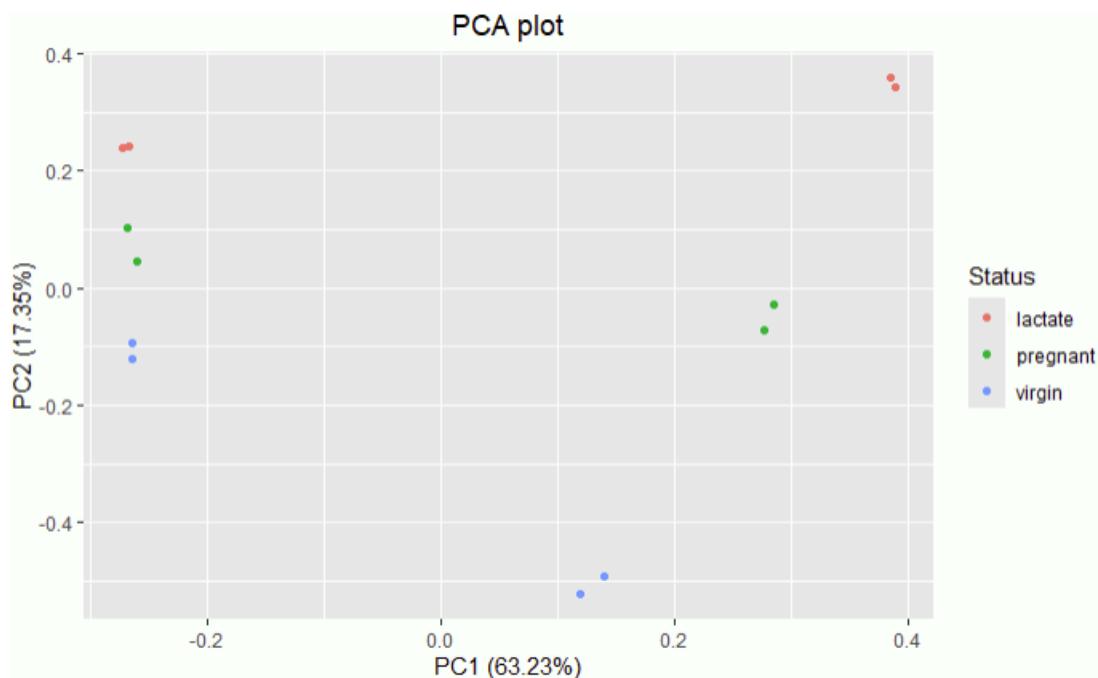


Figure 4.- PCA plot of the first two principal components. Each dot represents a single sample color-coded by Status.

It is possible to see that samples group closely together based on their status, which may indicate distinct gene expression profiles for each biological condition. Moreover, the pregnant group acts as a transition state between virgin and lactating states.

The separation observed in the PCA is likely due to the different cell type of each sample.

EXERCISE 3

In this exercise, a differential expression analysis (DEA) between virgin and pregnant mice is carried out assuming the virgin samples as the reference ones. This analysis identifies the genes whose expression levels differ significantly between the two groups by testing against the null hypothesis that the activity of the gene stays the same in two different conditions. Differentially Expressed Genes (DEGs) can be either upregulated (higher expression in one group) or downregulated (lower expression in one group).

It is important to mention that, in order to perform this analysis, the counts table should have the raw read counts as integers, without any normalization. Moreover, the row names from the metadata should match and should be in the same order as the column names from the counts matrix.

Afterwards, a volcano plot showing gene symbols and distinguishing between UP- and DOWN-regulated genes is performed.

The first step to perform this is to filter the metadata to keep only the samples from virgin and pregnant status. The file names from the filtered metadata are then used to filter the counts matrix to keep only the desired samples.

Then, a DESeqDataset object is constructed using both the counts and the metadata. This object contains all the information about the experimental setup, the read counts and the design formulas.

The DESeqDataset object (dds) is pre-filtered in order to remove genes that have almost no information in any given sample. Moreover, it is specified to use the virgin samples as the reference level, meaning that they will serve as the baseline or control group against which the gene expression levels of the pregnant mice samples are compared. Therefore, all the statistical tests will be calculated relative to this virgin group.

Next, DESeq() wrapper function is used to run the full DEA workflow, which includes:

1. Read counts normalization by computing size factors.
2. Dispersion estimation for each gene.
3. Negative binomial model fit for each gene.
4. Testing for differential expression.

The resulting output after the analysis contains 6 columns:

- baseMean: The average normalized gene counts across all samples.
- log2FoldChange: How much a gene's expression changes between two conditions (in this case "virgin" vs "pregnant"). Positive values mean the gene is more highly expressed in "virgin," and negative values mean it is more expressed in "pregnant." The values are on a log2 scale.
- IfcSE: The standard error for the fold change estimate.

- stat: The test statistic from the Wald test, which checks whether the fold change is significantly different from zero.
- pvalue: The probability that the observed difference is due to chance (from the Wald test) for each gene.
- padj: The p-value adjusted for multiple comparisons using the Benjamini-Hochberg method, which controls the false discovery rate (FDR).

Before visualizing and interpreting the results from the analysis, some diagnostic plots have been performed to improve the confidence about the quality of the data as well as the experimental setup.

In this case, the diagnostic plots have been a MA plot (*Figure 5*) and a p-value distribution plot (*Figure 6*).

The MA plot is a scatter plot where the X-axis corresponds to the average of normalized counts across samples and the Y-axis the log fold change. It is mainly used to check whether the data normalization worked. Following *Figure 6*, it is possible to see that most points are in the horizontal line (around 0). This was expected since most genes are not differentially expressed.

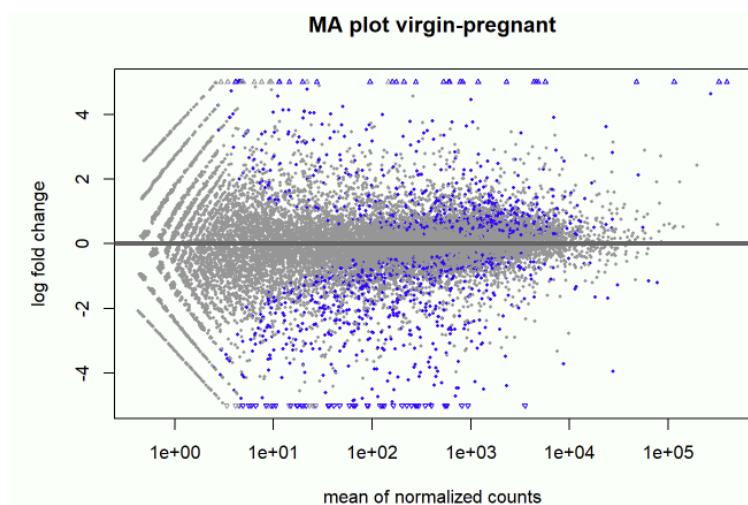


Figure 5.- MA plot for the virgin and pregnant samples for the dds object.

For each gene, the p-value comes from a hypothesis test that evaluates whether the observed expression difference (for example log2 fold change) is significantly different from zero or not.

If no genes are differentially expressed, the raw p-values distribution would be a flat line between 0 and 1, which is not the case.

The observable peak from *Figure 6* at very small p-values (around 0) suggests that many genes show statistically significant changes in their expression, which means the results have been meaningful.

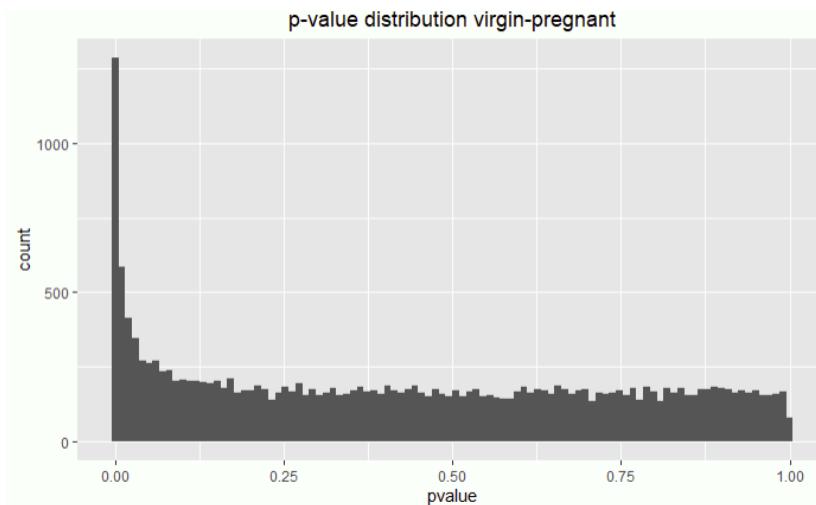


Figure 6.- p-value distribution of the DEA results for virgin and pregnant samples.

At this point, the DEA results are visualized through Volcano plots, which are scatter plots that provide a quick way to identify significant genes and visualize patterns in the data. The X-axis typically contains the Log2 Fold Change, representing a magnitude of change in gene expression between two conditions.

The Y-axis represents the adjusted p-value in the log scale, measuring the statistical significance of the change. In this case, genes with smaller p-values appear higher on the plot, indicating they are more likely to be differentially expressed. In summary, genes that appear in the right side of the plot are significantly upregulated because they have high log2 fold change and low p-value. In contrast, genes in the left are downregulated and the ones in the center do not experience any significant change.

Figure 7 represents the volcano plot for the virgin and pregnant samples. The thresholds *F2CLim* (Fold Change Limit) and *PadjLim* (Adjusted p-value Limit) were used to classify the genes into categories based on their statistical significance and magnitude of change.

A gene is classified as “up-regulated” if its p-value is lower than 0.05 and the fold change is greater than 2 (which means the expression is at least 4 times higher in one condition).

