Gene Expression Analysis of Hepatocellular Carcinoma using GEOQuery Dataset

Bioinformatics Course

Master's Degree in Engineering in Computer Science

Carmen Trovato – 2033862

Academic Year: 2024/2025

SAPIENZA
UNIVERSITÀ DI ROMA

Presentation Outline

Introduction

In this presentation, we will explore the identification of differentially expressed genes between case and control samples, with a focus on functional characterization. Specifically, we compare gene expression profiles of healthy individuals with those of patients affected by hepatocellular carcinoma (**HCC**).

Hepatocellular Carcinoma

Hepatocellular carcinoma (HCC), or hepatocellular carcinoma, is a malignant tumor that arises from liver cells, called hepatocytes.

It is the most common type of primary liver cancer and often develops in the presence of preexisting liver damage, such as cirrhosis.

Dataset Overview

For this analysis, the gene expression profiles from the **GSE22058** dataset were downloaded from the GEO (Gene Expression Omnibus) database using the **R programming language** and the **GEOquery** package.

The dataset includes microarray-based gene expression data from patients with **hepatocellular carcinoma (HCC)** and healthy controls.

It consists of a total of **96 samples**, including **tumor tissues (HCC)** and **adjacent non-tumor liver tissues** from the same individuals. Initially, there were **110 tumor samples** and **96 normal samples**, but only one sample per individual was retained to ensure unique pairing.

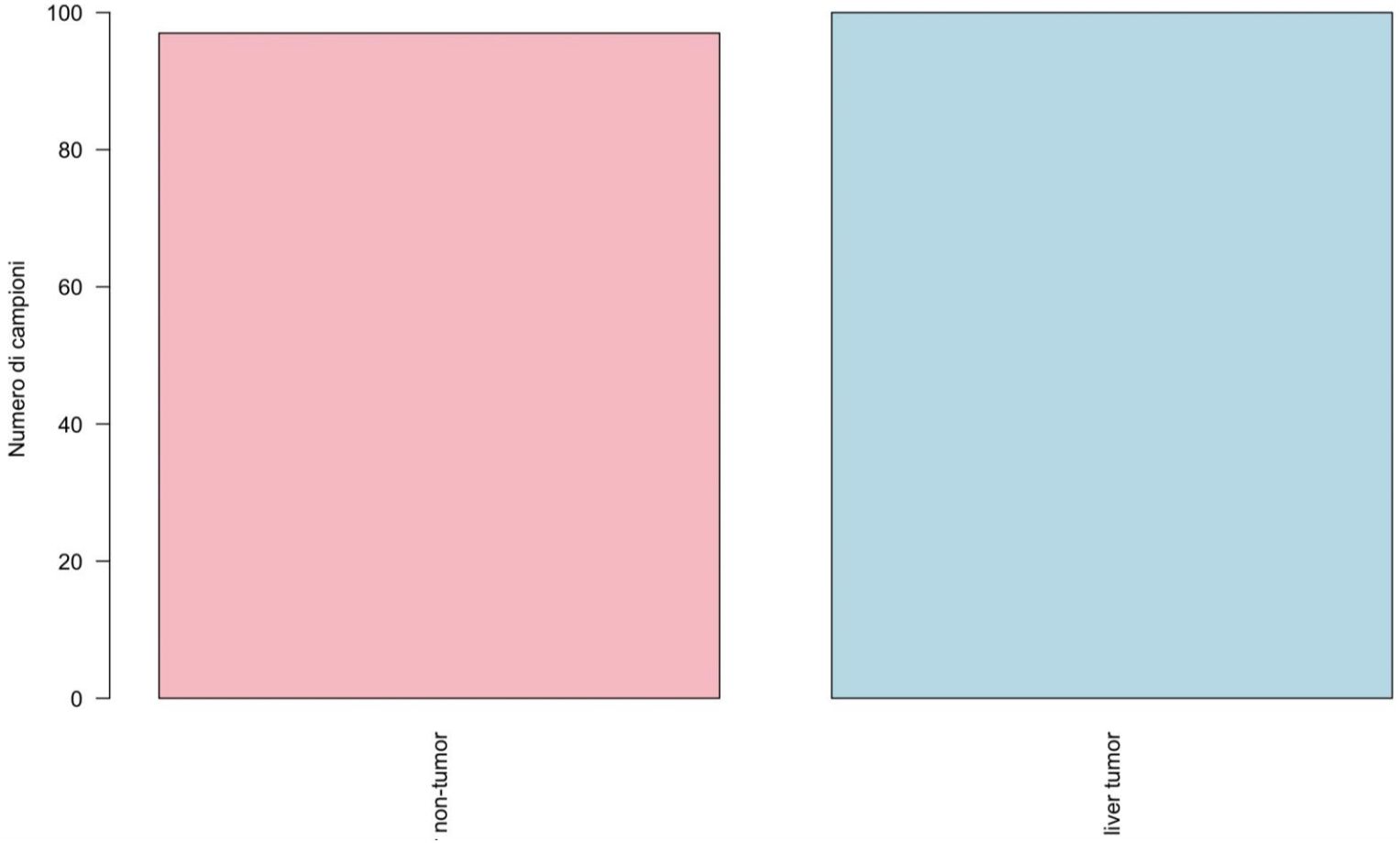
Dataset Overview

geo_accession	individual:ch1	tissue:ch1
GSM548340	21	adjacent liver non-tumor
GSM548341	40	adjacent liver non-tumor
GSM548342	73	liver tumor
GSM548343	73	adjacent liver non-tumor
GSM548344	74	liver tumor
GSM548345	76	liver tumor
GSM548346	76	adjacent liver non-tumor
GSM548347	77	liver tumor
GSM548348	77	adjacent liver non-tumor
GSM548349	78	liver tumor
GSM548350	78	adjacent liver non-tumor
GSM548351	79	liver tumor
GSM548352	79	adjacent liver non-tumor
GSM548353	80	liver tumor
GSM548354	80	adjacent liver non-tumor
GSM548355	82	liver tumor
GSM548356	82	adjacent liver non-tumor
GSM548357	84	liver tumor

