Gene Expression Analysis of Hepatocellular Carcinoma using GEOQuery Dataset

Carmen Trovato

2033862

Academic Year: 2024/2025

Bioinformatics Course

Master's Degree in Engineering in Computer Science

SAPIENZA
UNIVERSITÀ DI ROMA

Presentation Outline

- Introduction
- Dataset Overview
- Data Preprocessing
- Filtering
- Statistical Analysis
- Functional Enrichment Analysis
- Additional Analyses
- Literature Research
- Final Conclusions

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

The GSE22058 dataset was downloaded from the public GEO (Gene Expression Omnibus) database using the R programming language and the GEOquery package.

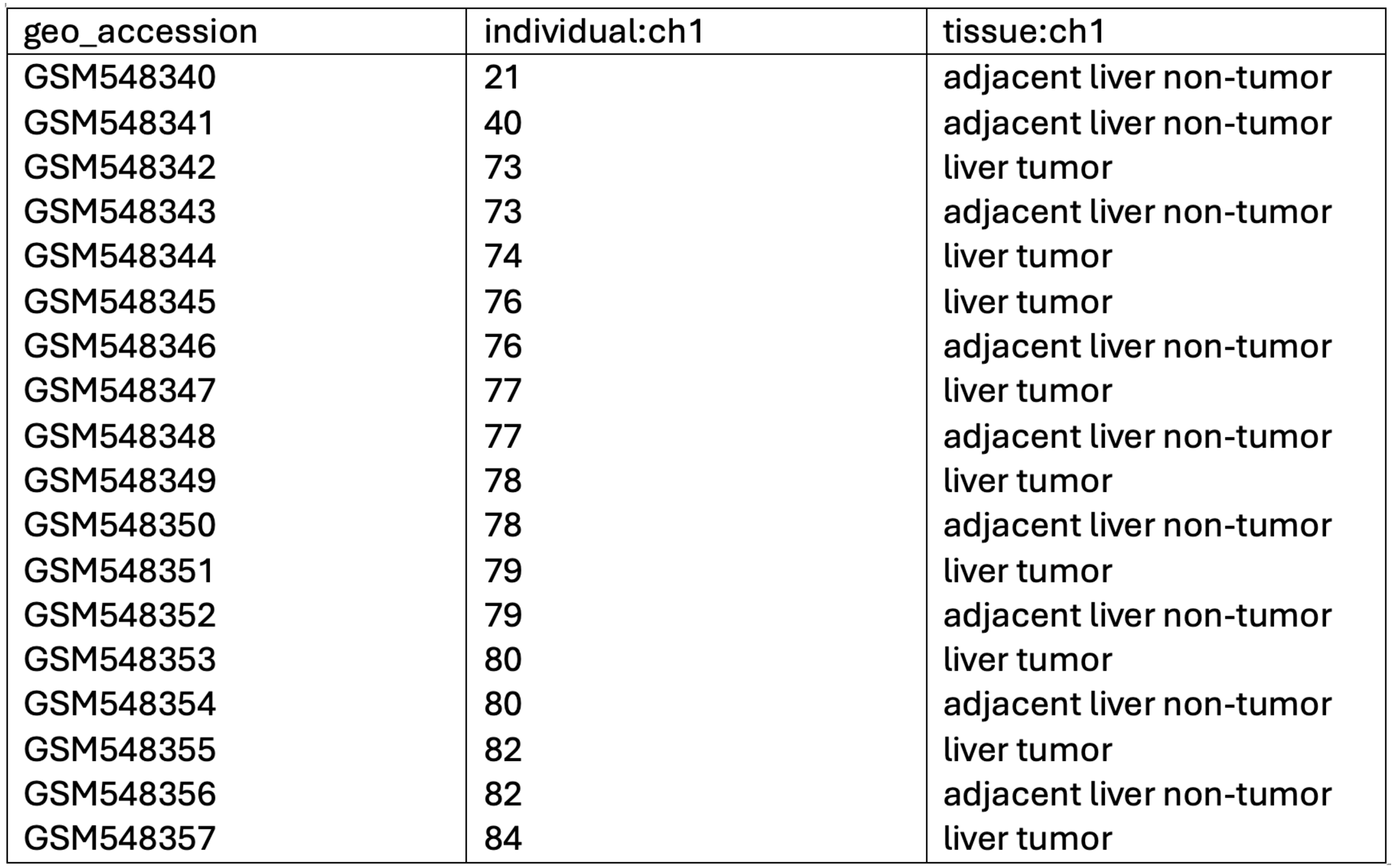
It includes **192 gene expression samples** obtained via ***microarray*** from **96 patients** diagnosed with hepatocellular carcinoma (HCC).

For each patient, **two tissue samples** were analyzed:

- **Tumor tissue (HCC)**
- **Adjacent non-tumor liver tissue (control)**

The dataset allows for **paired gene expression analysis** between tumor and healthy tissues from the same individual.

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
- **geo_accession** → Sample ID

Some results from metadata.txt

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 – IQR histogram

- The Interquartile Range (IQR) measures gene variability across all samples
- Most genes show **low variability (IQR < 1)**
- To reduce noise, **genes below the 10th percentile of IQR** were removed
- These low-variability genes are considered **uninformative** for differential expression

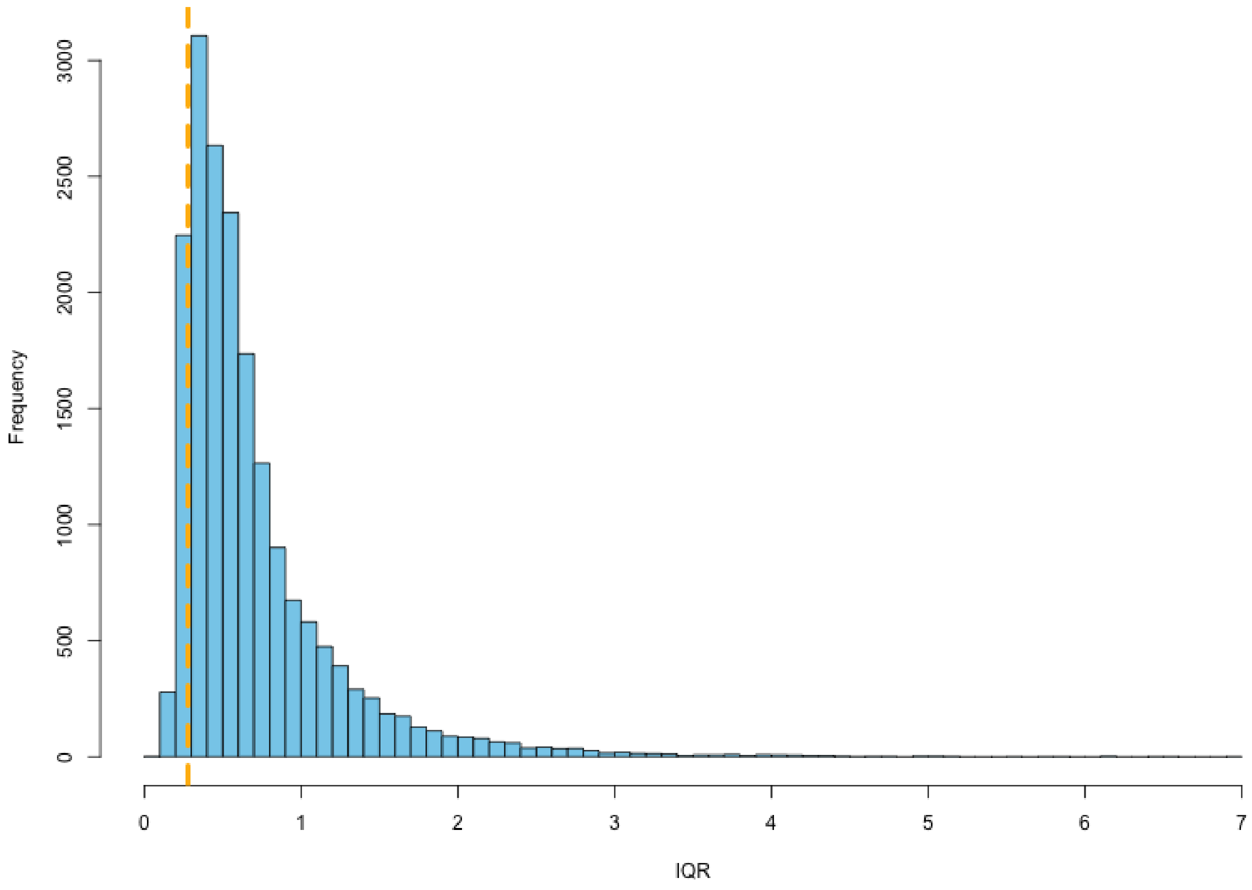
IQR frequency distribution

Data Preprocessing

Pre-processing – IQR Filtering of Genes

Histogram shows clustering of genes at low IQR values

- **Orange dashed line** : 10th percentile (***0.2780***)
- **Removed**: 1851
- **Retained:** 16652