However, if the gene is statistically significant also but it had a fold change lower than -2, it is classified as “down-regulated”. The rest of the genes that do not meet any criteria

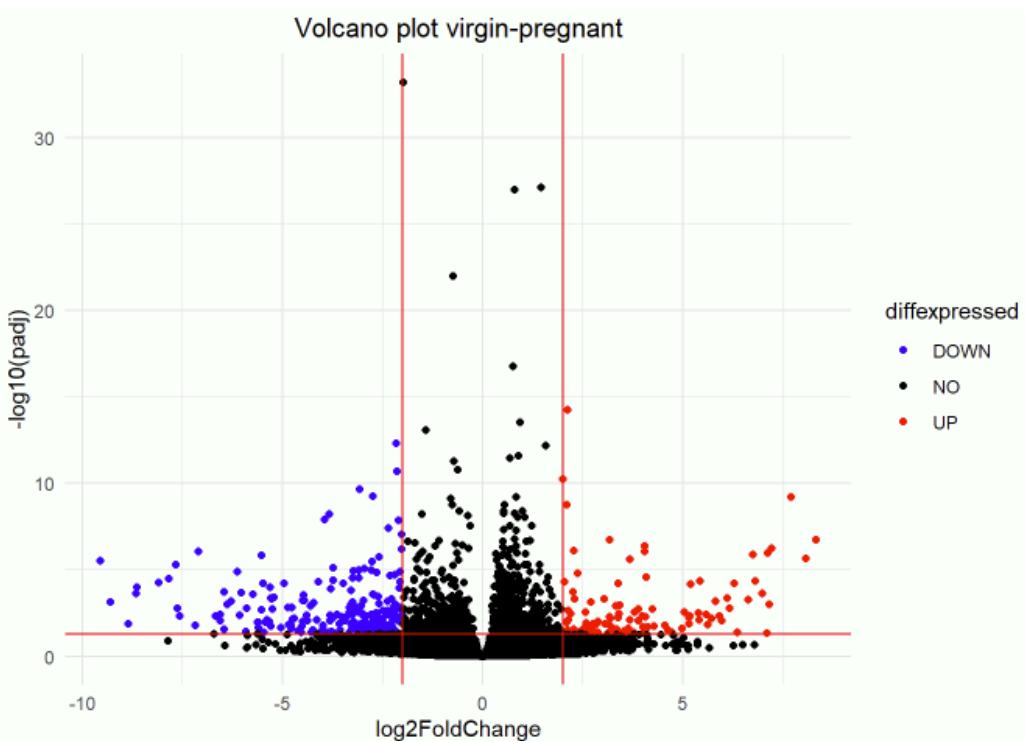


Figure 7.- Volcano plot for the virgin and pregnant samples, color-coding genes if they are down (blue), up (red) or not regulated (black).

In this context, the up-regulated genes are those expressed at higher levels in pregnant samples compared to virgin samples, and the down-regulated ones are expressed at lower levels. For the not regulated genes, they do not show significant differences in expression between virgin and pregnant samples.

Figure 8 contains the exact Volcano plot as before but with the respective annotated name for each gene. For this purpose, it was needed to load the *org.Mm.eg.db* package, which contains gene annotation data for the mouse genome. This database allows the mapping between different types of gene identifiers (Entrez IDs in this case). A new column named *GenSymbol*, which represents the gene symbol for each Entrez Gene ID, is used for the plotting.

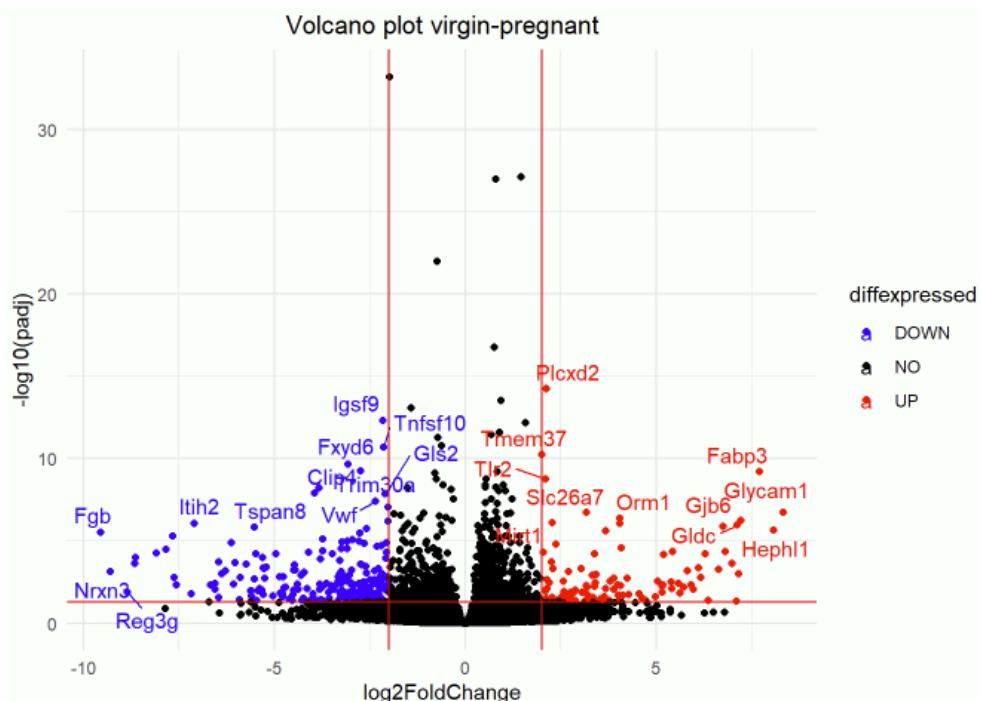


Figure 8- Volcano plot for the virgin and pregnant samples. Genes contain their respective name and are color-coded if they are differentially expressed.

A dataframe containing the results of up and down regulated genes is saved for downstream analysis.

EXERCISE 4

This section is devoted to performing an enrichment analysis of the significantly differentially expressed genes obtained from the previous exercise.

The enrichment analysis is useful to determine whether certain biological processes, molecular functions, or pathways are significantly associated with the up or down regulated genes. In a sense, it helps in connecting gene expression changes to higher-level biological insights.

The analysis works by comparing a list of differentially expressed genes with predefined gene sets from certain databases. In this exercise, Gene Ontology database is used for this end.

Gene Ontology (GO) provides a structured vocabulary that is very useful to describe the roles of genes and their products (like proteins) in a consistent and standardized way. It basically organizes biological knowledge into three main categories: biological process (BP), molecular function (MF) and cellular component (CC).

The main objective is to use GO terms to find out changes that can describe differences between the two groups of samples.

In this section, two GO enrichment analysis will be performed, one for down regulated genes and another for up regulated genes.

In both cases, the first step is to ensure that the gene identifiers used (EntrezGeneID) are valid and consistent with the mouse annotation database. To this end, org.Mm.egGO package is again loaded and stored in order to perform a mapping.

Then, by using the `enrichGO()` function, the enrichment analysis is performed. It is specified to filter out missing IDs and to retrieve all three GO ontologies mentioned before. The analysis tests whether the differentially expressed genes are associated with specific GO terms compared to the background gene universe.

The result of the analysis is a data object that contains enriched GO terms, the associated statistics, and the genes contributing to each term.

Finally, redundant GO terms are removed and then the `enrichplot` package is used to implement visualizations that can help interpreting the enrichment results.

Figure 9 and *Figure 10* contain the dot plots for down and up regulated genes. Each row of the Y-axis contains a specific enriched GO term and is divided into the 3 extracted categories (BP, CC and MF). If a specific GO term is associated with more genes in the input list than it would be expected by chance, it is considered as an enriched term.

Then, the X-axis reflects the gene ratio, calculated dividing the number of input genes associated with a GO term divided by the total input genes.

The size of the dots indicates the number of upregulated or downregulated genes associated with the term, whereas the dot color shows the statistical significance of the enrichment.

The plots display the top 5 enriched terms in each ontology.

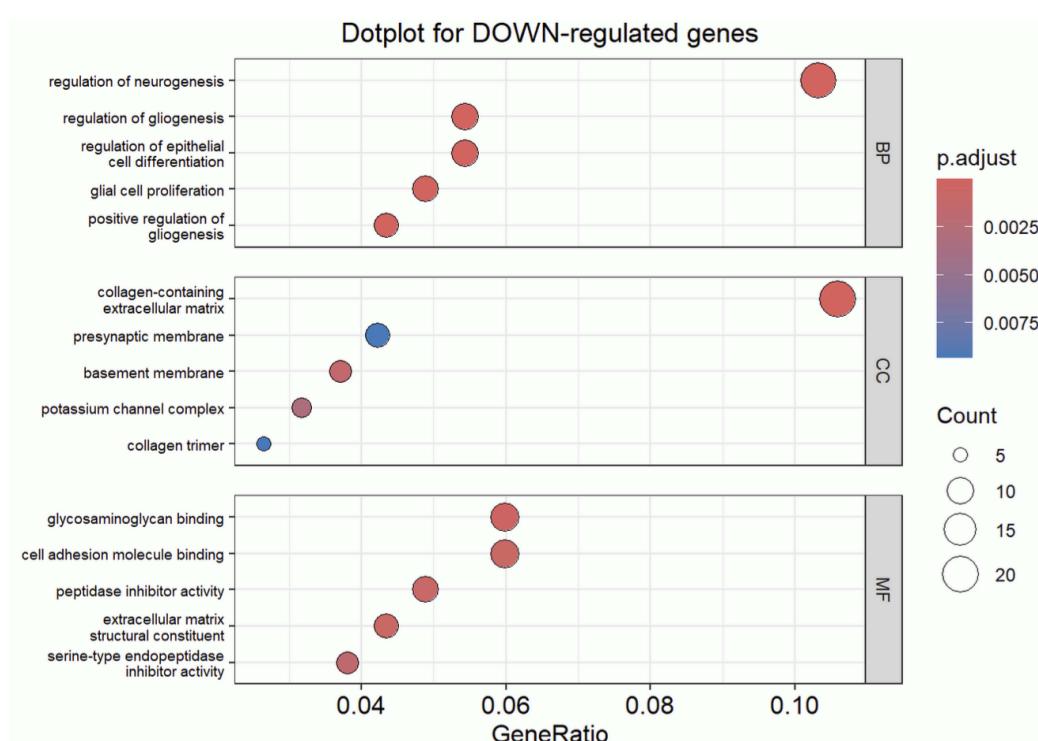


Figure 9- Dot plot for down regulated genes of virgin and pregnant samples.