- **individual:ch1->** Patient ID
  tissue:ch1 -> Tissue type

Some results from metadata.txt

Dataset Overview



Sample distribution by tissue type

Data Preprocessing

Pre-processing

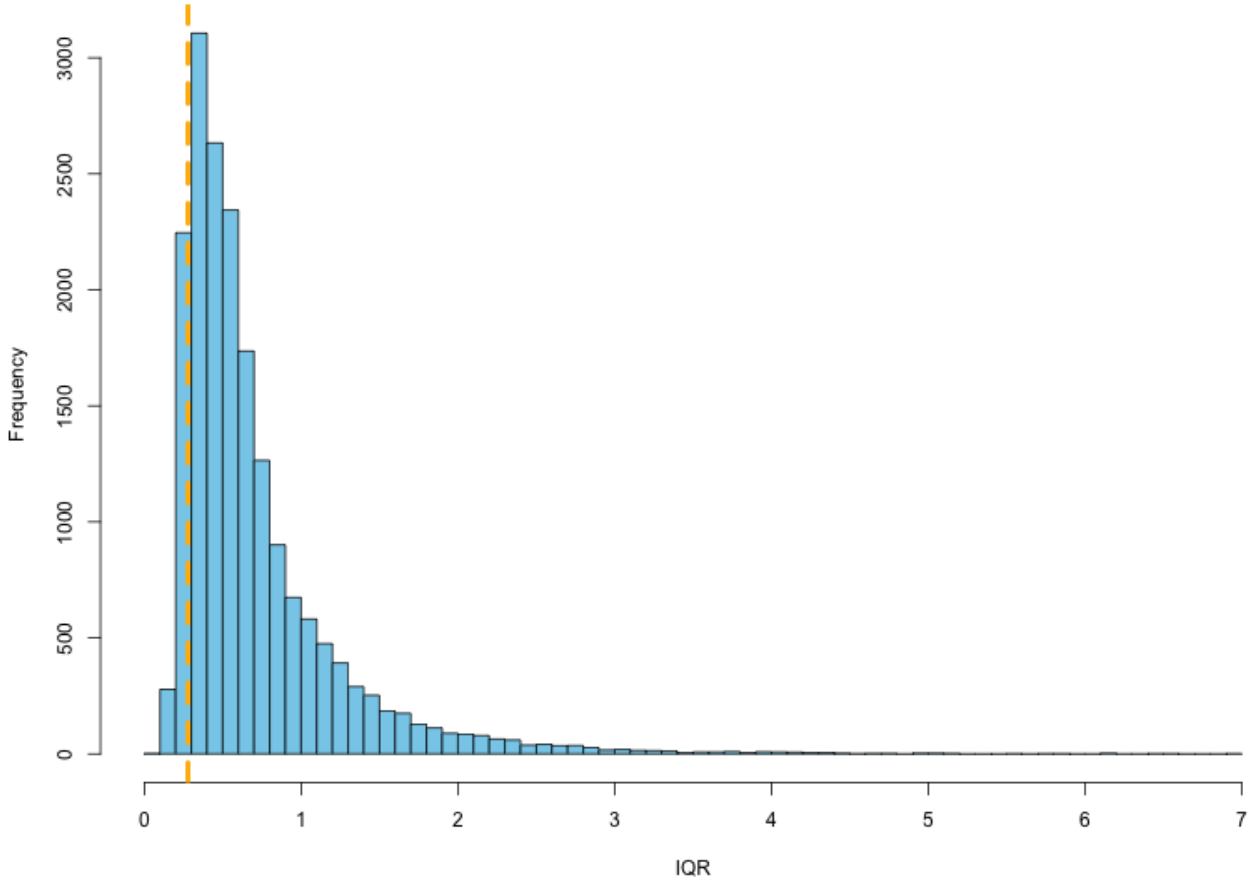
Before performing differential expression analysis, raw expression data must be preprocessed to improve robustness and reduce noise. The main steps include log transformation, variability filtering, and removal of uninformative genes.

- Log Transformation: Stabilize variance across expression levels
- Removal of Non-Expressed Genes: genes with **mean expression = 0** in both groups (Tumor and Normal) were removed
- IQR Filtering (Interquartile Range): measures how variable each gene is across all samples

Data Preprocessing

Pre-processing

The IQR histogram shows that only a small fraction of the genes have enough variability to be informative. By applying a 10th percentile filter, we selected the most variable genes, improving the sensitivity of the differential analysis between Tumor and Normal samples. This step is essential to achieve the goal of the analysis: identifying genes actually involved in HCC.