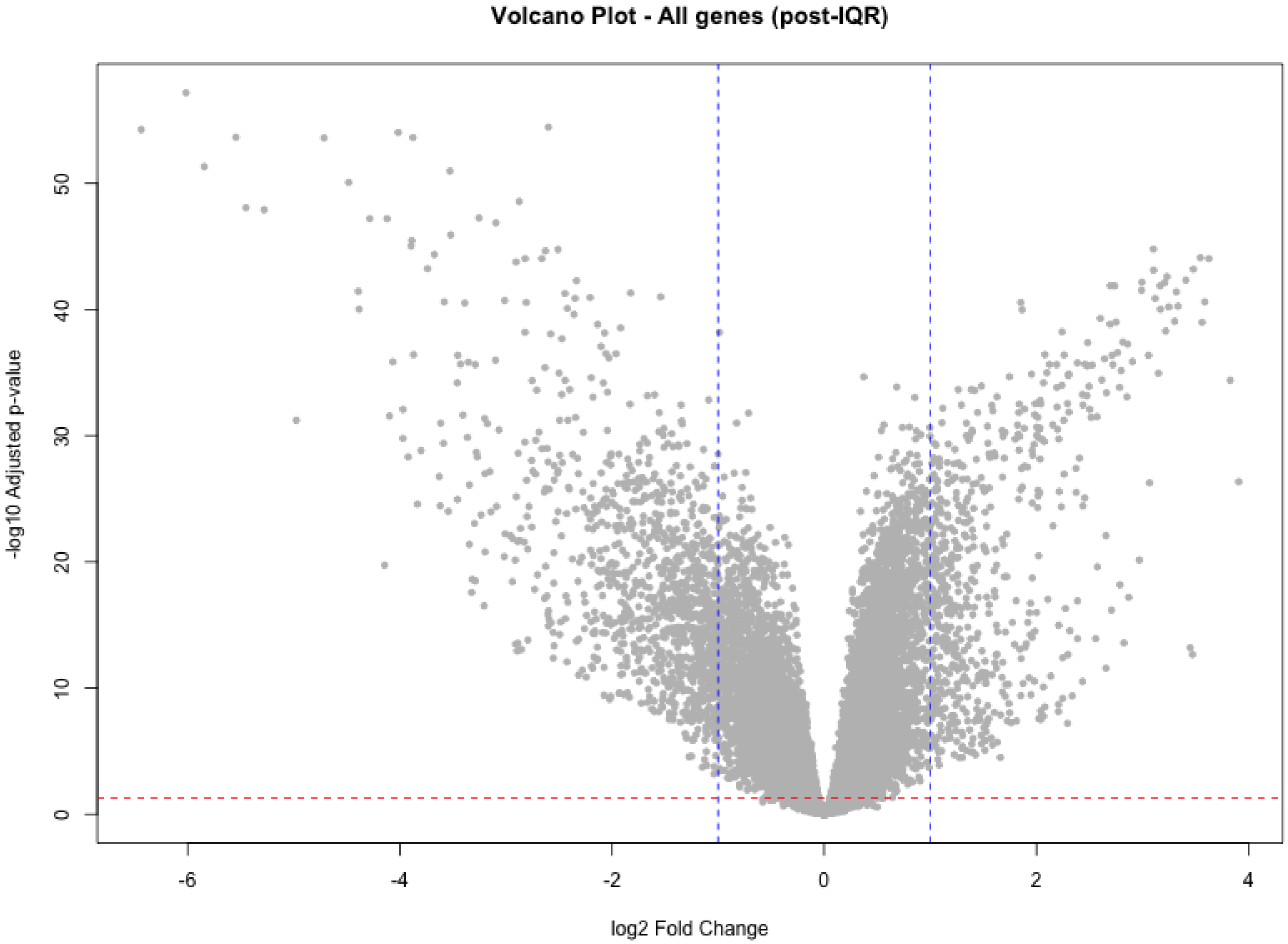
IQR frequency distribution

Filtering

After filtering out uninformative and low-variability genes, I proceed with differential expression analysis to identify genes significantly deregulated in hepatocellular carcinoma compared to healthy tissue. We can understand the results of the differential analysis by looking at the boxplot below, which allows us to refocus our attention on how gene expression changes between the two conditions (cancer/healthy).

Filtering

Vulcano-Plot before filtering

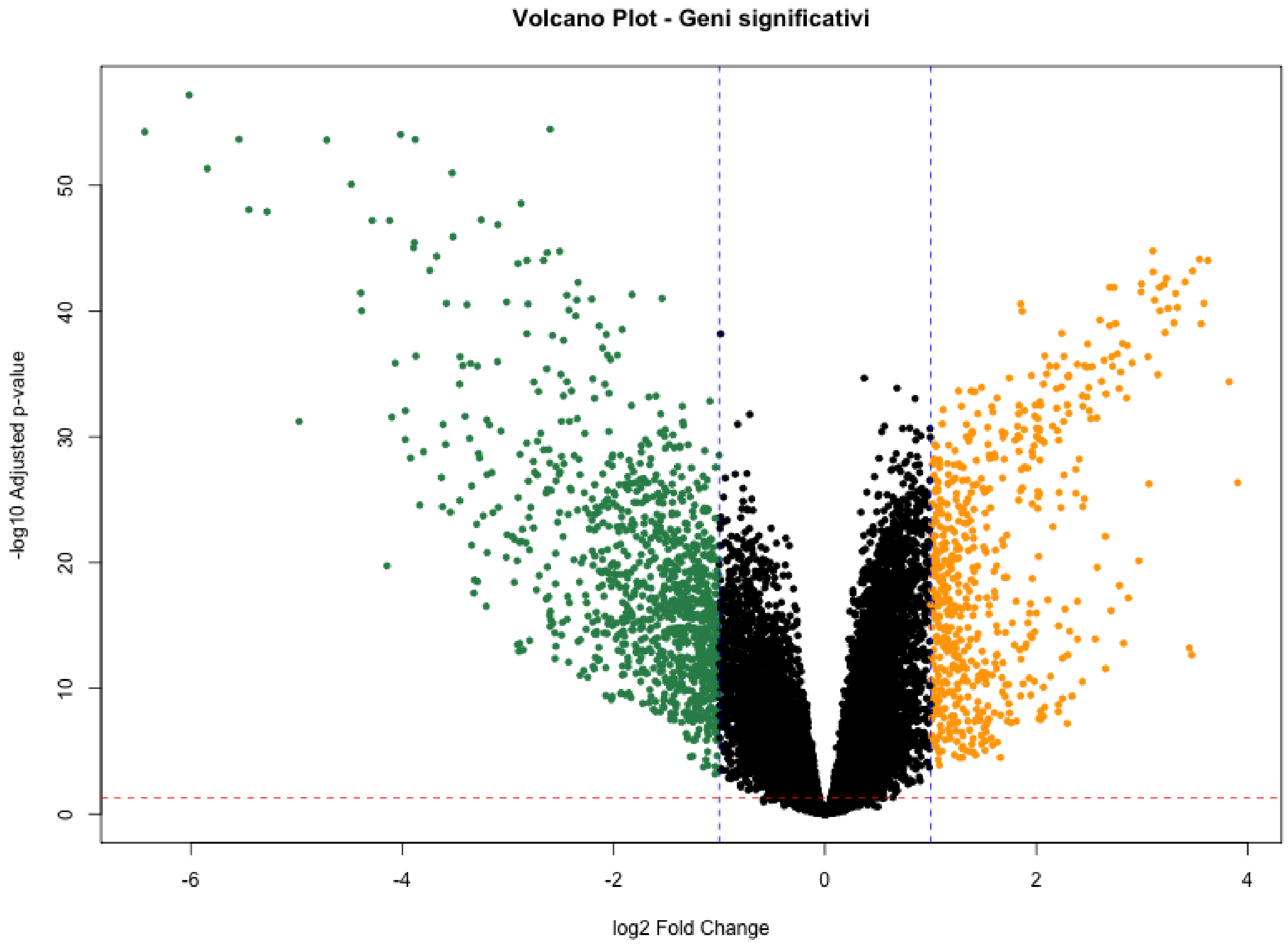


Filtering

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.