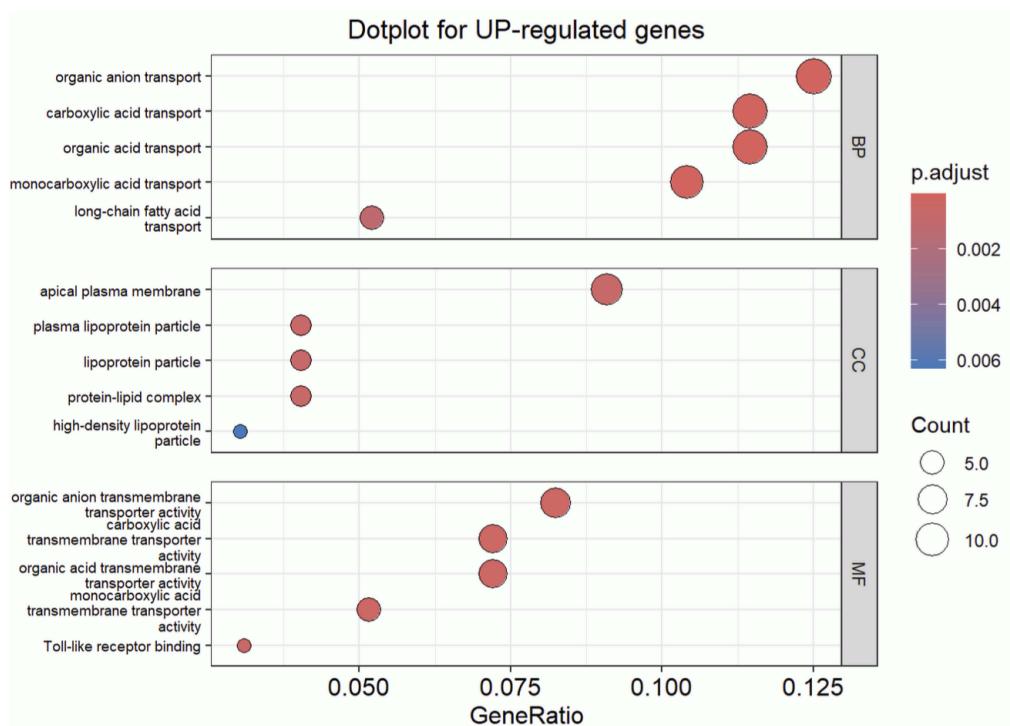


Figure 10- Dot plot for upregulated genes of virgin and pregnant samples.

After observing both dot plots, it is possible to say that down regulated (more common in virgin samples) genes are mainly associated with neural processes, extracellular matrix, and cell adhesion. On the other hand, upregulated genes (more common in pregnant samples) participate in processes like lipid transport, energy metabolism and immune preparation.

This comparison may be useful to understand what functions do mammary cells prioritize in the different status. For instance, in the pregnant state the mice might be preparing for lactation and therefore there is an increased lipid metabolism and transport, whereas the neurogenesis might not be as important.

Figure 11 and *Figure 12* represent the gene concept network for the top 3 enriched GO terms obtained with *cnetplot()*. This visualization can complement the dot plot because it shows which genes are involved in the significant terms. The colored lines indicate relationships between genes and GO terms and the size of these nodes corresponds to the number of genes related to each term.

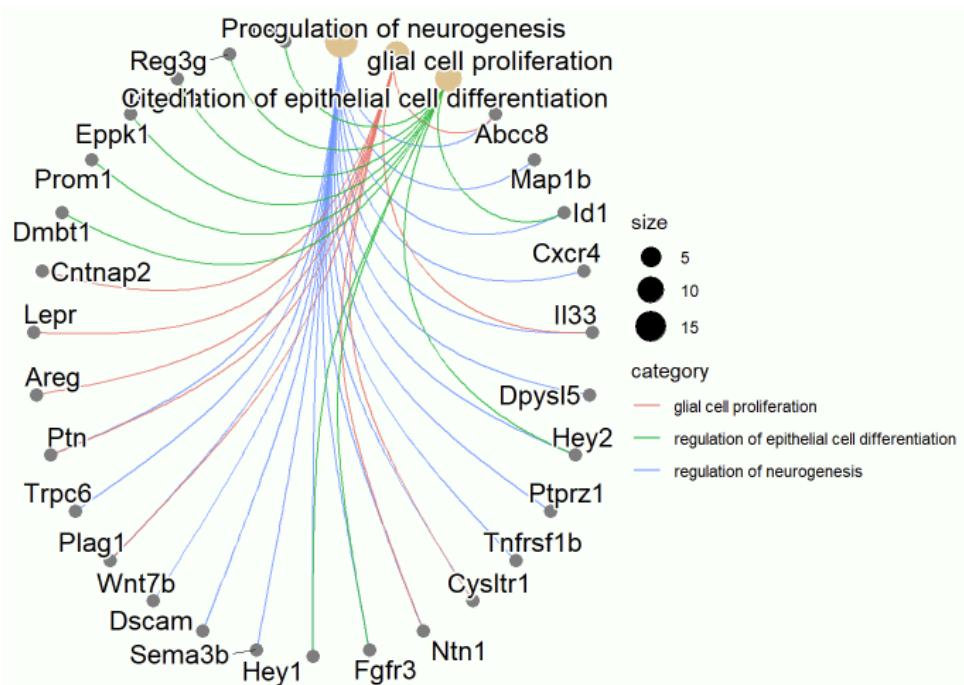


Figure 11- Gene concept network for down regulated genes of virgin and pregnant samples.

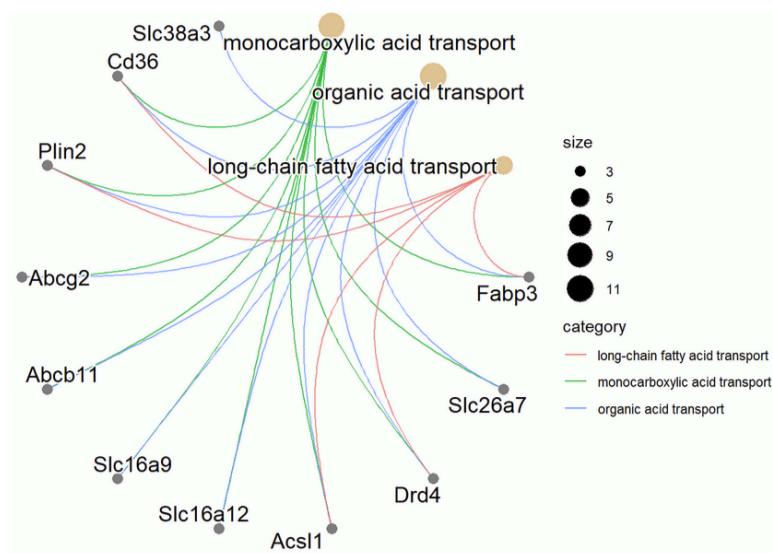


Figure 12- Gene concept network for up regulated genes of virgin and pregnant samples.

The following figures contain UpSet plots, which visualizes the overlap of genes across GO terms. The vertical bars represent the number of genes associated with specific intersections, helping us understand if genes are shared or unique among different GO terms.

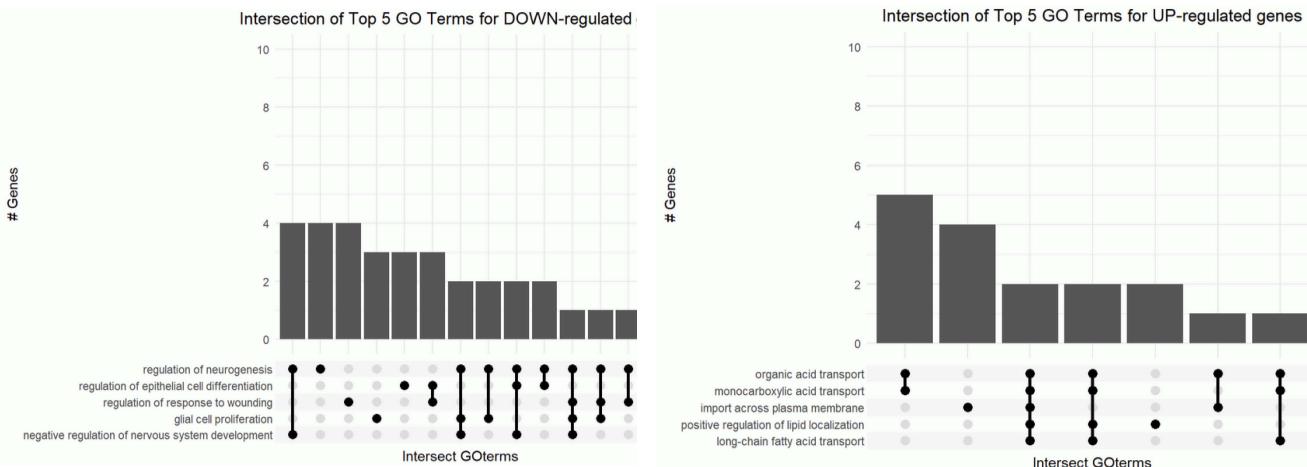


Figure 13- UpSet plot for down regulated genes (left) and up-regulated genes (right).

The last important visualization is the tree plot, which is a way to visualize hierarchically clustered enriched GO terms based on their similarity. The plot is divided into colored clusters, and each cluster groups enriched terms that are related. On the other hand, the size of the dots represents the number of genes associated with each term and the color the statistical significance.