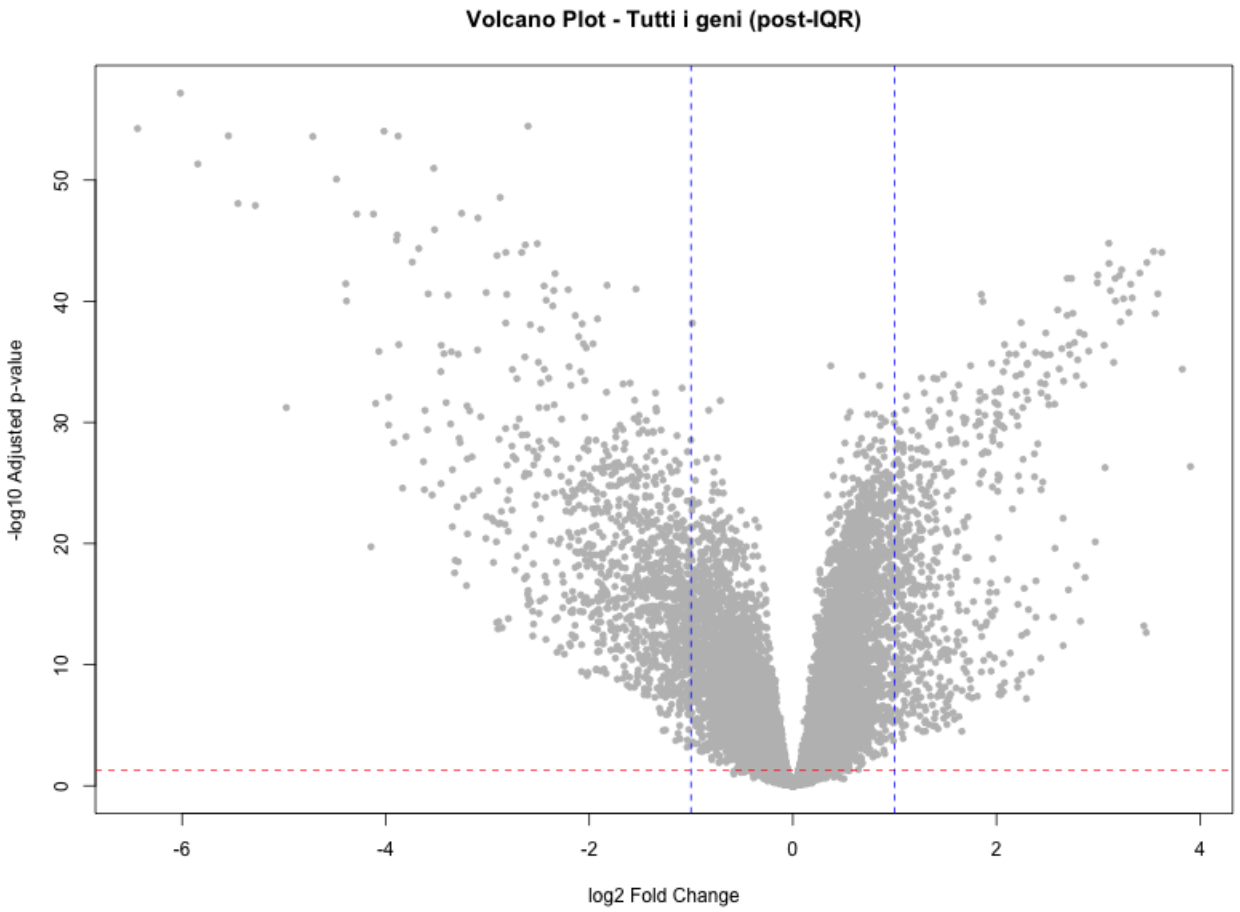
IQR frequency distribution

Filtering

The volcano plot summarizes the results of differential expression analysis by plotting the log2 fold change (x-axis) against the adjusted p-value (y-axis).
Genes with a log2FC > 1 or < -1 and an adjusted p-value < 0.05 are considered significantly differentially expressed and are highlighted in orange (upregulated) and green (downregulated), respectively.
This visualization confirms that several genes show statistically significant changes in expression between hepatocellular carcinoma patients and healthy controls, supporting the biological relevance of the analysis. Following it showed the graphic of volcano plot.

Filtering

Vulcano-Plot before filtering



Filtering

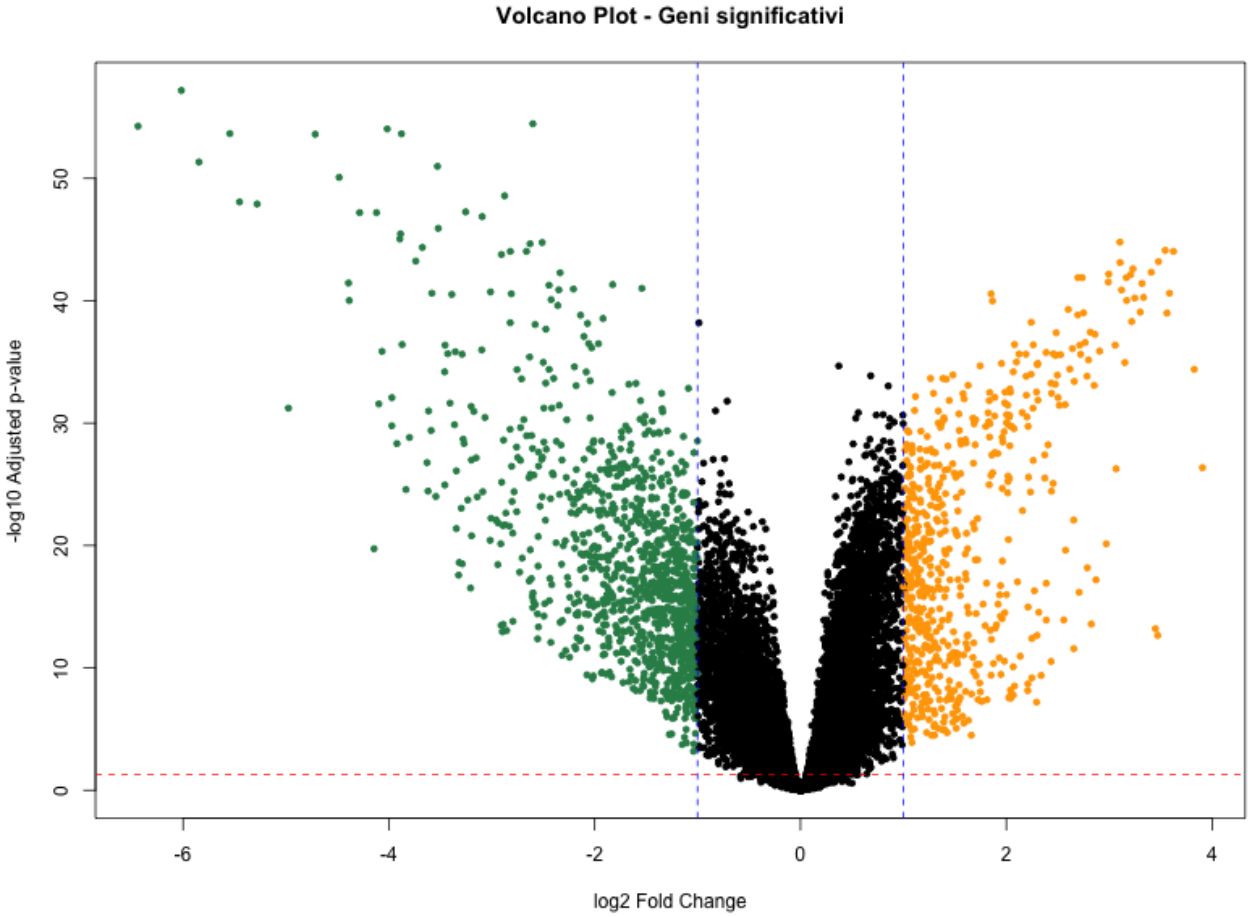
Selection of significant genes: logFC and FDR