Filtering

Vulcano-plot post filtering



🟧 **Upregulated | 603 genes**

🟩 **Downregulated | 1065**

⬛ **Non-significant genes | 14984**

(low difference + high p-value)

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

Resume about Vulcano-plot

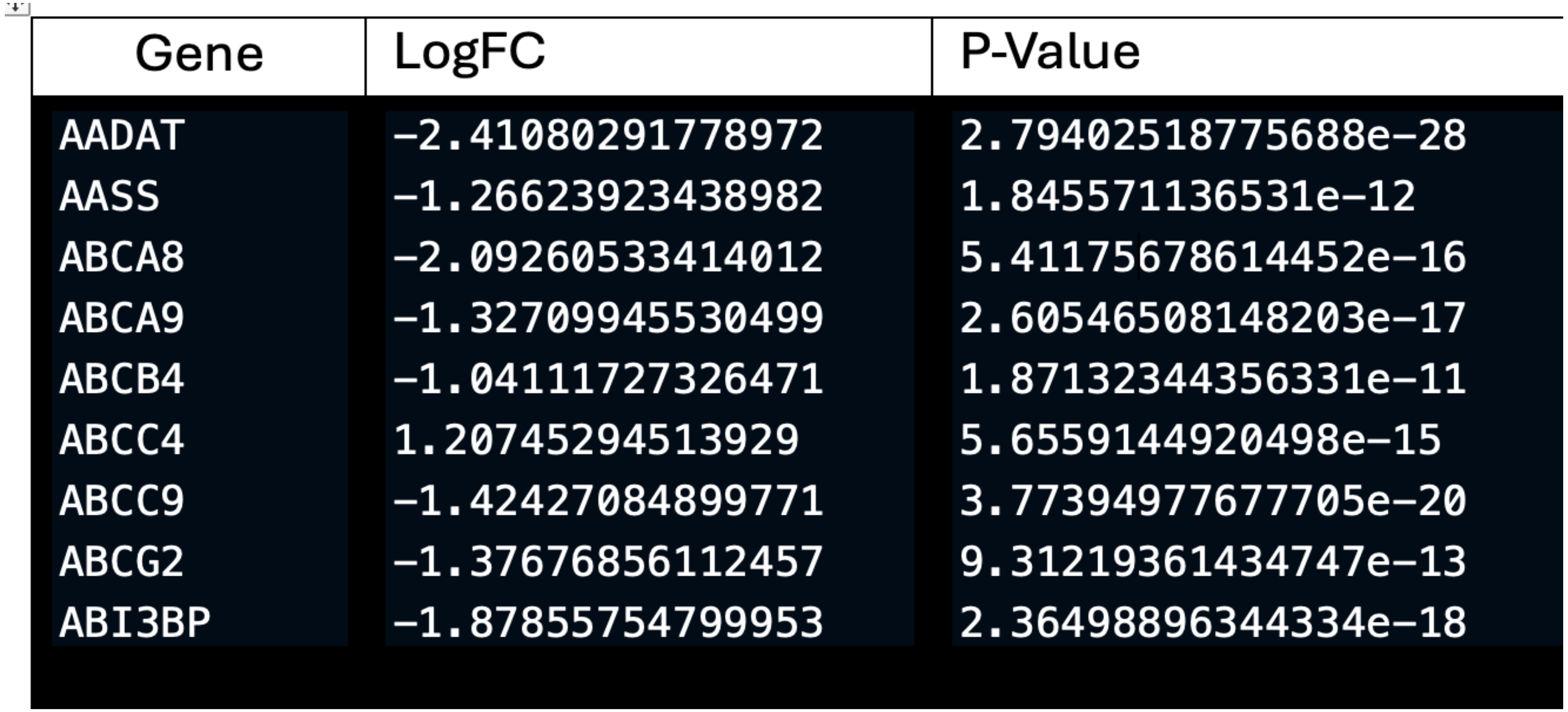
In this step we applied a combined filter on log fold change and FDR, selecting only genes with strong variation and statistical significance. This step is crucial to avoid false positives and focus on the most relevant genes for the difference Tumor vs Normal.

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

DEGs Filtered Data

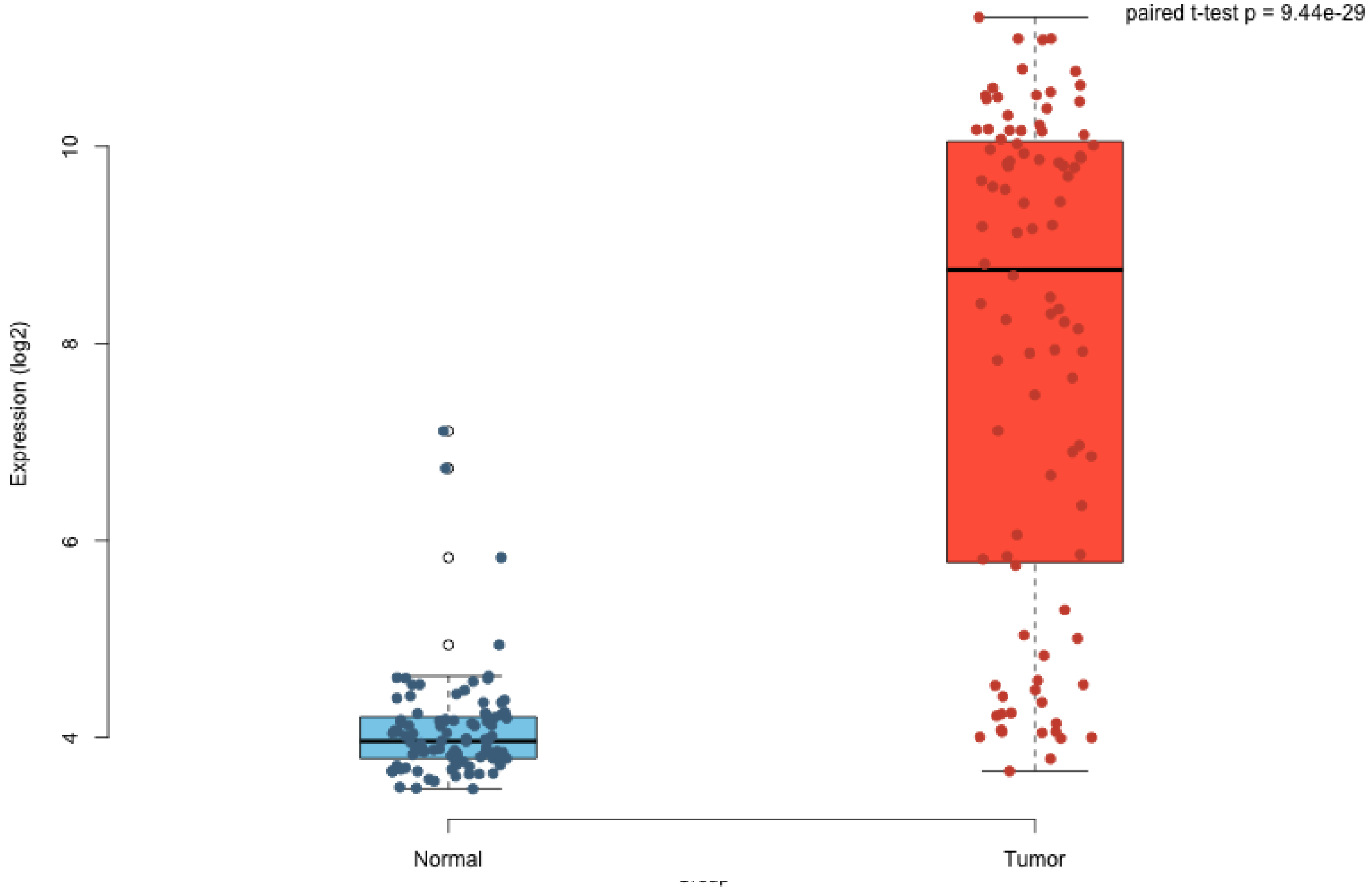


Gene	LogFC	P-Value
AADAT	−2.41080291778972	2.79402518775688e−28
AASS	−1.26623923438982	1.845571136531e−12
ABCA8	−2.09260533414012	5.41175678614452e−16
ABCA9	−1.32709945530499	2.60546508148203e−17
ABCB4	−1.04111727326471	1.87132344356331e−11
ABCC4	1.20745294513929	5.6559144920498e−15
ABCC9	−1.42427084899771	3.77394977677705e−20
ABCG2	−1.37676856112457	9.31219361434747e−13
ABI3BP	−1.87855754799953	2.36498896344334e−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.