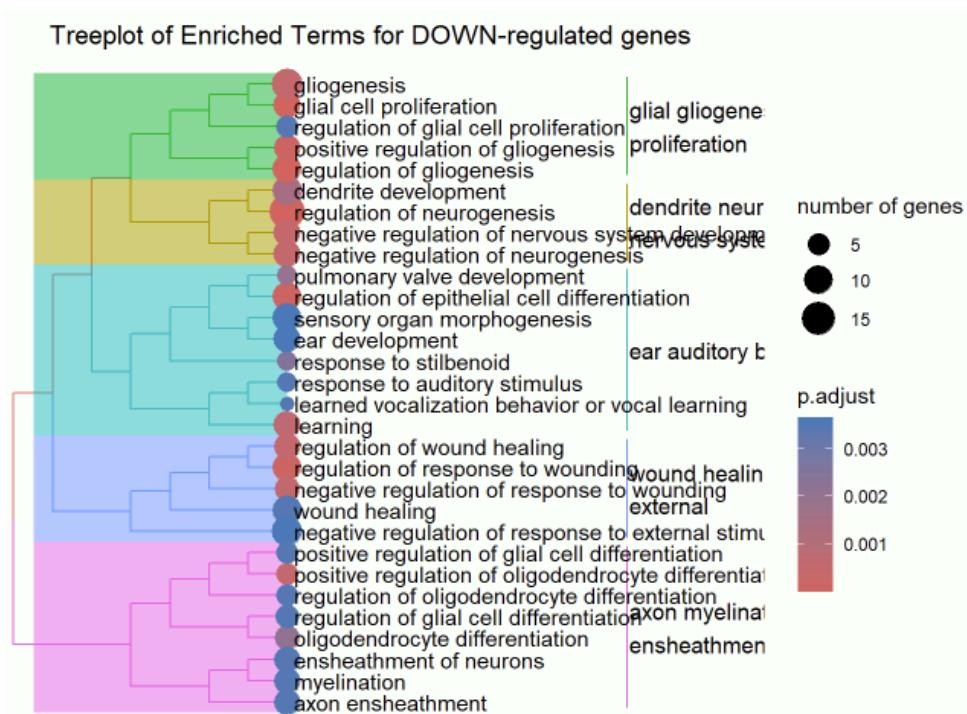


Figure 14- Tree plot for down regulated genes in virgin and pregnant samples.

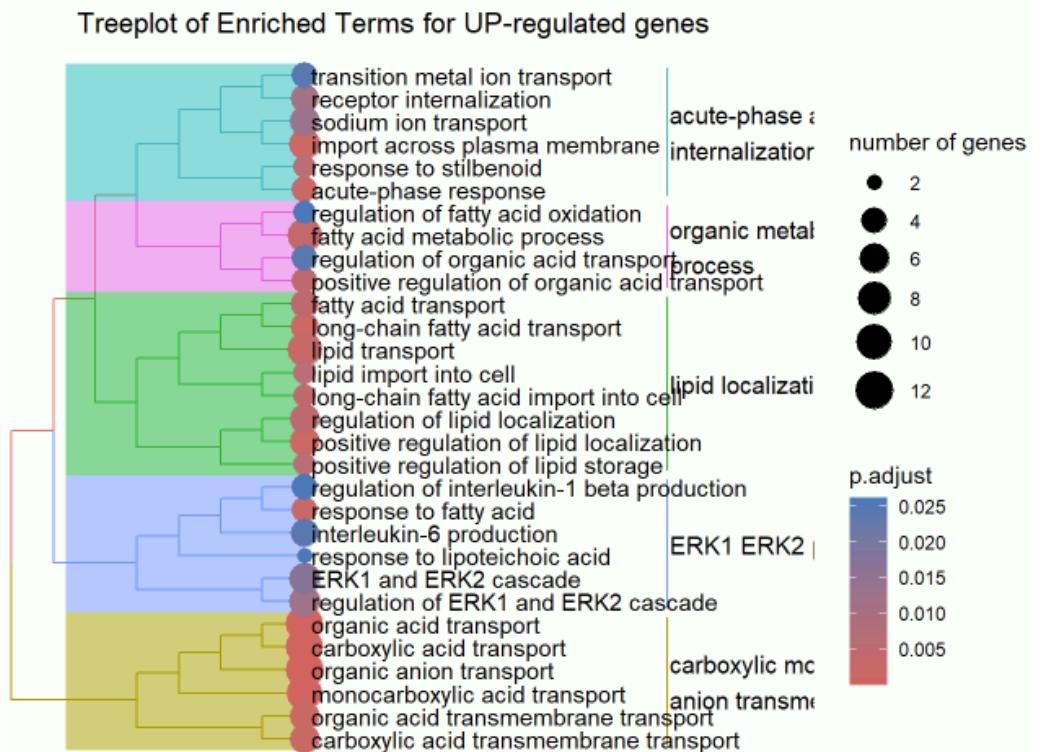


Figure 15- Tree plot for up regulated genes in virgin and pregnant samples.

EXERCISE 5

The aim of this section is to compare and comment deeper on the difference between GO terms of up and down regulated genes for virgin and pregnant samples.

The following table (*Table 1*) contains an easier comparison between the three GO terms observed in down and up regulated genes after the enrichment analysis.

The GO term BP refers to the broader biological objectives or tasks that a group of genes contribute to, whereas CC describes the location within the cell where a gene product is active. Finally, MF refers to the specific biochemical activity of a gene product, such as binding, catalysis, or transport, among others.

GO Term	Down regulated genes (more present in virgin samples)	Up regulated genes (more present in pregnant samples)
Biological process (BP)	Neurogenesis, extracellular matrix	Lipid metabolism, immune response

Cellular components (CC)	Collagen matrix, neural structures	Lipoprotein particles, plasma membrane
Molecular functions (MF)	Binding and structural integrity	Transporter activity, fatty acid binding

Table 1-Comparison table of GO terms between up and down regulated genes for virgin and pregnant samples.

The observed differences in the GO terms are aligned with the physiological changes that occur in the mammary gland during pregnancy.

It is possible to observe that down-regulated genes, which have a lower expression in pregnant samples compared to virgin samples, are more related to neurogenesis and structural functions. At this stage, the gland focuses on maintaining its ductal structures as well as a rigid extracellular matrix. These functions observed in down-regulated genes can be considered as the baseline state of the mammary gland before it undergoes a remodeling and functional differentiation during pregnancy.

On the other hand, during pregnancy, the focus is shifted to milk production, immune protection, and nutrient transport. Lipids are a major component of milk, so it makes sense that genes involved in lipid metabolism are upregulated in order to prepare the mammary gland for the production of milk. This metabolic change ensures that essential milk components such as triglycerides and fatty acids are synthesized, and the nutrient transport delivers the necessary components for its synthesis. Moreover, the mammary gland recruits immune cells during pregnancy to prepare for lactation and protect the gland from pathogens.

EXERCISE 6

In this section, a differential expression analysis, as well as an enrichment analysis is performed but only taking into account samples from virgin and lactating mice.

All the steps followed in order to complete this exercise are exactly the same as the ones explained in Exercises 3 and 4, with the difference that in this case the metadata and counts matrix is filtered to contain only samples corresponding to virgin and lactating mice.

The reference level is also set to the virgin samples, since that is our control or baseline group, and we want to study genes that are up or down regulated in lactating mice compared to virgin mice.

The diagnostic plots for this exercise can be found in the Rcode and show again that the data is meaningful.

Figure 16 represents the volcano plot for these samples. The thresholds to build the plots remained the same as for virgin and pregnant samples.

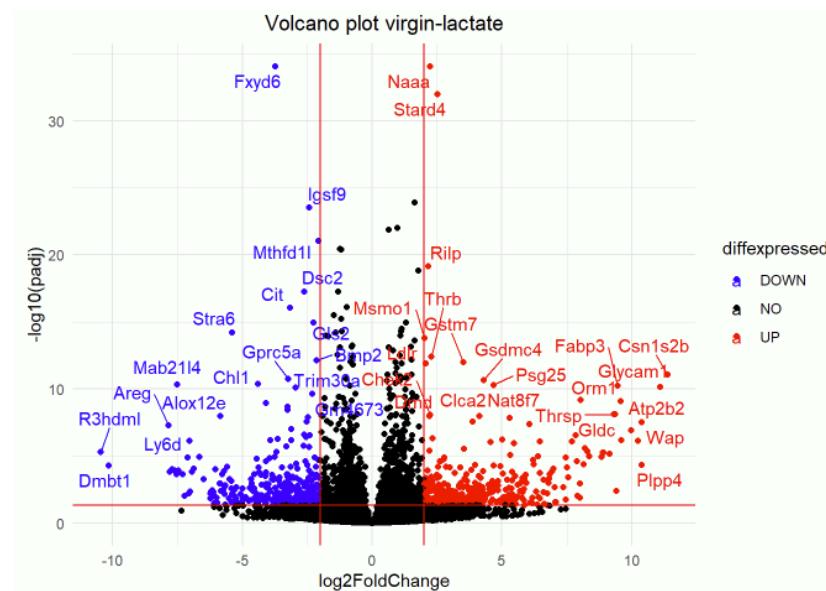


Figure 16- Volcano plot for the virgin and lactate samples. Genes contain their respective name and are color-coded if they are differentially expressed.

A dataframe containing the results of up and down regulated genes is saved for the enrichment analysis.

The enrichment analysis was also performed with enrichGO() function and using the mouse genome from a reference database.

Figure 17 and Figure 18 contain the dot plots of down and up regulated genes in this case.