Two selection criteria were applied to identify differentially expressed genes (DEGs):
|log2 Fold Change| > 1 → significant expression change
FDR < 0.05 → multiple error control (corrected tests)
After filtering, the following remain: **1668 significant genes**

These genes represent the most reliable candidates for functional analysis.

Filtering

Vulcano-plot post filtering



Upregulated | 603 genes

Downregulated | 1065

Non-significant genes | 14984
(low difference + high p-value)

Filtering

Resume about Vulcano-plot

In this step we applied a combined filter on log fold change and FDR, selecting only genes with strong variation and statistical significance. This step is crucial to avoid false positives and focus on the most relevant genes for the difference Tumor vs Normal.

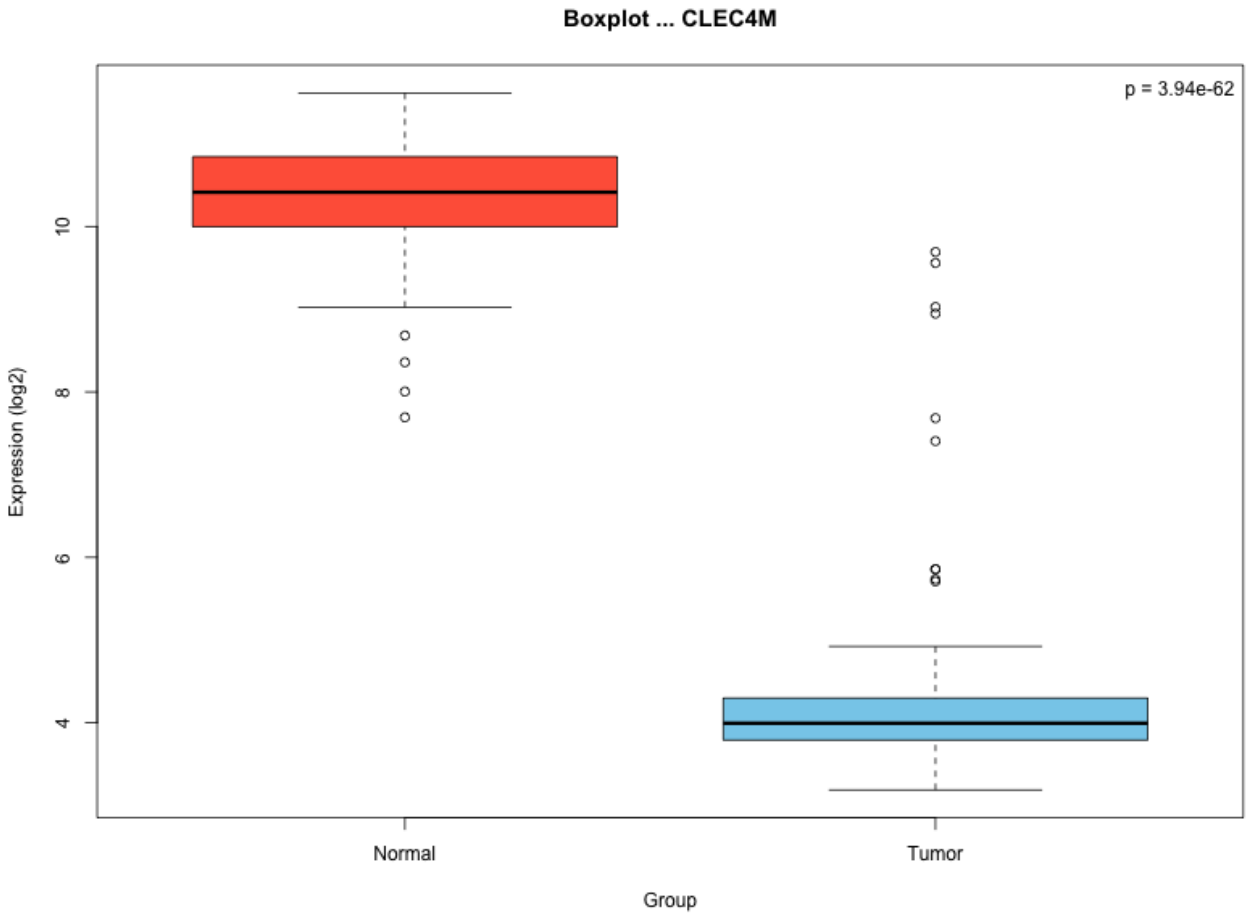
Statistical Analysis

After filtering with |logFC| > 1 and FDR < 0.05, 1668 genes were selected. These genes show the most marked differences between Tumor and Normal tissues. To explore the separation between groups and identify global patterns, we performed:

- Boxplot of a representative genes
- Heatmap of DEGs
- Clustering of genes and samples
- PCA of significant genes

