SAPIENZA
UNIVERSITÀ DI ROMA

DIGITAL EPIDEMIOLOGY AND PERCISION MEDICINE

Differential Gene Expression Analysis in Cholangiocarcinoma through TCGA Data Exploration

Mohammadmehdi Razavi

2023856

2023-Dec

Abstract

This study addresses the critical scientific issue of understanding the molecular underpinnings of Cholangiocarcinoma, a complex and aggressive bile duct cancer. Our aim was to unravel the distinctive patterns of gene expression that could underpin its pathogenesis and progression. Utilizing the comprehensive repository of The Cancer Genome Atlas (TCGA), we performed a rigorous differential gene expression analysis to identify key genes and pathways altered in Cholangiocarcinoma. Our methodology involved advanced bioinformatics tools for data processing, normalization, and statistical analysis to ensure robust results. The study's main result was the identification of a set of differentially expressed genes, which not only delineated the unique transcriptomic signature of Cholangiocarcinoma but also highlighted potential therapeutic targets and diagnostic biomarkers. Furthermore, pathway analysis revealed significant perturbations in crucial signaling pathways, providing insights into the disease mechanism. In conclusion, our findings contribute to a deeper understanding of Cholangiocarcinoma at the molecular level, offering new avenues for research and potential clinical applications in diagnosis and therapy.

Introduction

Cholangiocarcinoma, a form of cancer that originates in the bile ducts, represents a significant scientific and medical challenge due to its typically late diagnosis and limited treatment options. This malignancy, though relatively rare, is characterized by its aggressive nature and poor prognosis, necessitating a deeper understanding of its molecular and genetic underpinnings. Recent advances in genomic sequencing and bioinformatics have paved the way for comprehensive genomic studies.

The state-of-the-art research in Cholangiocarcinoma primarily focuses on identifying key genetic alterations, understanding the pathways involved in tumorigenesis, and exploring potential therapeutic targets. Studies have highlighted the heterogeneity of this cancer, with significant variations in genetic mutations and molecular profiles. Key findings include the identification of mutations in genes like IDH1/2, KRAS, and TP53, and the role of these mutations in cancer development and progression.

Building upon this foundation, the present study hypothesizes that a detailed analysis of TCGA data specific to Cholangiocarcinoma will reveal further insights into its molecular subtypes, prognostic biomarkers, and potential therapeutic targets. This hypothesis is grounded in the belief that a more nuanced understanding of the genomic alterations in Cholangiocarcinoma can lead to better diagnostic, prognostic, and treatment strategies, ultimately improving patient outcomes.

Data Acquisition and Preprocessing:

The study utilized publicly available genomic data from The Cancer Genome Atlas (TCGA) for Cholangiocarcinoma. Our focus was on Transcriptome Profiling data category, Gene Expression Quantification data type, and STAR-Counts workflow type. Our data was divided into cancer and normal samples. For the Normal samples we have 9 samples and 60660 genes. For the cancer samples we had 35 samples and 60660 genes. For the better accuracy and sake of comparison between the two groups we only keep the samples which we have both normal and cancer samples. The intersect of two groups included 8 samples and 606660 genes. Next, we clean the data with respect to the proper labels. We also

filtered the genes that have at least 10 counts on 90% patients. For the next step we normalize our data with Deseq2 median of Ratios methods

Methods:

The Deseq2 library facilitated the identification of Differentially Expressed Genes (DEGs) with an adjusted p-value threshold of 0.05 and an absolute Fold Change (FC) greater than 1.2. Employing these DEGs, we computed gene co-expression networks for the two conditions (cancerous and normal) using the Spearman measure of similarity. We then constructed binary adjacency matrices for both cancerous and normal samples based on Spearman similarity results. Thereafter, we identified network hubs (top 5% of nodes by degree values) and compared the differential co-expression network (Cancer vs. Normal) by computing the degree index. Finally, we analyzed the Patient Similarity Network using cancer gene expression profiles and performed community detection.

Results

Upon executing differential gene expression analysis, we identified 4,412 genes, comprising 2,131 up-regulated and 2,281 down-regulated genes. The volcano plot (Figure 1) visually represents these results, illustrating the significant alterations in gene expression.