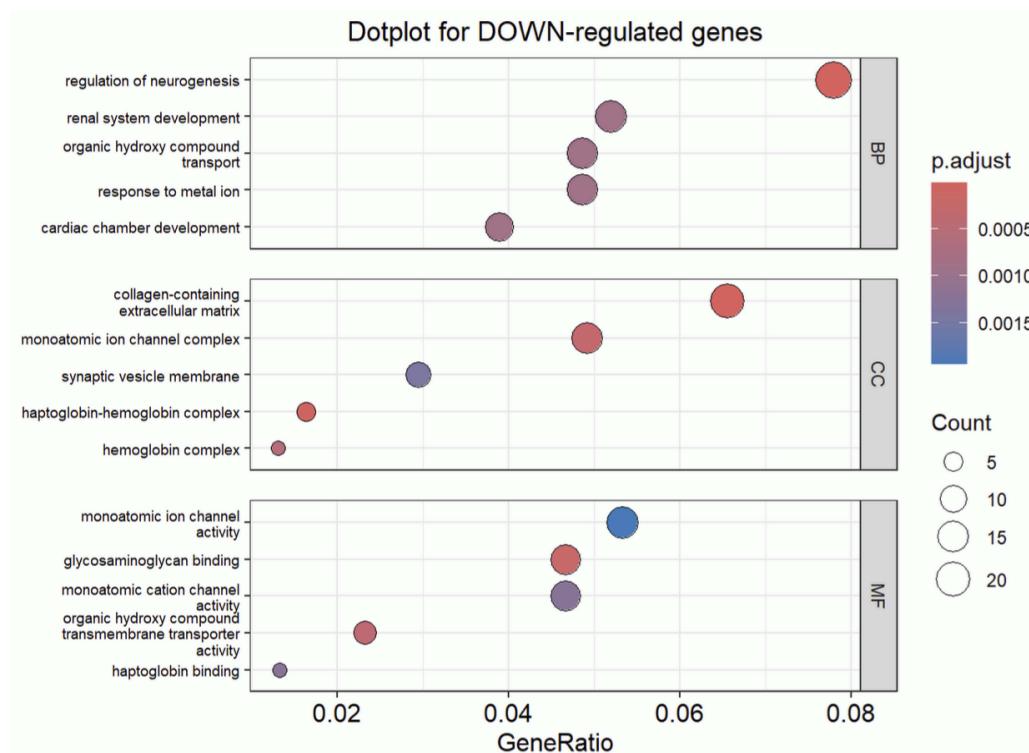


Figure 17- Dot plot for down regulated genes of virgin and lactating samples.

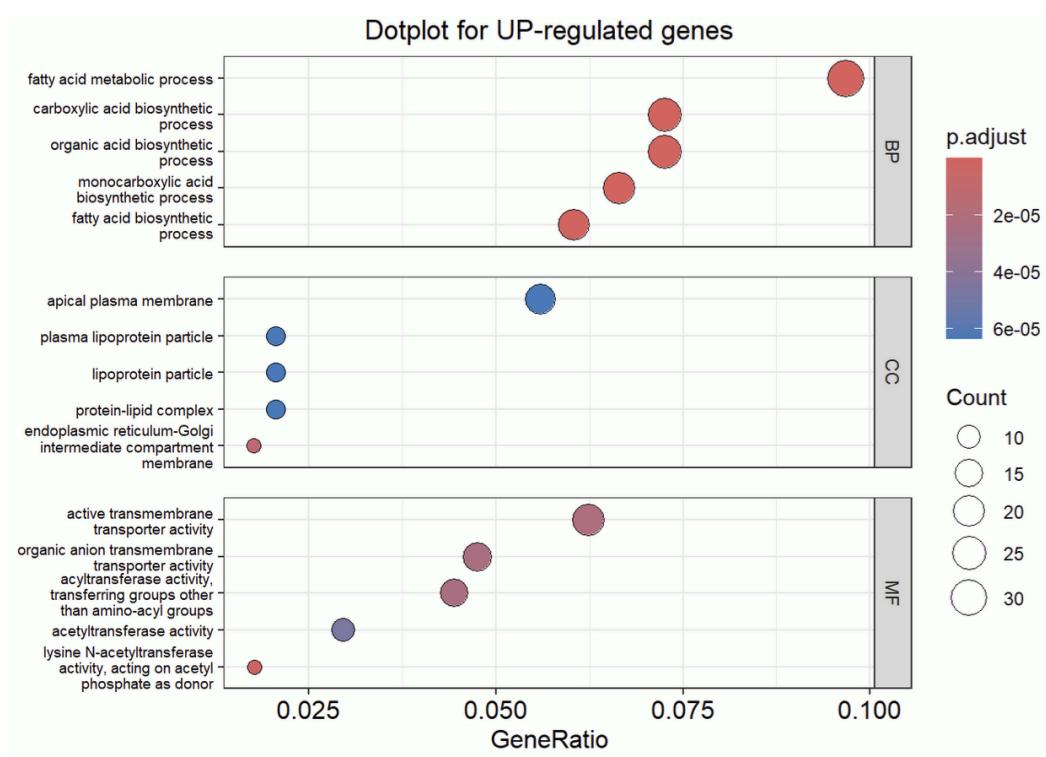


Figure 18- Dot plot for up regulated genes of virgin and lactating samples.

The dot plots give us an overview of the top 5 GO terms for each category in both down and regulated genes and is a way to illustrate the results of the enrichment analysis.

In this context, down-regulated genes are those that have lower expression in lactating samples compared to virgin samples, thus reflecting the functions that are less critical during lactation and more characteristic of a virgin state. Up-regulated genes, on the other hand, would be those that have a higher expression in lactating samples compared to virgin samples.

From *Figure 17*, it is possible to see that, during the virgin state, the mammary gland has higher expression of genes associated with neurogenesis, development, structure and sensory functions. Virgin mammary glands also express genes involved in ion transport and structural binding, which maintains homeostasis and ductal integrity. Nevertheless, these functions are less emphasized during the lactation state, where metabolic processes dominate. The main cellular component observed is the apical plasma membrane, which is the site of milk secretion in alveolar cells. Moreover, lipoproteins deliver cholesterol and lipids to the mammary gland, therefore providing the “raw materials” for milk fat production. At a molecular function level, the upregulation of transporter genes is essential for the import of nutrients.

In contrast, from *Figure 18* it is observed that fatty-acid metabolic processes dominate in up-regulated genes, instead of neurogenesis processes. It was already expected that the lactating mammary gland would be highly active in synthesizing and secreting milk fats. The upregulation of these pathways is essential for producing energy-rich triglycerides and lipids needed for milk.

Figure 19 and Figure 20 contain the three main GO terms for down and up regulated genes, highlighting the name of each gene and the different relations for a deeper inspection.

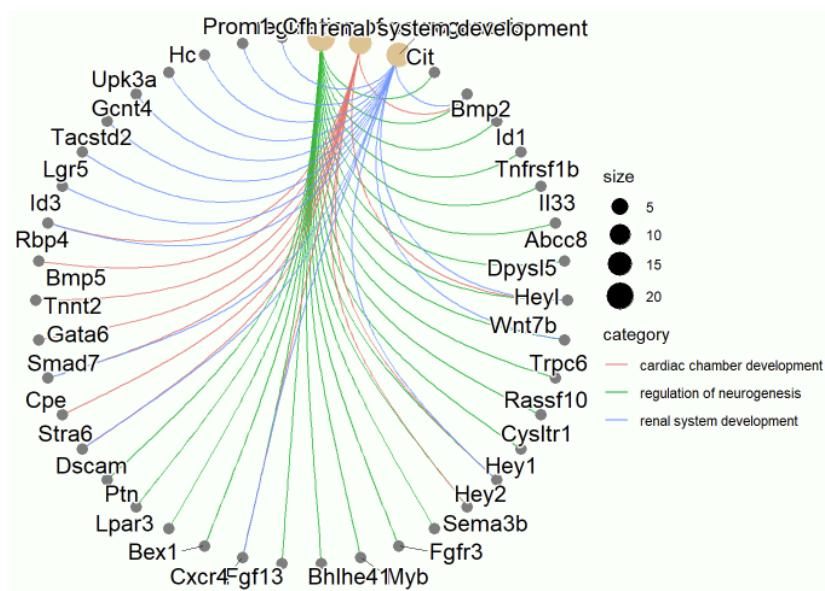


Figure 19.- Gene concept network for down regulated genes of virgin and lactating samples.

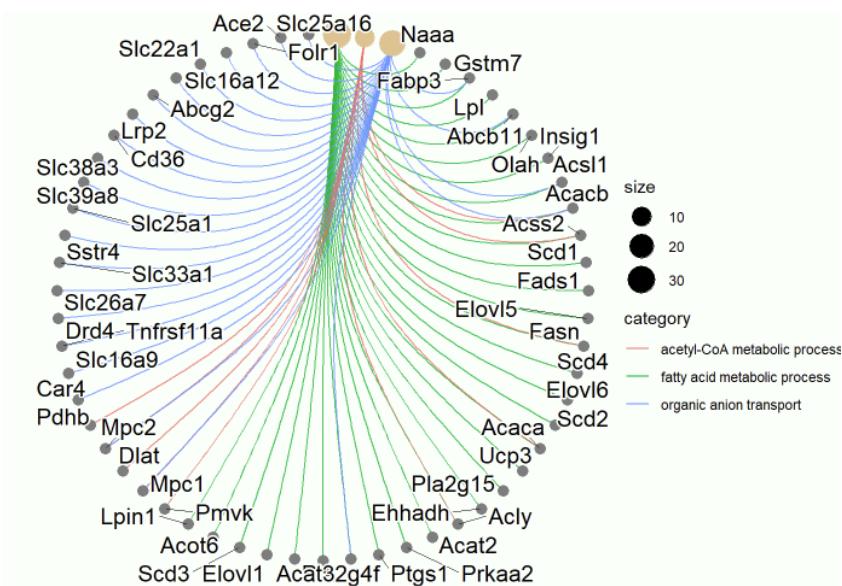


Figure 20.- Gene concept network for upregulated genes of virgin and lactating samples.