Statistical Analysis

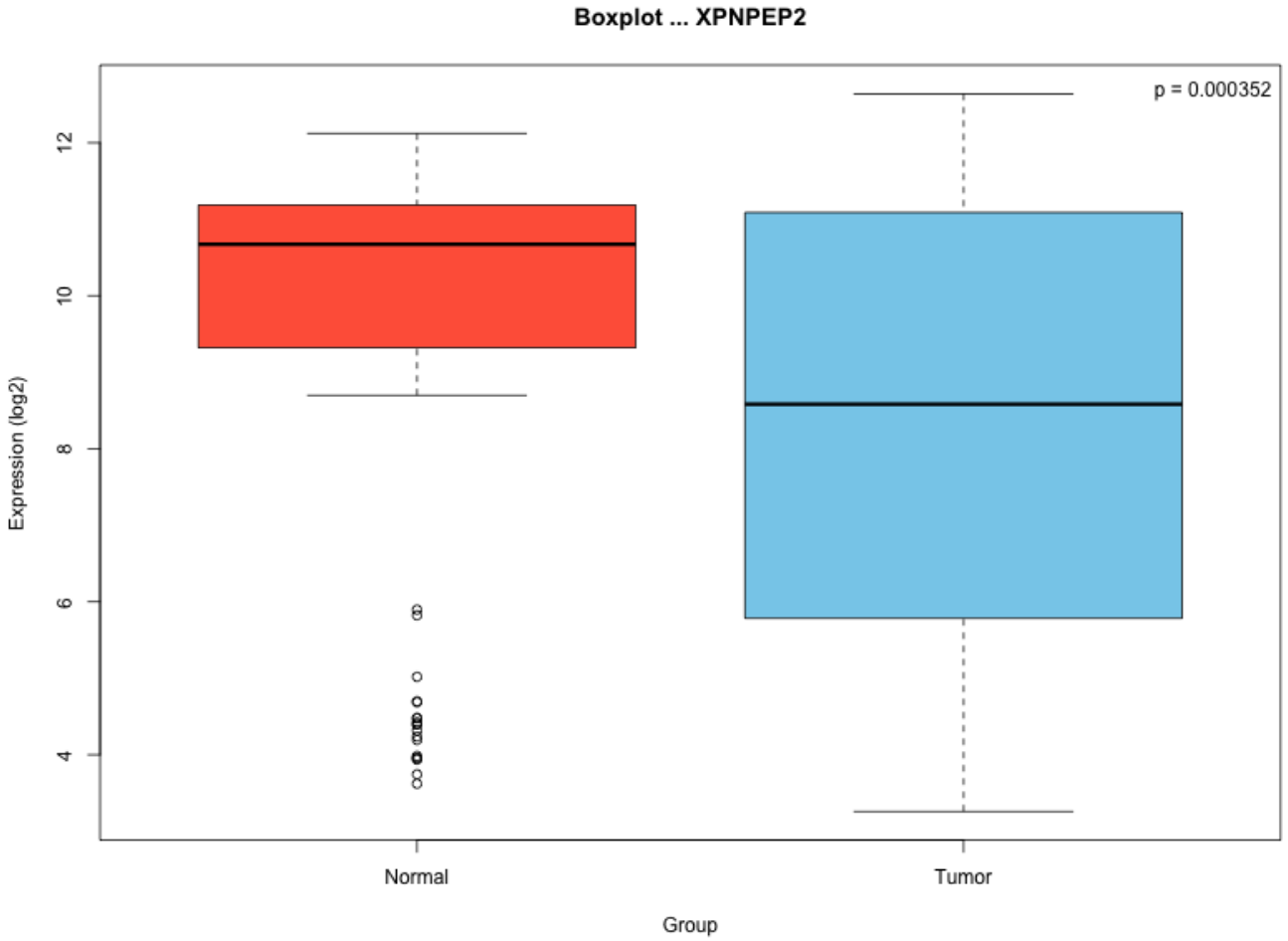
Box-plot with expression levels of the most down-regulated



- Gene selected: **CLEC4M**
- Significant difference in expression between Tumor and Normal
- P-value calculated with paired t-test: **3.935148e-62**

Statistical Analysis

Box-plot with expression levels of the most up-regulated



- Gene selected : **XPNPEP2**
- Significant difference in expression between Tumor and Normal
- P-value calculated with paired t-test: **0.0003519497**

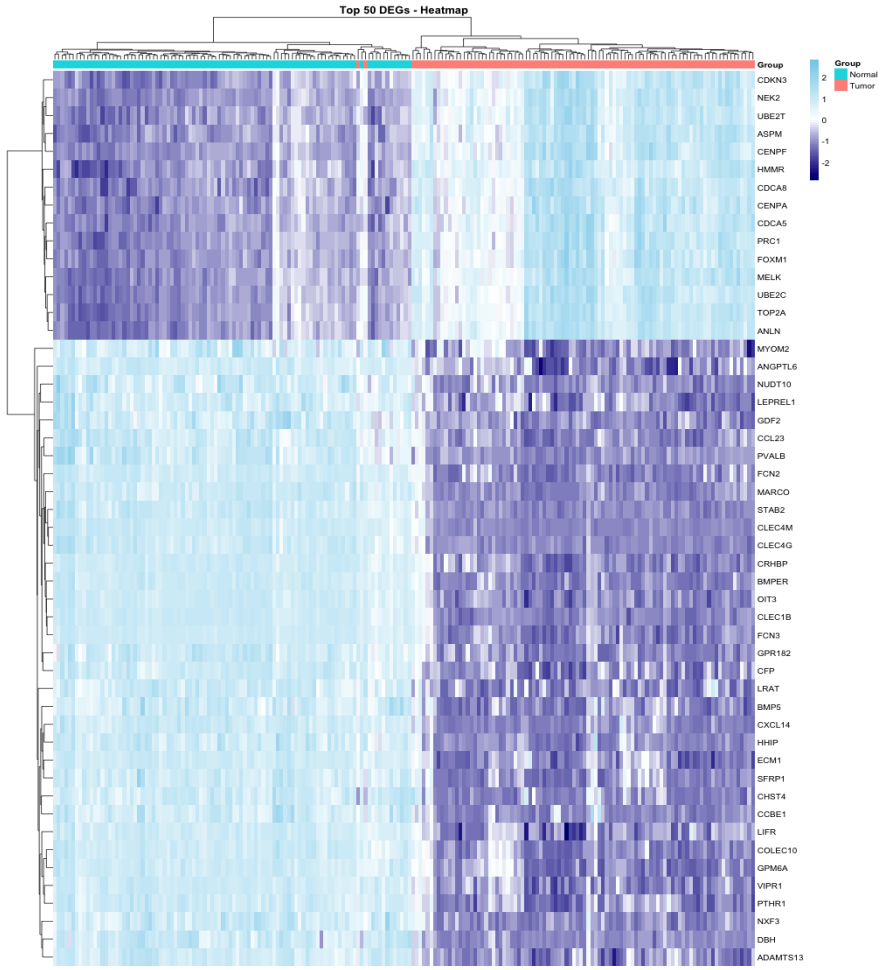
Statistical Analysis

Box-plot conclusion

These genes represent promising candidates for biomarker discovery or therapeutic targeting, given the strong and significant expression differences observed.