Statistical Analysis

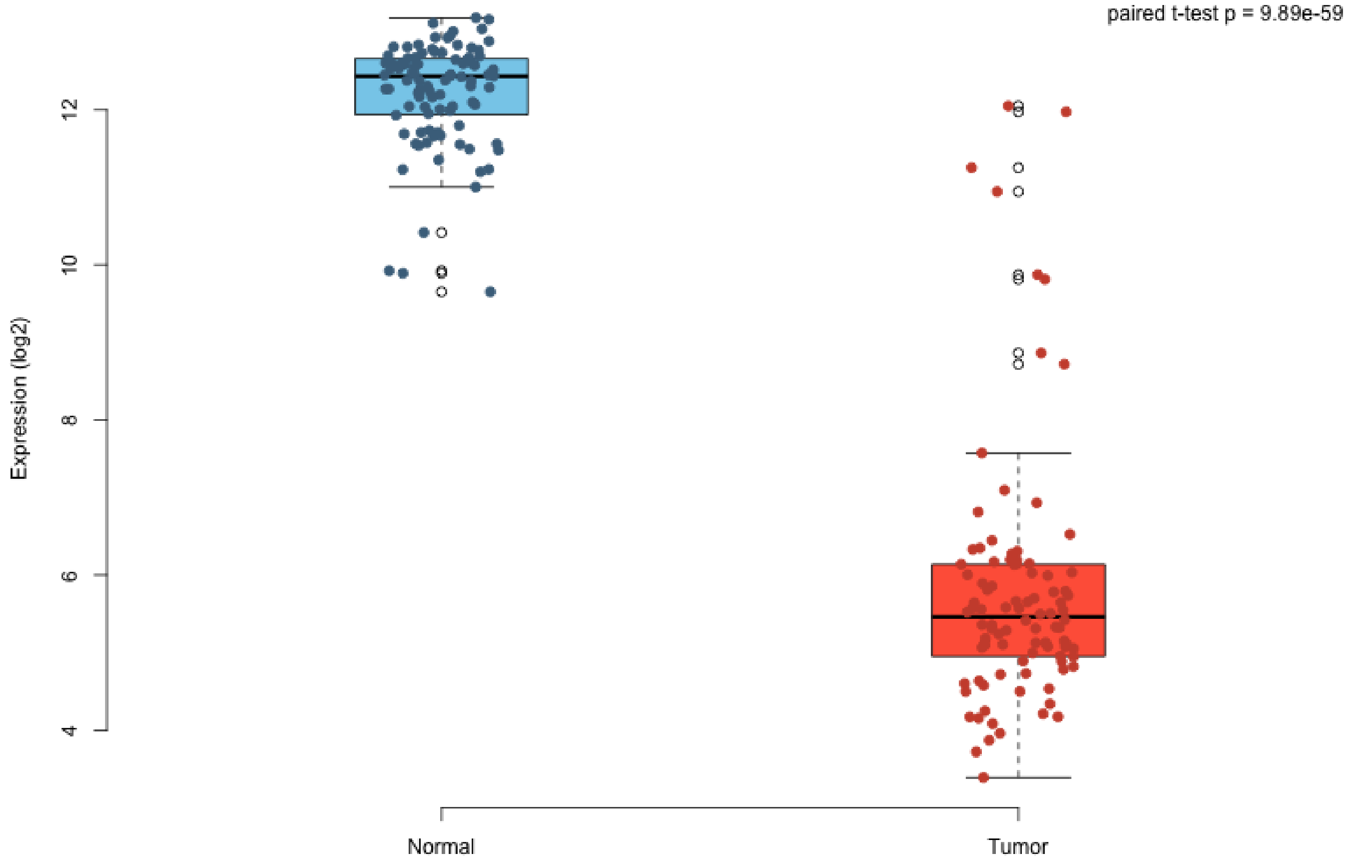
Boxplot of the most upregulated gene



- Gene selected: *ZIC2*
- Pval ≈ 9.4e-29
- Adj-pval ≈ 4.4e-27
- logFC ≈ 3.9

Statistical Analysis

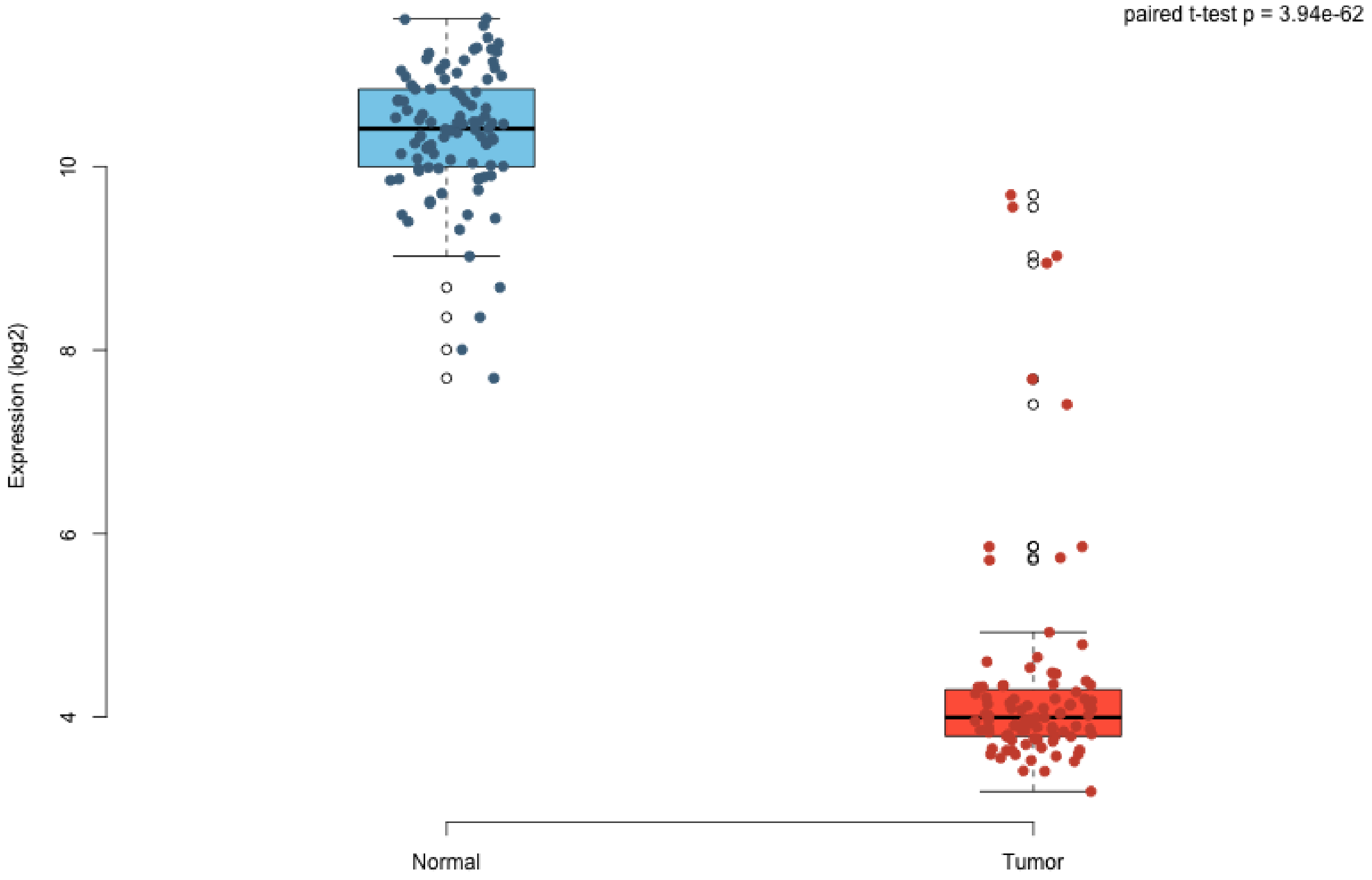
Boxplot of the most downregulated gene



- Gene selected: *CLEC1B*
- Pval ≈ 9.9e-59
- Adj-pval ≈ 5.5e-55
- logFC ≈ -6.4

Statistical Analysis

Boxplot of the most significant gene



- Gene selected: *CLEC4M*
- Pval ≈ 3.9e-62
- Adj-pval ≈ 6.6e-58
- logFC ≈ -6.01

Statistical Analysis

Box-plot conclusions

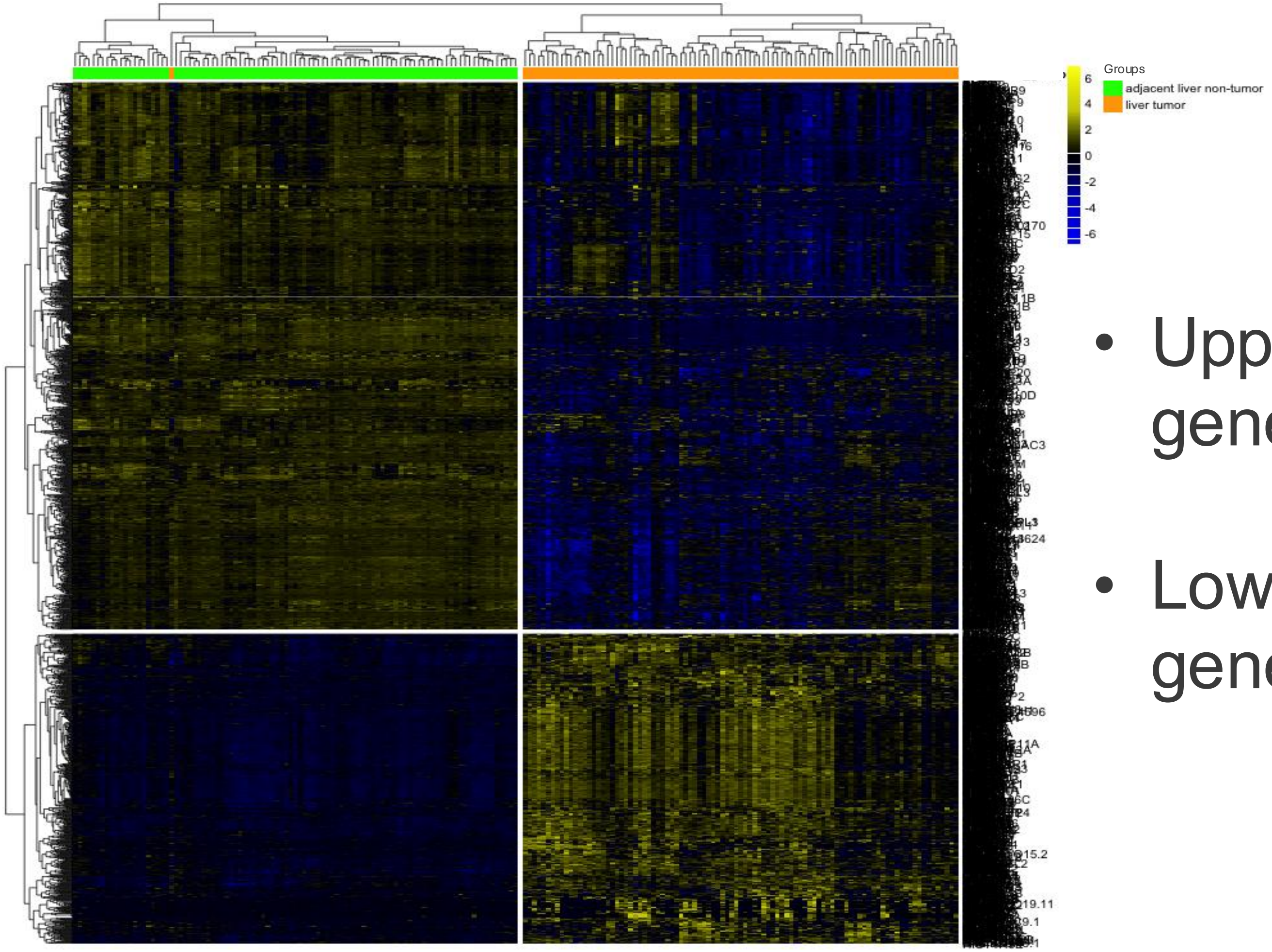