Figure 21 contains the UpSet plots for virgin and lactating samples:

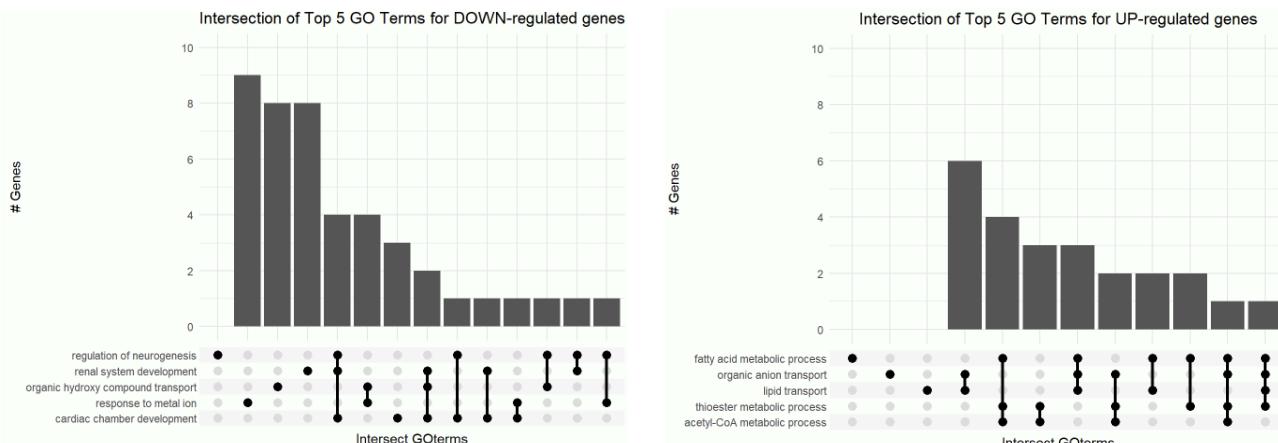


Figure 21- UpSet plot for down regulated genes (left) and up-regulated genes (right).

Figure 22 and Figure 23 contain the respective tree plots, highlighting the GO terms that are similar in each case.

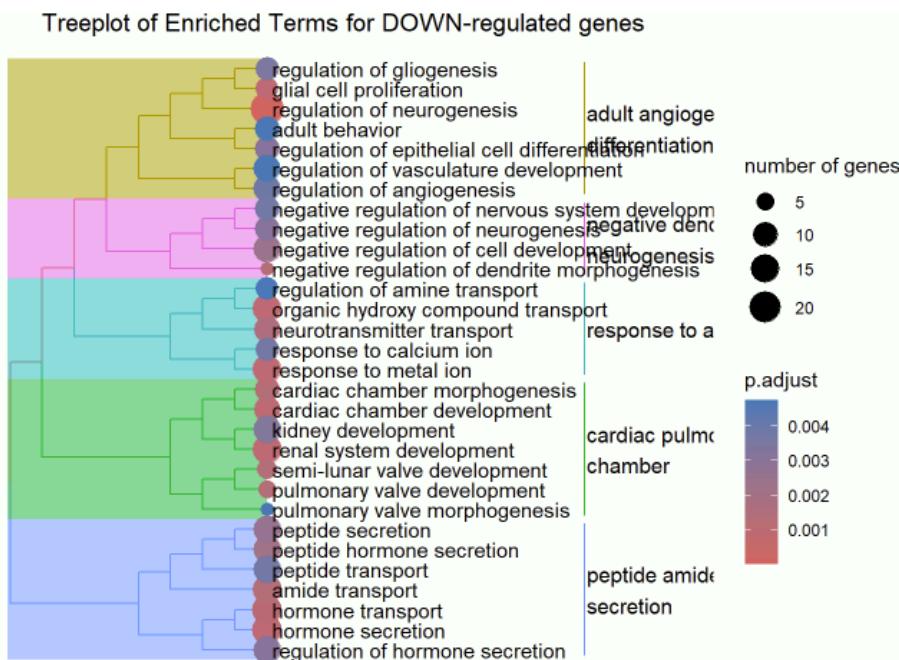


Figure 22- Tree plot for down regulated genes in virgin and lactating samples.

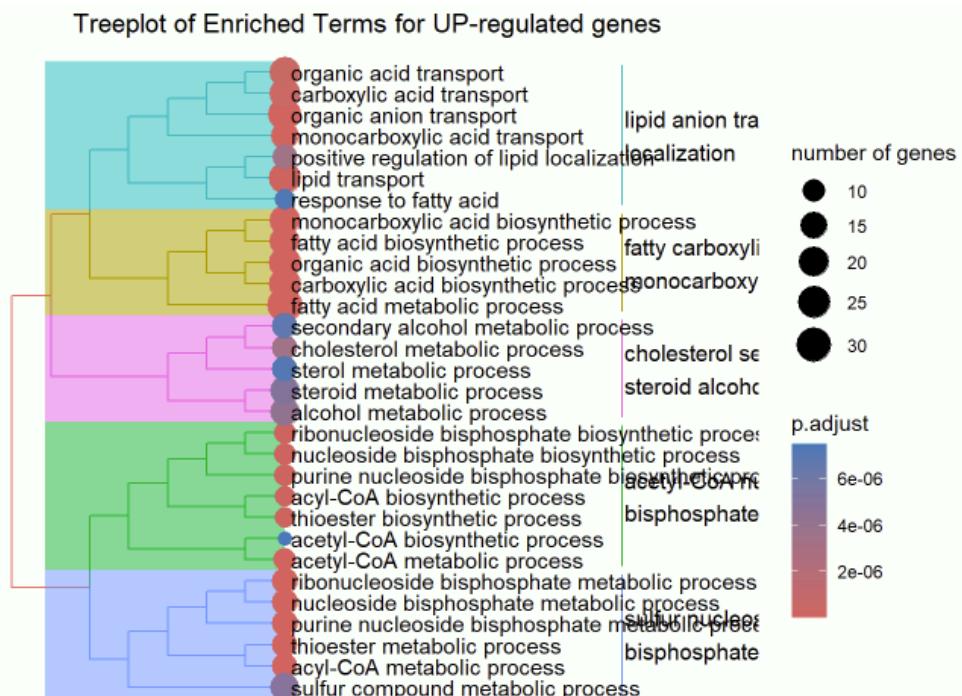


Figure 23- Tree plot for upregulated genes in virgin and lactating samples.

EXERCISE 7

In this section, a differential expression analysis, as well as an enrichment analysis is performed but only taking into account samples from pregnant and lactating mice.

All the steps followed in order to complete this exercise are exactly the same as the ones explained in Exercises 3 and 4, with the difference that in this case the metadata and counts matrix is filtered to contain only samples corresponding to pregnant and lactating mice.

Another difference is that the reference level is set to the pregnant samples, since in this case the goal is to understand how gene expression changes during the transition from pregnancy to lactation. By setting pregnancy as the reference, it will be possible to identify genes that are upregulated or downregulated in lactation relative to pregnancy.

The diagnostic plots for this exercise can be found in the Rcode and show again that the data is meaningful.

Figure 24 represents the volcano plot for these samples. The thresholds to build the plots remained the same as for virgin and pregnant samples.

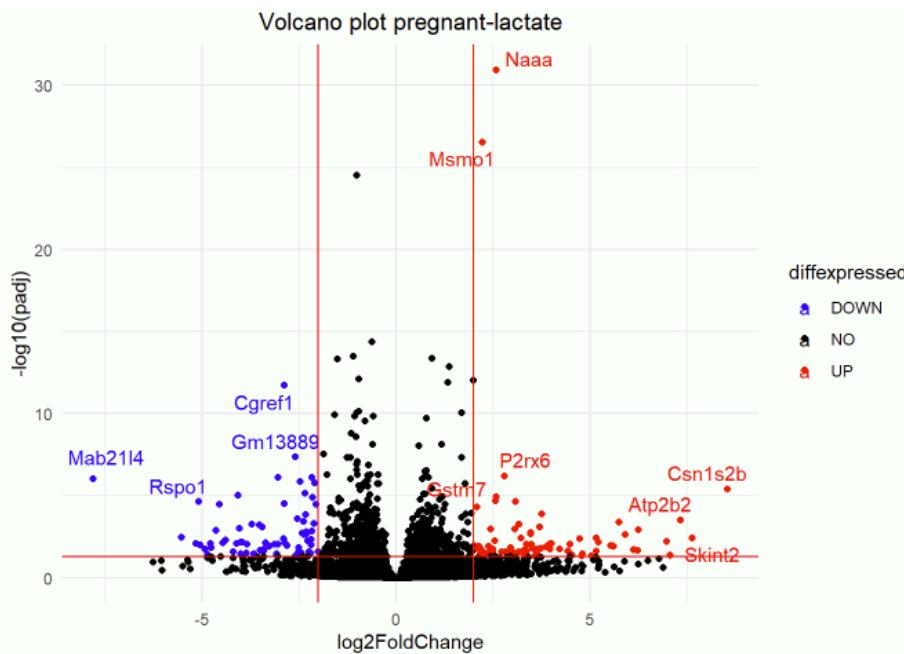


Figure 24- Volcano plot for the pregnant and lactate samples. Genes contain their respective name and are color-coded if they are differentially expressed.

The enrichment analysis was also performed with `enrichGO()` function and using the mouse genome from a reference database.

Figure 25 and *Figure 26* contain the dot plots of down and up regulated genes in this case. In this case, upregulating genes are those with higher expression in lactating samples compared to pregnant ones, and downregulated genes the opposite.