Statistical Analysis

Heatmap of DEGs



Were hierarchically clustered **the** Genes (rows) and samples (columns) then log-transformed and Z-score normalized the expression values are across genes.

The separation between tumor and normal samples is visible, indicating consistent expression changes

Statistical Analysis

PCA of significant genes

The tumor and normal samples are clearly separated in the graph.
This means that the gene expressions of the two groups are very different and that the DEGs "capture" this difference well, highlighting a strong biological signal

Statistical Analysis

Box-plot conclusion

These genes represent promising candidates for biomarker discovery or therapeutic targeting, given the strong and significant expression differences observed.

DEGs Filtered Data

Gene	LogFC	P-Value
AADAT	-2.41080291778972	2.7940251877568e-28
AASS	-1.26623923438982	1.845571136531e-12
ABCA8	-2.09260533414012	5.411756786144 52e-16
ABCA9	-1.32709945530499	2.60546508148203e-17
ABCB4	-1.04111727326471	1.871323443563 31e-11
ABCC4	1.20745294513929	5.6559144920498e-15
ABCC9	-1.42427084899771	3.773949776777 05e-20
ABCG2	-1.37676856112457	9.312193614347 47e-13
ABI3BP	-1.87855754799953	2.364988963443 34e-18

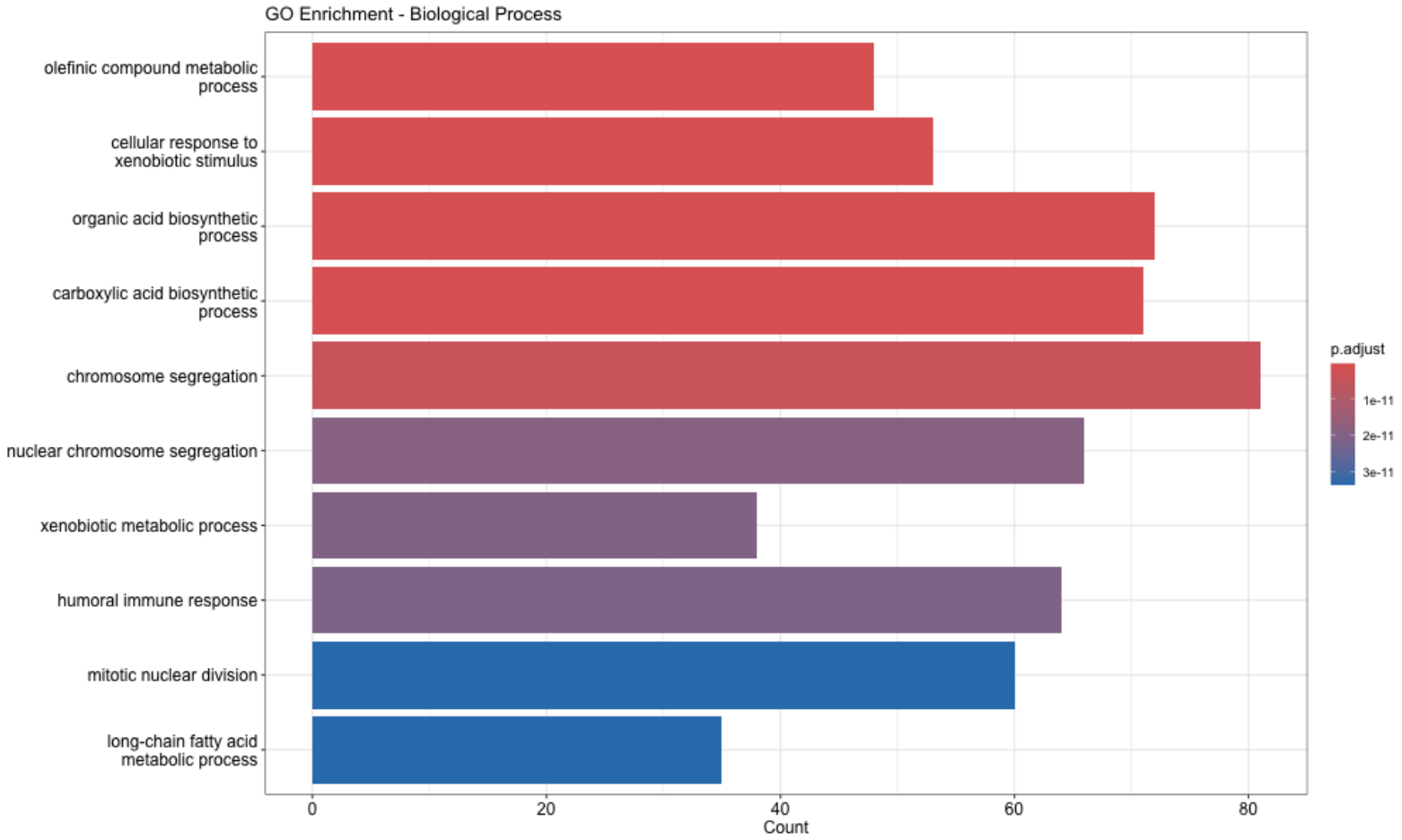
In the following table we can see some of the differentially expressed genes that were exported into the DEGs_filtered.txt file to allow further functional analysis.