***ZIC2* -** logFC, ≈15-fold more expressed in the tumor. Highly heterogeneous tumor expression, so it may indicate a gene involved only in tumor subgroups

***CLEC4M* -** logFC, ≈87-fold less expressed in the tumor. Markedly and consistently reduced expression in all tumors. Indicates a possible silenced tumor suppressor gene in the tumor.

***CLEC4M* -** underexpressed in the tumor ≈65 –fold less. Reduced expression in nearly all tumor samples supports the idea that it is involved in tumor biology

Statistical Analysis

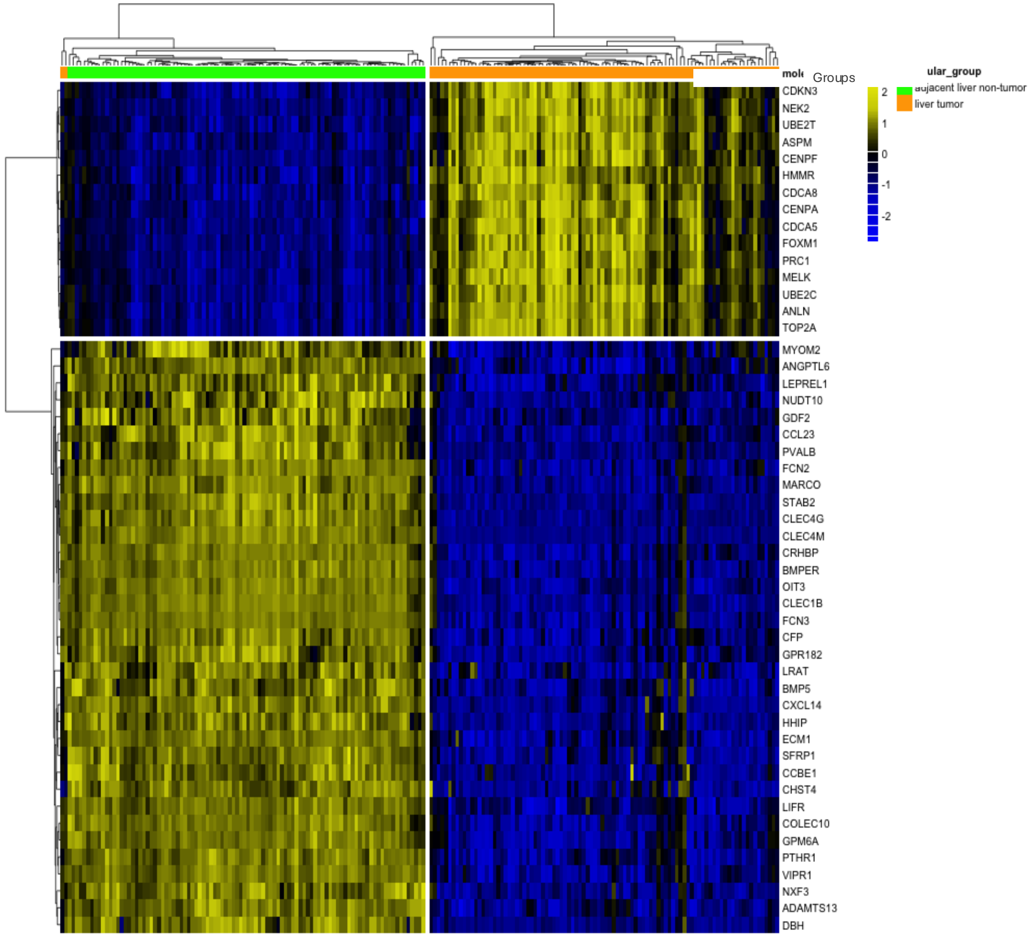
Heatmap of DEGs



- Upper part (darker in tumors) → genes overexpressed in tumors
- Lower part (darker in normals) → genes underexpressed in tumors.