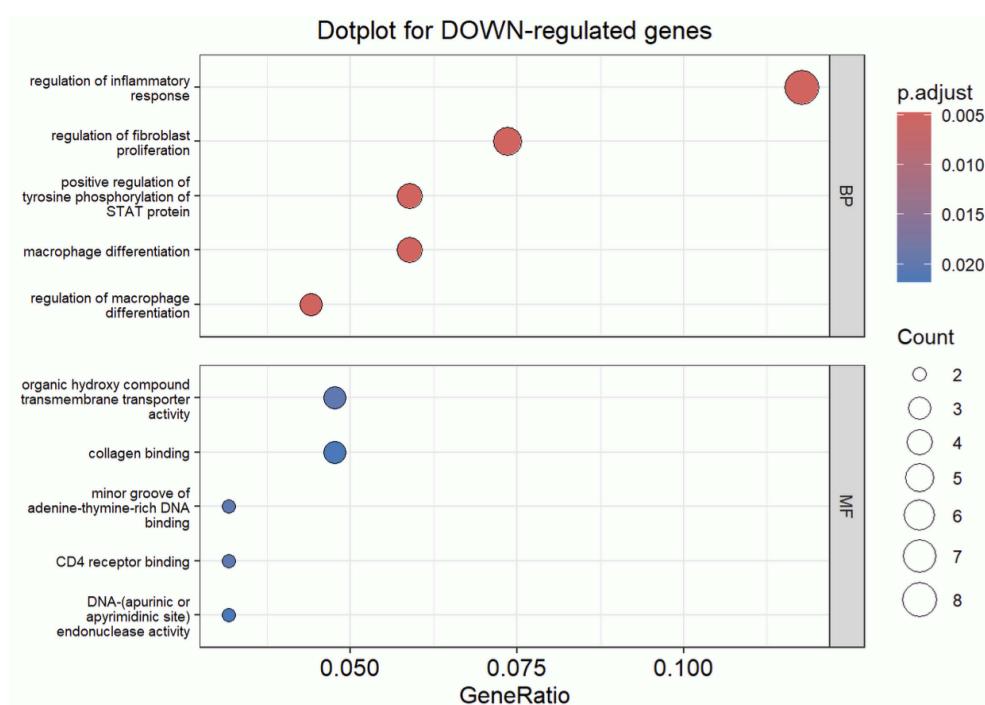


Figure 25- Dot plot for down regulated genes of pregnant and lactating samples.

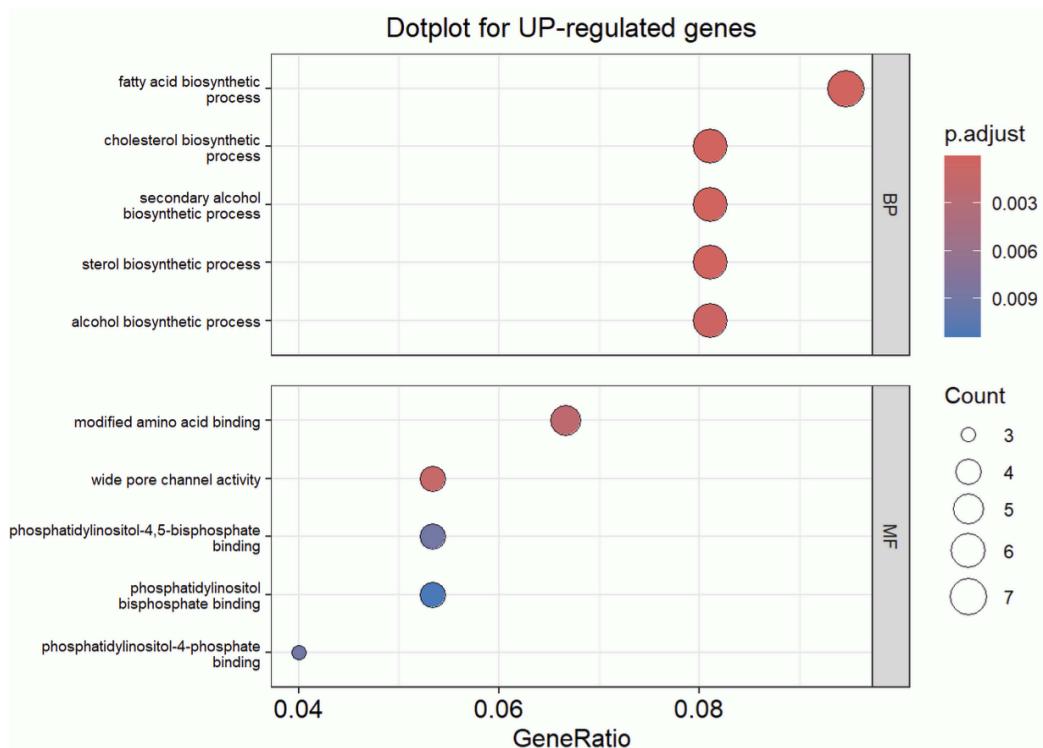


Figure 26- Dot plot for up regulated genes of pregnant and lactating samples.

From *Figure 25*, it is possible to state that, during pregnancy, immune regulation is crucial to support tissue remodeling, angiogenesis, and the protection of the fetus and mammary gland. The main focus for the mammary gland in this stage is to ensure structural preparation and immune regulation.

However, these processes are less important during lactation (as seen in *Figure 26*), where the focus shifts to milk secretion. During lactation it is observed an upregulation of lipid biosynthesis, transporter activity, and signaling pathways, which facilitates the production and secretion of milk components as it has been seen before.

The cellular components GO terms have not achieved statistical significance for enrichment and therefore are not present in the dot plots.

Figure 27 and *Figure 28* contain the three main GO terms for down and up regulated genes. It is possible to see a smaller quantity of genes compared to the rest of the analysis.

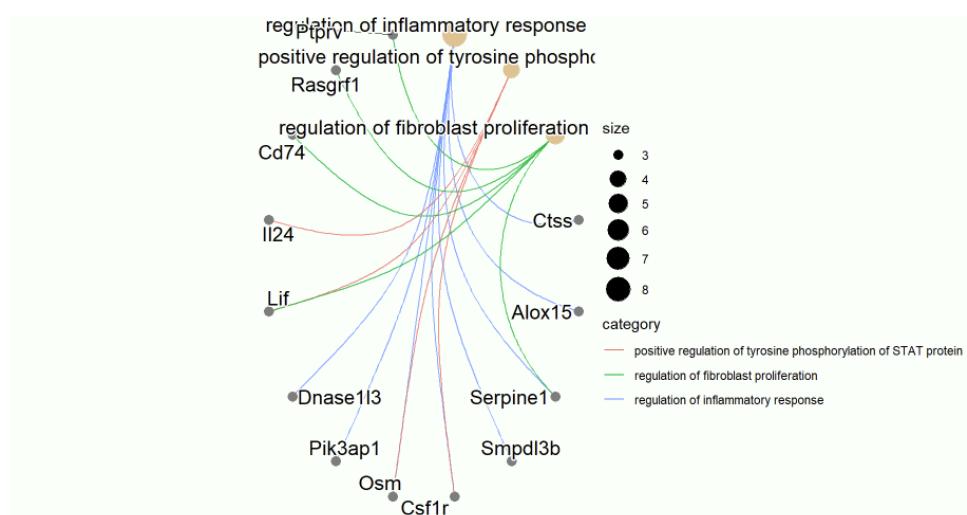


Figure 27.- Gene concept network for down regulated genes of pregnant and lactating samples.

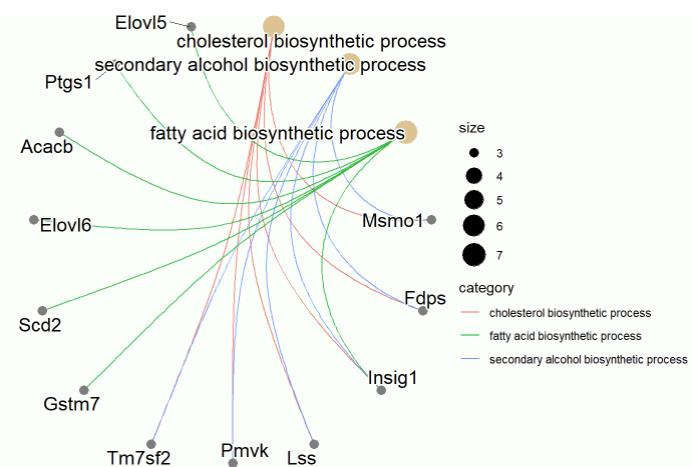


Figure 28.- Gene concept network for upregulated genes of pregnant and lactating samples.

Figure 29 contains the UpSet plots for pregnant and lactating samples:

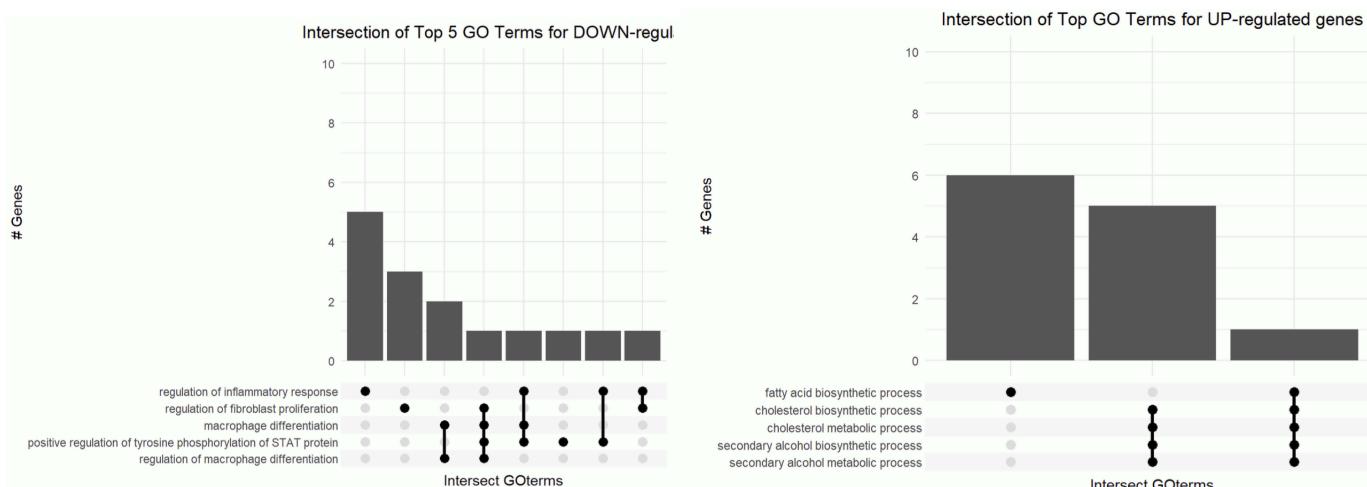


Figure 29- UpSet plot for down regulated genes (left) and up-regulated genes (right).