Functional Enrichment Analysis

To understand the biological significance of DEGs, I performed a functional enrichment analysis focusing on the following contents:

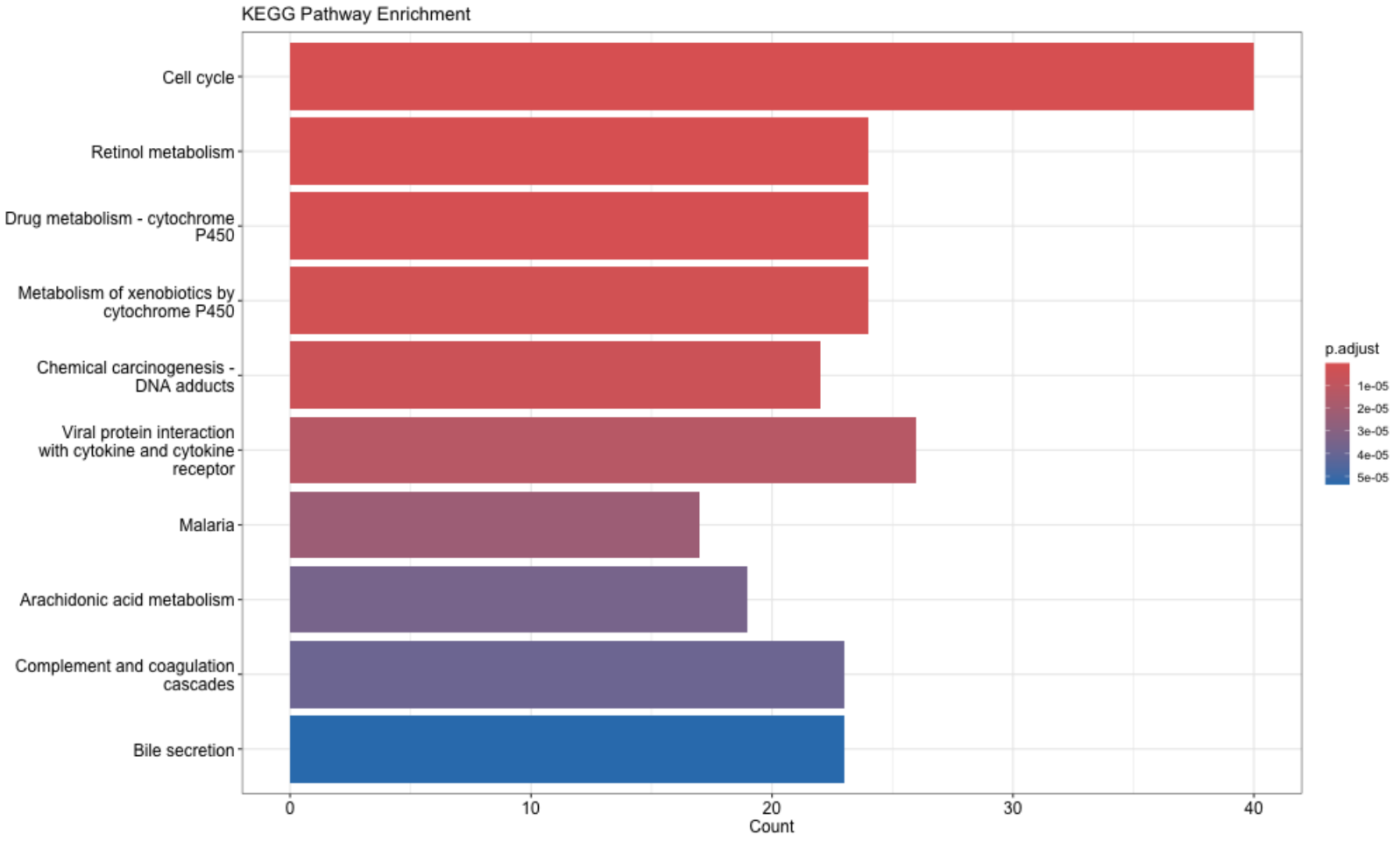
- Gene Ontology (GO): associated biological functions
- KEGG Pathways: cellular processes involved

BarPlot - Top GO terms



This graph shows us how differentially expressed genes are involved in key processes such as immune response, cell adhesion and regulation of apoptosis.

KEGG pathway enrichment



The identified genes participate in critical cancer-related pathways. These findings help to understand the molecular mechanisms of HCC progression.