Statistical Analysis

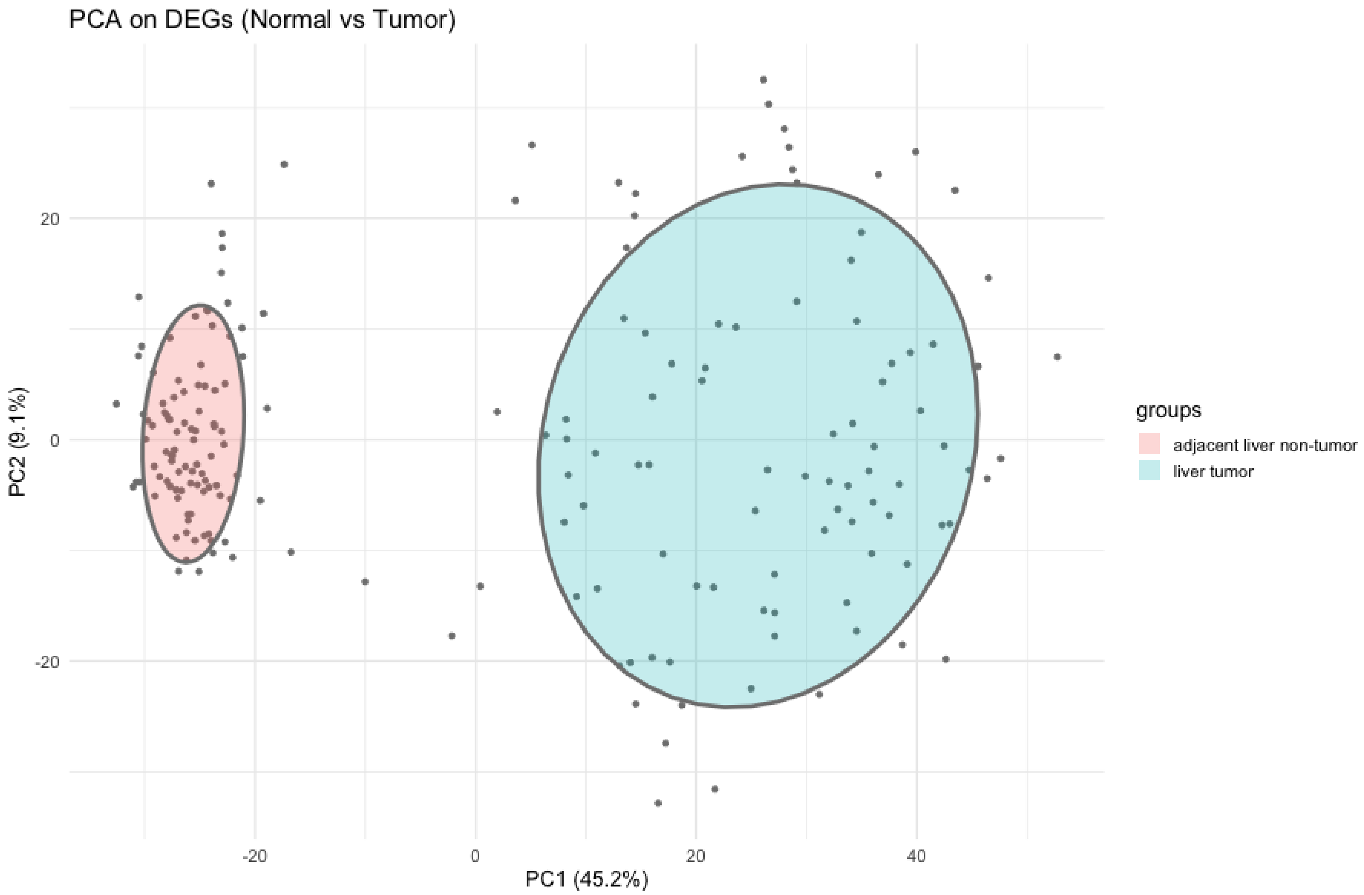
Heatmap of top 50 DEGs



Tumor samples (orange color) and adjacent non-tumor samples (green) tend to be clearly separated, confirming that gene expression robustly distinguishes the two groups.

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

PCA conclusion

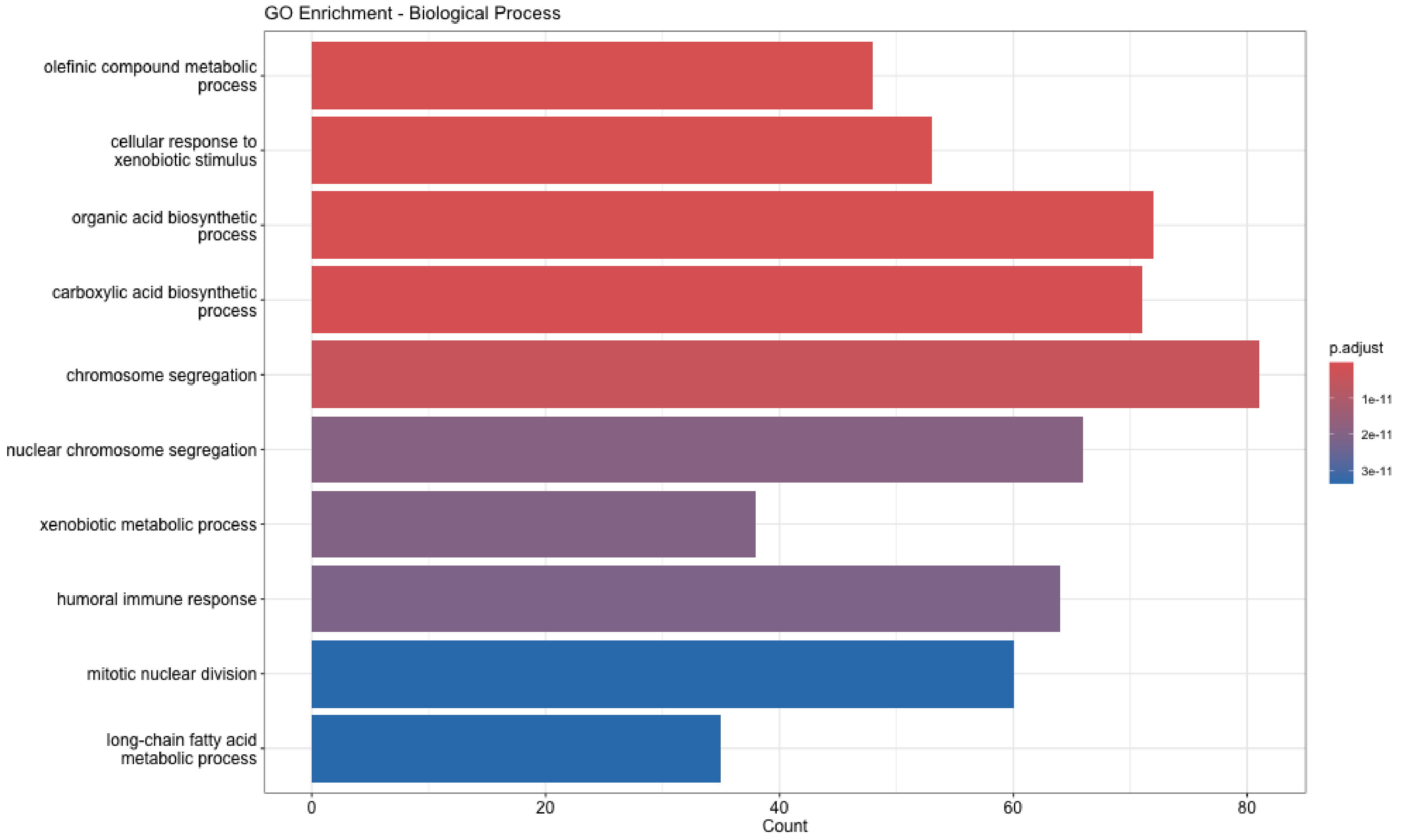
These genes represent promising candidates for biomarker discovery or therapeutic targeting, given the strong and significant expression differences observed.

Functional Enrichment Analysis

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

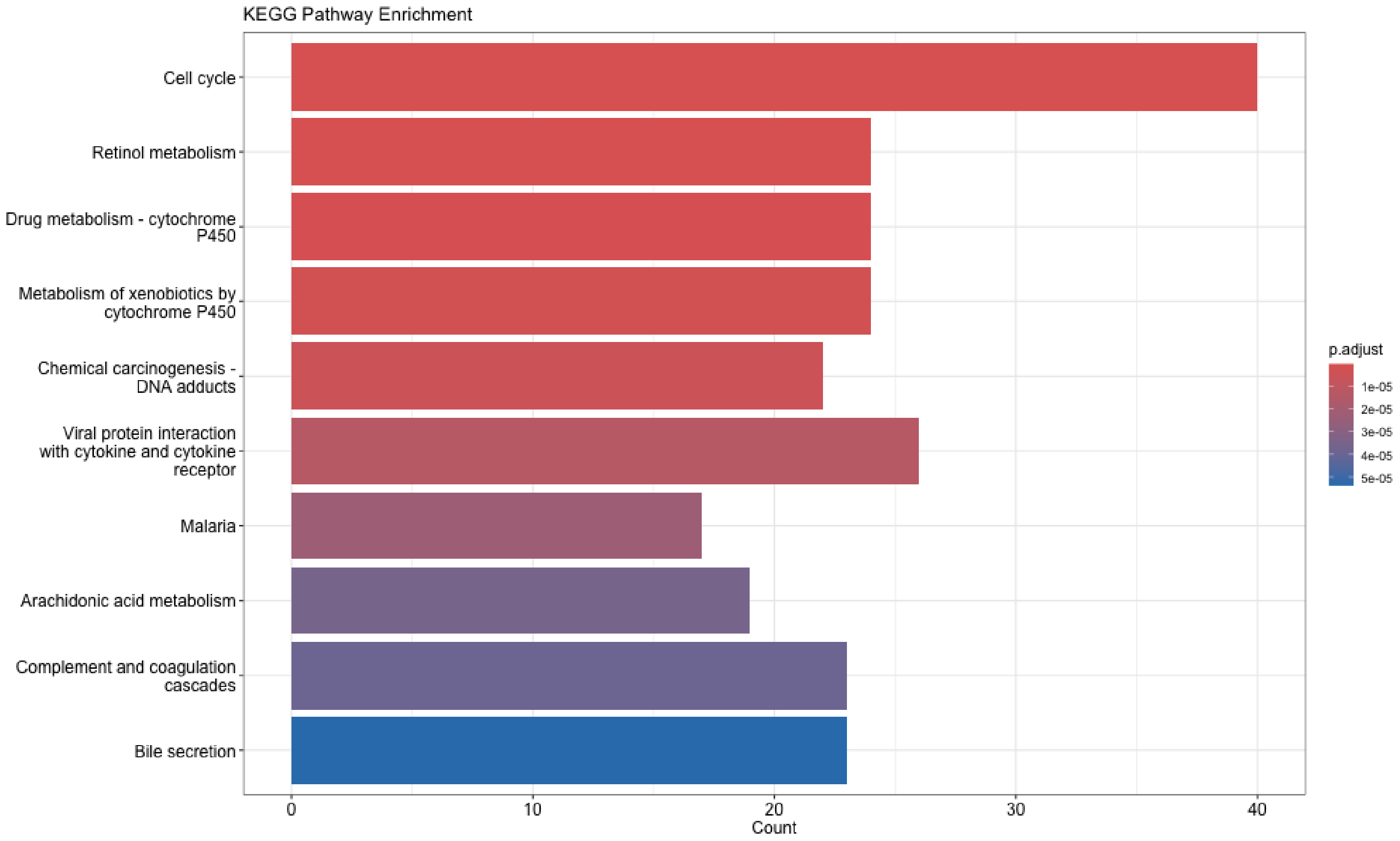
- **Gene Ontology** (GO): associated biological functions
- **KEGG Pathways**: cellular processes involved

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.