Finally, *Figure 30* and *Figure 31* contain the respective tree plots, highlighting the GO terms that are similar in each case.

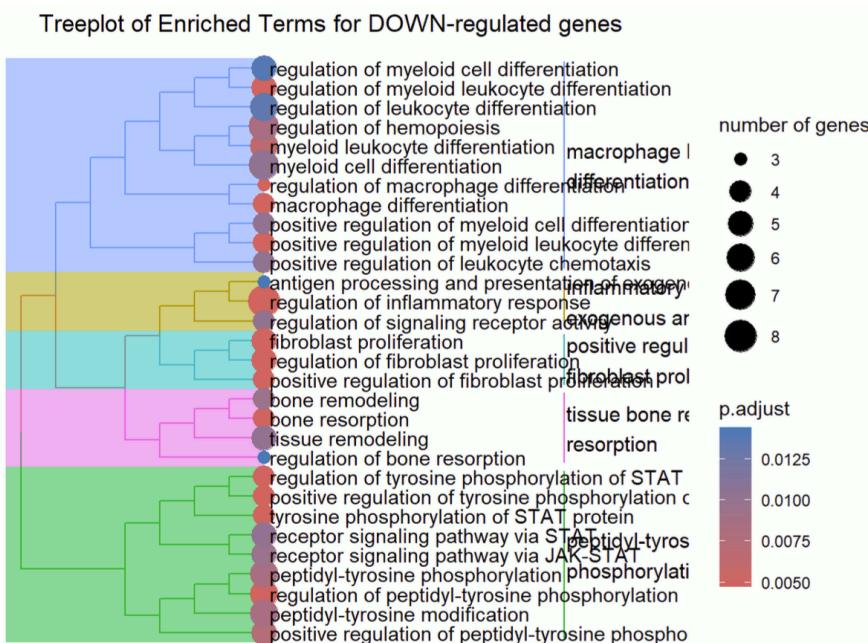


Figure 30- Tree plot for down regulated genes in pregnant and lactating samples.

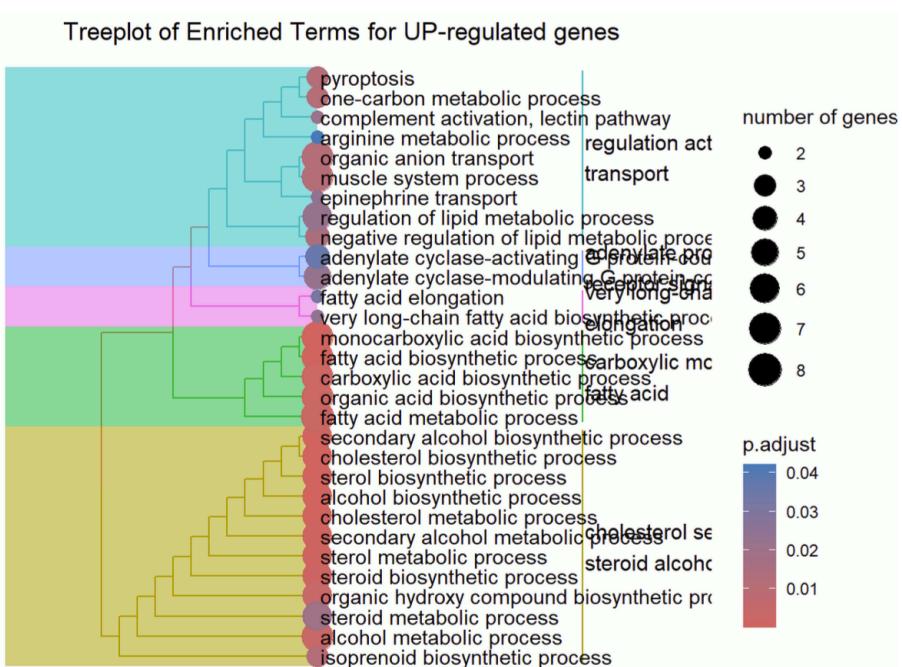


Figure 31- Tree plot for up regulated genes in pregnant and lactating samples.

EXERCISE 8

The goal of this last exercise is to compare and comment on the GO terms obtained across the whole analysis.

Table 2 contains a summary table of the comparisons between the GO terms of the three groups (virgin vs pregnant, virgin vs lactating and pregnant vs lactating).

Comparison group	GO category		
	Biological processes	Molecular function	Cellular component
Virgin vs pregnant	<ul style="list-style-type: none"> - Up in Pregnant: Tissue morphogenesis, cell proliferation, lipid metabolism - Down in Pregnant: Neurogenesis, structural functions 	<ul style="list-style-type: none"> - Up in Pregnant: Transporter activity, fatty acid binding - Down in Pregnant: Binding, structural integrity 	<ul style="list-style-type: none"> - Up in Pregnant: Lipoprotein particles, plasma membrane - Down in Pregnant: Neural structures, collagen matrix
Virgin vs lactating	<ul style="list-style-type: none"> - Up in Lactating: Fatty acid metabolic processes, lipid biosynthesis - Down in Lactating: Neurogenesis, organ and system development and regulation 	<ul style="list-style-type: none"> - Up in Lactating: Transporter activity, transferase enzyme activity - Down in Lactating: Structural binding, ion channel activity 	<ul style="list-style-type: none"> - Up in Lactating: Apical plasma membrane, lipoprotein particles - Down in Lactating: Collagen matrix, structural complexes
Pregnant vs lactating	<ul style="list-style-type: none"> - Up in Lactating: Fatty acid biosynthesis, milk secretion pathways - Down in Lactating: Inflammatory response, macrophage activity, fibroblast proliferation 	<ul style="list-style-type: none"> - Up in Lactating: Wide pore channel activity, lipid and amino acid binding - Down in Lactating: Collagen binding, CD4 receptor binding 	No statistically significant data

Table 2.- Final comparison table of GO terms across the three comparison groups. GO terms are summarized and divided into up or down regulated genes.

Through the differences observed in *Table 2*, it has been proven that the mammary gland goes through significant morphological changes during puberty, pregnancy and lactation. Each stage is characterized by distinct processes such as ductal elongation (puberty), alveolar formation and differentiation (pregnancy), and secretory activation (lactation) [1].

In virgin samples, the mammary gland is in a baseline or developmental state, and its main focus is to maintain a structural integrity and support the neurogenesis and extracellular matrix organization. These processes are critical for ductal elongation, which prepares the gland for future differentiation and adaptation [2]. The GO terms that dominate in this stage are those related to collagen matrix, neural structures, and binding/structural integrity.

In pregnant samples, on the other hand, the mammary experiments a significant remodeling and differentiation in order to prepare for lactation. GO terms enriched in pregnant samples reflect processes that include cell proliferation, alveolar morphogenesis, and extracellular matrix (ECM) remodeling, which are essential for expanding the glandular tissue and creating milk-producing alveoli. Moreover, this remodeling process supports the generation of new blood vessels, as well as infiltration of immune and inflammatory cells and fibroblast reorganization [3]. In this stage, it is also possible to see that the lipid metabolism is rising to prepare for milk demand.

The main function of lactation is to synthesize and release milk, so the mammary glands need to change to meet the demand for milk production. During lactation, the mammary gland is the most active site for both lipid synthesis and fatty acid esterification [7]. The lactating mammary gland operates at a high metabolic rate, since it requires an efficient transport and synthesis of milk components. This is shown in the upregulation of lipid metabolism, transporter activity, and biosynthetic processes, which is supported by literature [7]. Conversely, processes like inflammatory response and macrophage activity, which were common during pregnancy for tissue remodeling, are downregulated in this stage.

The commented results show the difference in gene expression for the mammary gland in each stage, highlighting its ability to shift focus depending on the current necessity.

REFERENCES

1. Fu NY, Rios AC, Pal B, Soetanto R, Lun ATL, Liu K, et al. EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival. *Nat Cell Biol.* 2015 Apr;17(4):365-75. doi: 10.1038/ncb3117.
2. Fata JE, Werb Z, Bissell MJ. Regulation of mammary gland branching morphogenesis by the extracellular matrix and its remodeling enzymes. *Breast Cancer Res.* 2003 Aug 19;6(1):1-11. doi: 10.1186/bcr634.
3. Biswas SK, Banerjee S, Baker GW, Kuo CY, Chowdhury I. The mammary gland: basic structure and molecular signaling during development. *Int J Mol Sci.* 2022 Mar 31;23(7):3883. doi: 10.3390/ijms23073883.
4. Duttaroy AK, Basak S. Maternal fatty acid metabolism in pregnancy and its consequences in the feto-placental development. *Front Physiol.* 2022 Jan 20;12:787848. doi: 10.3389/fphys.2021.787848.
5. Plaks V, Boldajipour B, Linnemann JR, Nguyen NH, Kersten K, Wolf Y, et al. Adaptive immune regulation of mammary postnatal organogenesis. *Dev Cell.* 2015 Sep 14;34(5):493-504. doi: 10.1016/j.devcel.2015.07.015.
6. Rudolph MC, McManaman JL, Phang TL, Russell T, Kominsky DJ, Serkova NJ, et al. Metabolic regulation in the lactating mammary gland: a lipid synthesizing machine. *Physiol Genomics.* 2007 Feb 12;28(3):323-36. doi: 10.1152/physiolgenomics.00020.2006.
7. Canul-Medina G, Fernandez-Mejia C. Morphological, hormonal, and molecular changes in different maternal tissues during lactation and post-lactation. *J Physiol Sci.* 2019;69:825–835. doi: 10.1007/s12576-019-00687-y.
8. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. *Nat Genet.* 2000 May;25(1):25-9. doi: 10.1038/75556.
<https://geneontology.org/docs/go-enrichment-analysis/>