Additional Analysis

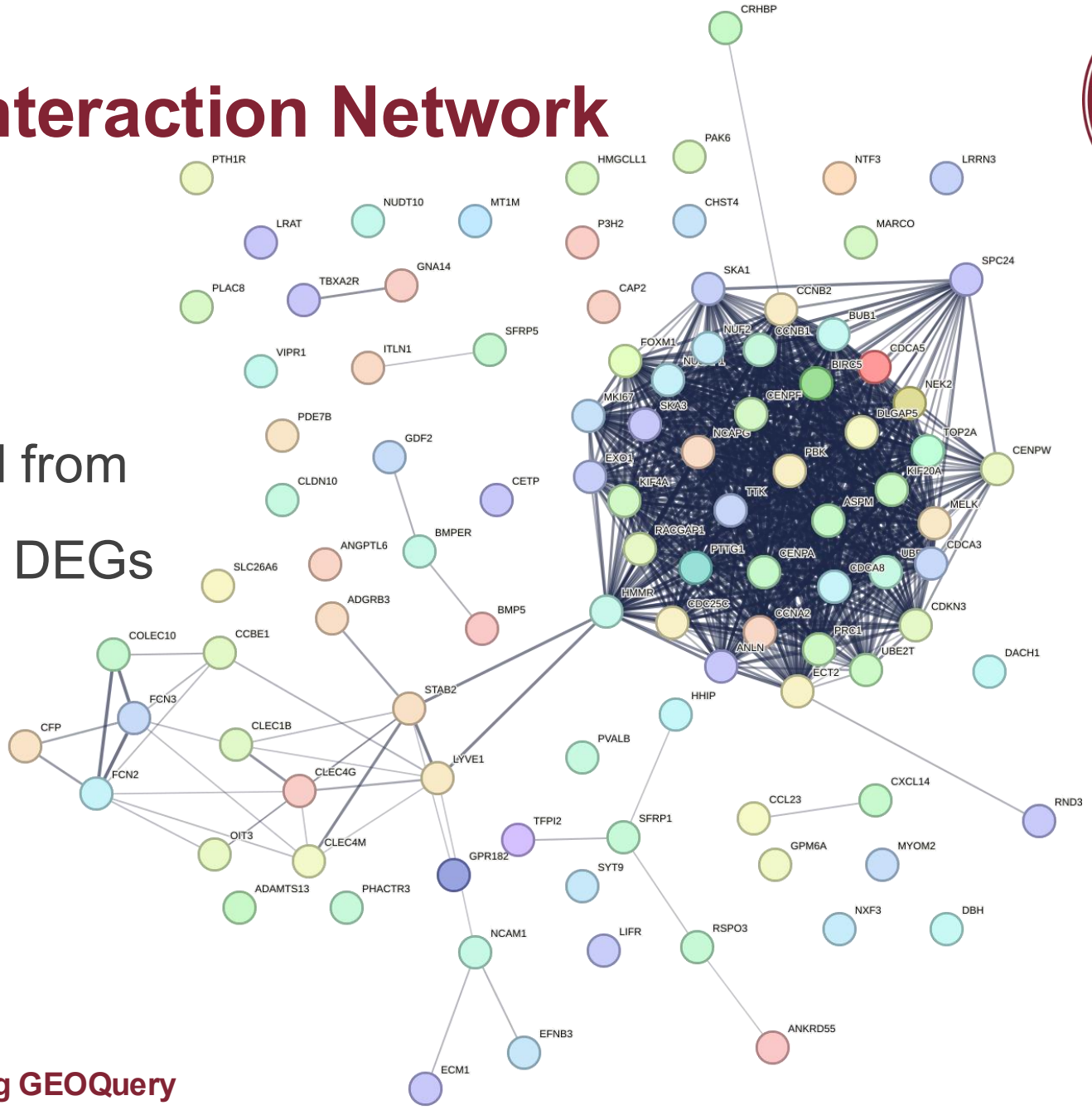
To further investigate the biological relevance of DEGs, I analyzed their protein-protein interaction (PPI) network using the STRING database.

The resulting network highlights genes that are closely related in their function and potential key regulators (“hub genes”).

This analysis helps identify **molecular modules** involved in HCC progression.

Protein-Protein Interaction Network

Network picture obtained from STRING-db with top 100 DEGs



miRNA Analysis

Some of the DEGs identified (e.g., CYP3A4, GPC3) are known to be regulated by specific microRNAs such as **miR-122**, **miR-21**, or **miR-199a**, according to the miRTarBase and TargetScan databases.
This suggests that deregulated miRNAs may contribute to the altered gene expression observed in HCC.

miRNA Analysis

ID	Species (miRNA)	Species (Target)	miRNA	Target	Validation methods: Strong evidence: Reporter assay	Strong evidence: Western blot	Strong evidence: qPCR	Less strong evidence: Microarray	Less strong evidence: NGS	Less strong evidence: pSILAC	Less strong evidence: Other	Less strong evidence: CLIP-Seq	Sum	# of papers
MIRT000012	Homo sapiens	Homo sapiens	hsa-miR-122-5p	CYP7A1	✔		✔				✔		3	1
MIRT000364	Homo sapiens	Homo sapiens	hsa-miR-122-5p	IGF1R	✔	✔	✔				✔		4	3
MIRT000365	Homo sapiens	Homo sapiens	hsa-miR-122-5p	SRF	✔	✔	✔				✔		4	1
MIRT000663	Homo sapiens	Homo sapiens	hsa-miR-122-5p	RAC1	✔	✔	✔	✔			✔		5	2
MIRT000717	Homo sapiens	Homo sapiens	hsa-miR-122-5p	RHOA	✔						✔		2	2
MIRT003006	Homo sapiens	Homo sapiens	hsa-miR-122-5p	CCNG1							✔		7	5
MIRT003075	Mus musculus	Mus musculus	mmu-miR-122-5p	Tmem50b							✔		1	1
MIRT003076	Mus musculus	Mus musculus	mmu-miR-122-5p	Lass6							✔		1	1
MIRT003077	Mus musculus	Mus musculus	mmu-miR-122-5p	Gpx7							✔		1	1
MIRT003079	Homo sapiens	Homo sapiens	hsa-miR-122-5p	GTF2B			✔				✔		2	1
MIRT003080	Homo sapiens	Homo sapiens	hsa-miR-122-5p	GYS1		✔	✔	✔			✔		4	2
MIRT003081	Homo sapiens	Homo sapiens	hsa-miR-122-5p	ANK2	✔		✔				✔		3	1

Some of the DEGs regulated by microRNA miR-122 obtained from the database https://www.targetscan.org/.

miRNA Analysis

The miR-122 analyzed in the previous slide is a liver-specific miRNA and a key regulator in liver development and liver diseases; its loss is associated with hepatitis C virus (HCV) infection , hepatocellular carcinoma (HCC) [2], and treatment resistance of HCC.
([PubMed Central])

Literature Research

To further support our computational results, I conducted a literature-based validation focused on previously analyzed miR-122. Using TargetScan, I identified that miR-122 potentially regulates several DEGs found in the previous analysis, including CYP3A4, GPC3.

In addition, recent studies on PubMed Central, confirm the downregulation of miR-122 in HCC tissues and its association with tumor progression, metastasis and chemoresistance.

Final Conclusions

Summarizing the analyses covered in the presentation including:

- Visualization via volcano plots and heat maps confirmed clear differences in expression between case and control samples.

- Further validation via STRING and bibliographic databases (TargetScan, PubMed) supported the biological significance of selected genes and regulatory miRNAs (e.g., miR-122);

We conclude that HCC involves significant alterations in gene expression that can be interpreted biologically and clinically, thus validating the power of differential expression analysis for cancer studies.