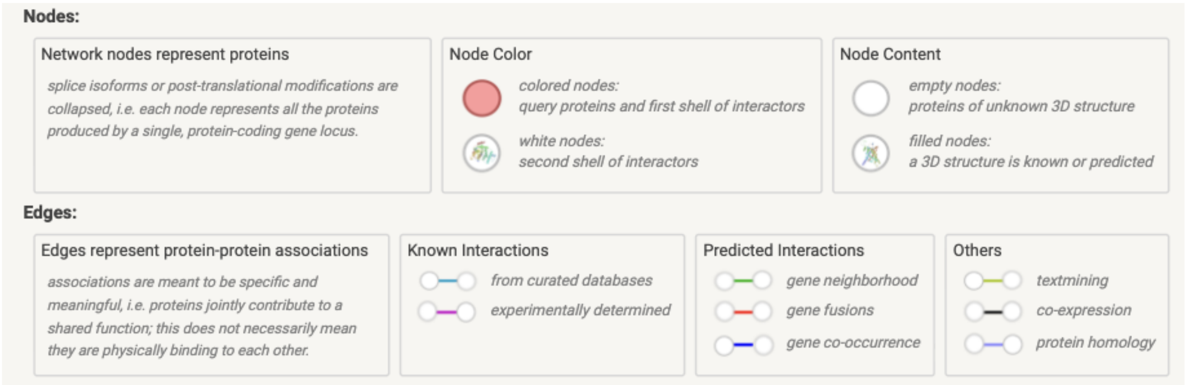
The resulting network highlights genes that are closely related in their function and potential key regulators (“hub genes”).

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

Protein-Protein Interaction Network

Network picture obtained from

STRING-db with top 100 DEGs



Nodes:

Network nodes represent proteins

splice isoforms or post-translational modifications are collapsed, i.e. each node represents all the proteins produced by a single, protein-coding gene locus.

Node Color

🟥 *colored nodes:* *query proteins and first shell of interactors*

○ *white nodes:* *second shell of interactors*

Node Content

○ *empty nodes:* *proteins of unknown 3D structure*

● *filled nodes:* *a 3D structure is known or predicted*

Edges:

Edges represent protein-protein associations

associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding to each other.

Known Interactions

○—○ *from curated databases*

○—○ *experimentally determined*

Predicted Interactions

○—○ *gene neighborhood*

○—○ *gene fusions*

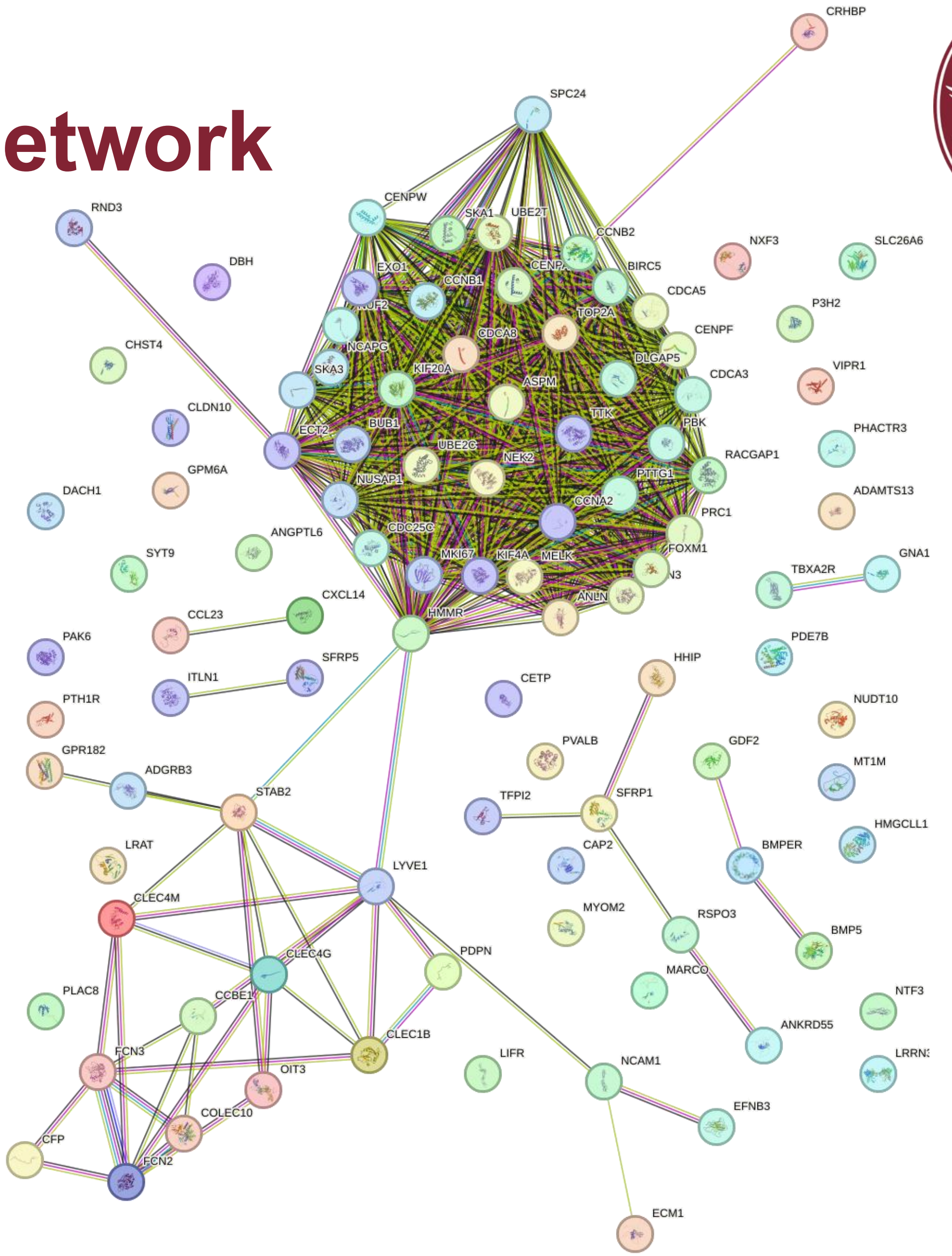
○—○ *gene co-occurrence*

Others

○—○ *textmining*

○—○ *co-expression*

○—○ *protein homology*



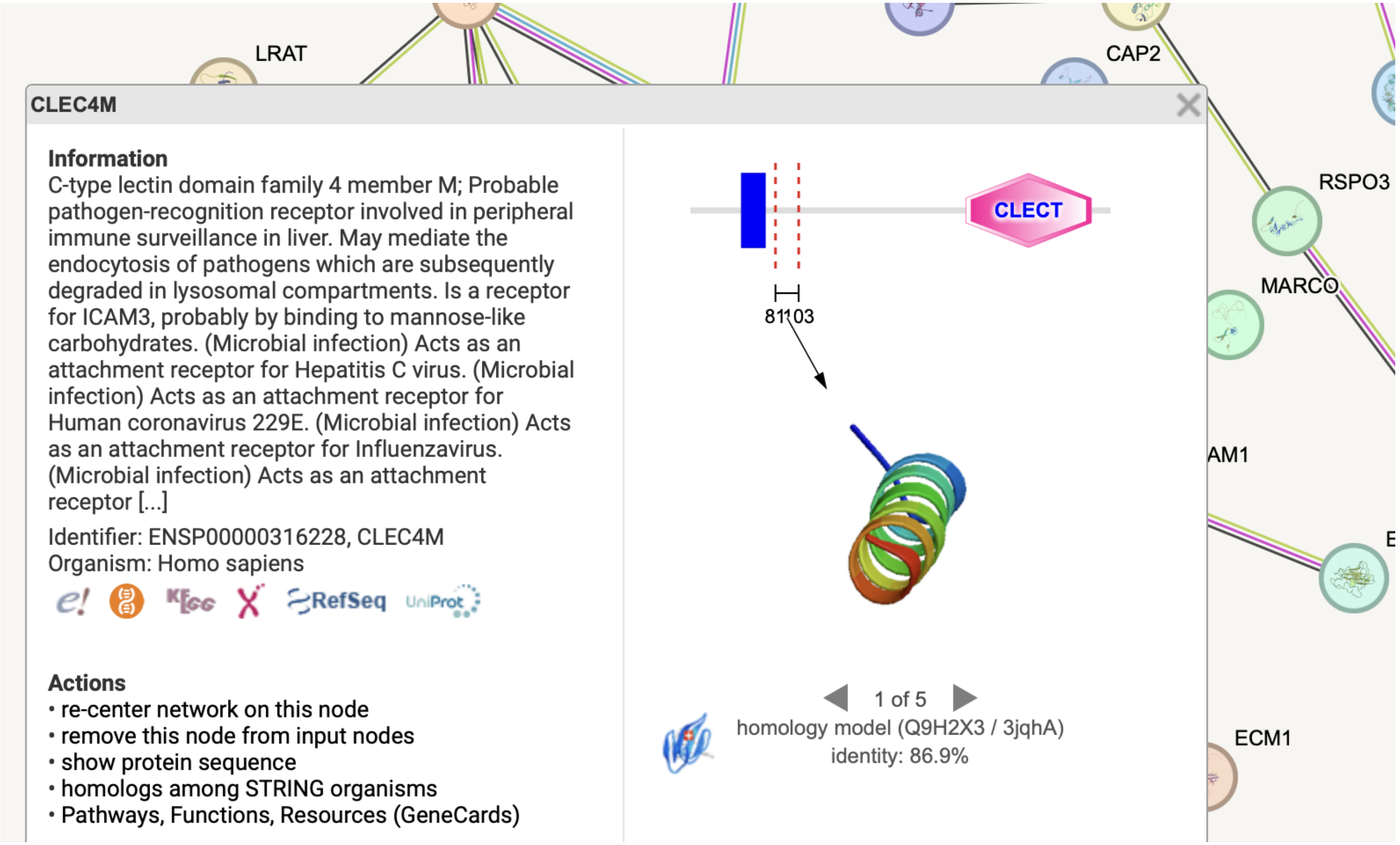
https://string-db.org/cgi/network?taskId=bF87efluz9nO&sessionId=bu2HSoyhrTGG

Protein-Protein Interaction Network

The Protein–Protein Interaction (PPI) network of the top 100 DEGs, obtained with STRING-db, shows that genes are not isolated but form functional modules. In particular, clusters linked to cell proliferation (e.g., CCNB1, CDCA8, UBE2C) and clusters associated with liver and immune functions (e.g., CLEC4M, MARCO, STAB2) are highlighted. These results suggest that the DEGs participate in common biological pathways, confirming the role of proliferative and immunometabolic processes in the development of HCC.

https://string-db.org/cgi/network?taskId=bF87efluz9nO&sessionId=bu2HSoyhrTGG

Protein-Protein Interaction Network



CLEC4M

Information
C-type lectin domain family 4 member M; Probable pathogen-recognition receptor involved in peripheral immune surveillance in liver. May mediate the endocytosis of pathogens which are subsequently degraded in lysosomal compartments. Is a receptor for ICAM3, probably by binding to mannose-like carbohydrates. (Microbial infection) Acts as an attachment receptor for Hepatitis C virus. (Microbial infection) Acts as an attachment receptor for Human coronavirus 229E. (Microbial infection) Acts as an attachment receptor for Influenzavirus. (Microbial infection) Acts as an attachment receptor [...]

Identifier: ENSP00000316228, CLEC4M
Organism: Homo sapiens

e! KEGG RefSeq UniProt

Actions
- re-center network on this node
- remove this node from input nodes
- show protein sequence
- homologs among STRING organisms
- Pathways, Functions, Resources (GeneCards)

◀ 1 of 5 ▶
homology model (Q9H2X3 / 3jqhA)
identity: 86.9%

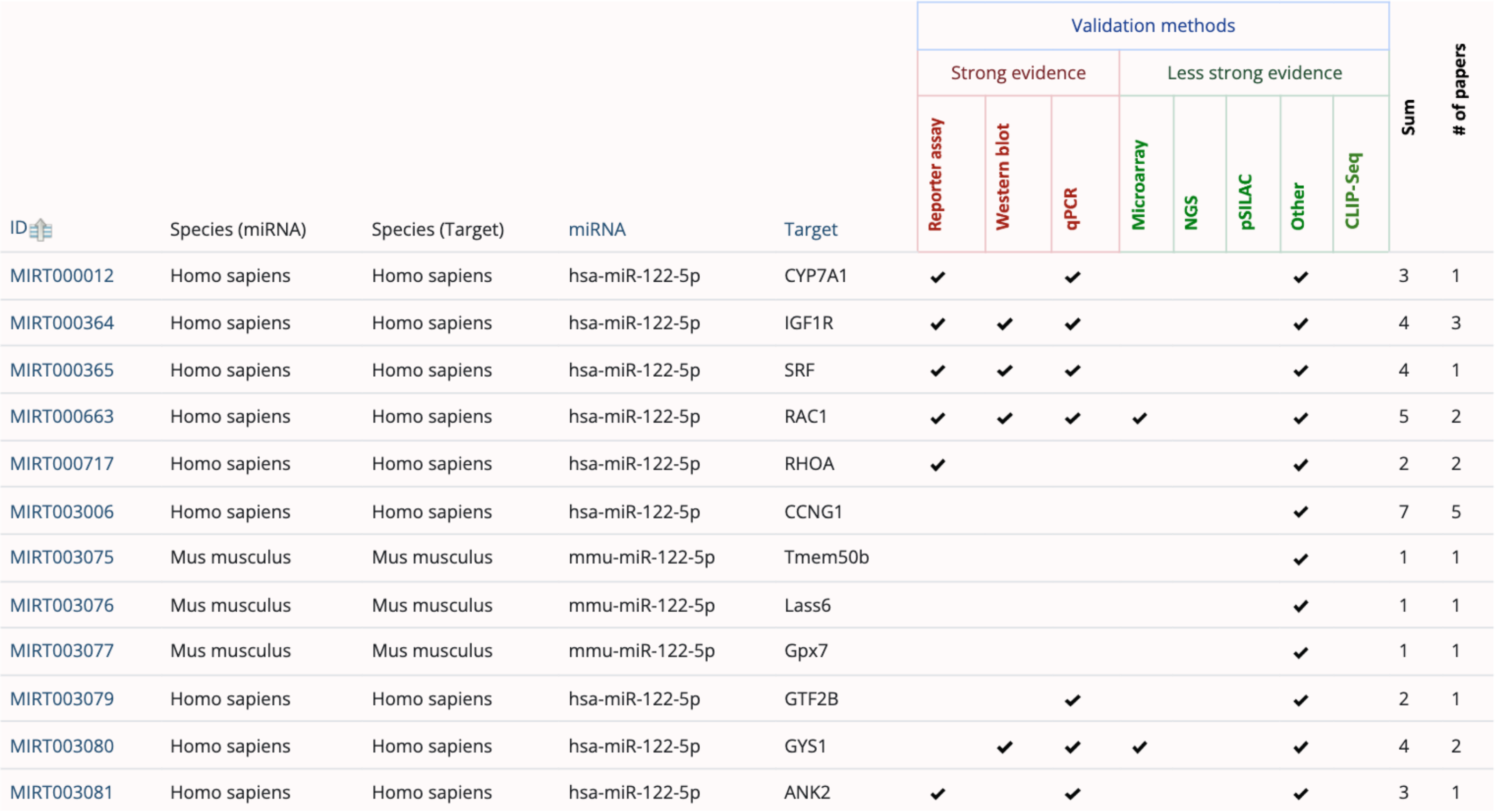
CLEC4M was found to be the most significantly differential gene. Its underexpression in tumor samples suggests altered immune surveillance in hepatocellular carcinoma.

https://string-db.org/cgi/network?taskId=bF87efluz9nO&sessionId=bu2HSoyhrTGG

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



					Validation methods									
					Strong evidence			Less strong evidence						
ID	Species (miRNA)	Species (Target)	miRNA	Target	Reporter assay	Western blot	qPCR	Microarray	NGS	pSILAC	Other	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 (regulator of GPC3) 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

Several independent studies confirm that miR-122 is consistently downregulated in hepatocellular carcinoma (HCC) tissues. This loss is not only a hallmark of liver tumorigenesis, but is also associated with disease progression, metastatic potential, and resistance to treatment.

Literature Research

In line with our computational analysis, which identified GPC3 and CYP3A4 among the putative targets of miR-122, the literature strongly supports the role of this microRNA as a key regulator of liver homeostasis. Altogether, these findings highlight how the deregulation of miR-122 contributes both to the altered gene expression patterns observed in HCC and to the aggressive clinical behavior of the tumor

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.