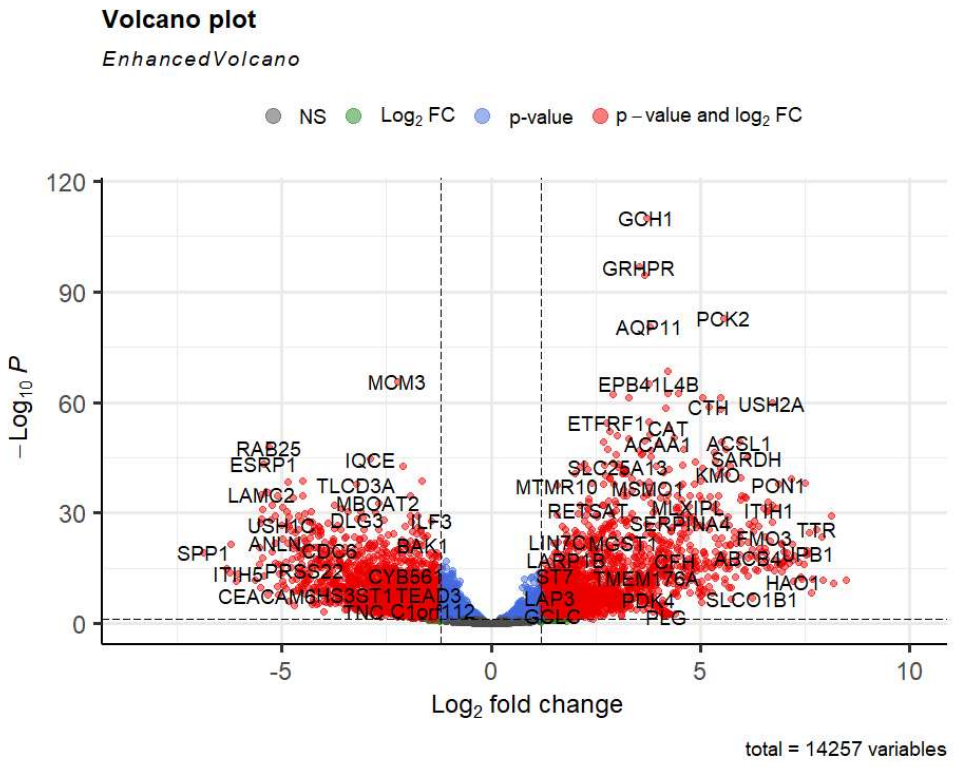


Figure 1 Volcano Plot (adjusted p-value<0.05 and |FC| < 1.2)

The construction of gene co-expression networks for both cancerous and normal samples revealed a network density of 0.00289630 for cancerous samples and 0.002459133 for normal samples. The degree distribution of nodes in each network, which adheres to a power-law distribution indicating a scale-free network, is depicted in Figure 2.

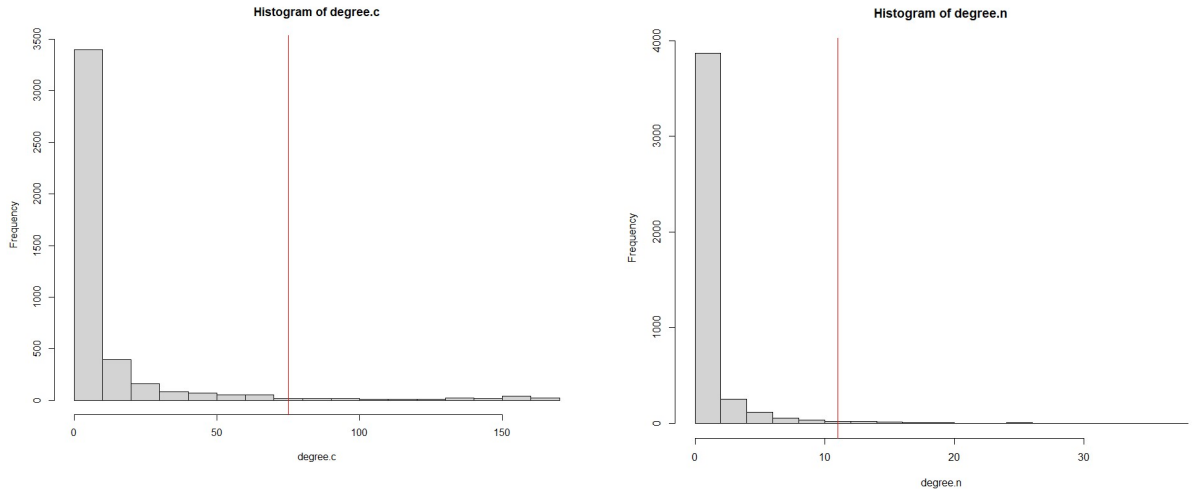


Figure 2Degree distribution of Cancer Samples (Left) and Normal Samples (Right)

For these networks, Pearson's correlation was employed. After filtering the data for |p| < 0.9, binary adjacency matrices were constructed. Figure 3 shows the cancerous and normal networks based on these matrices.

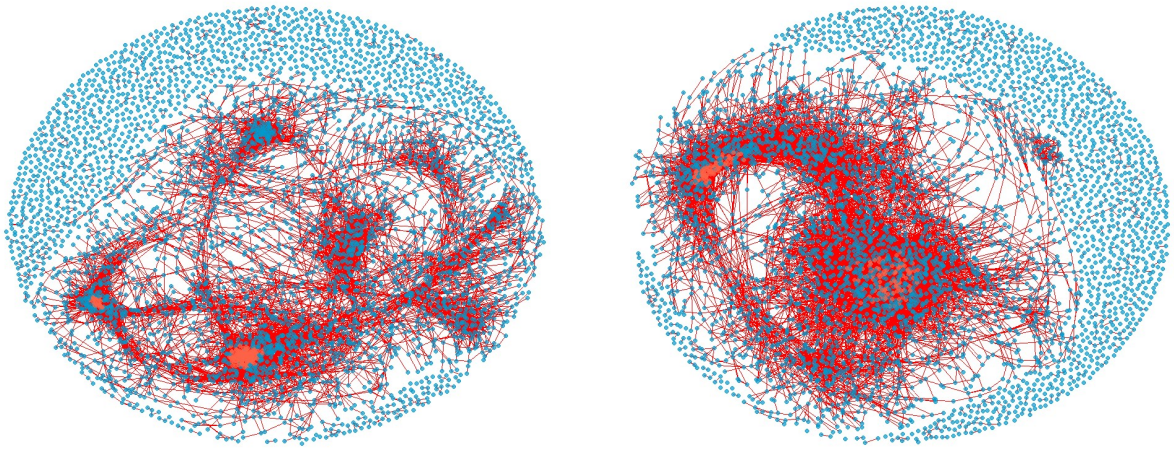
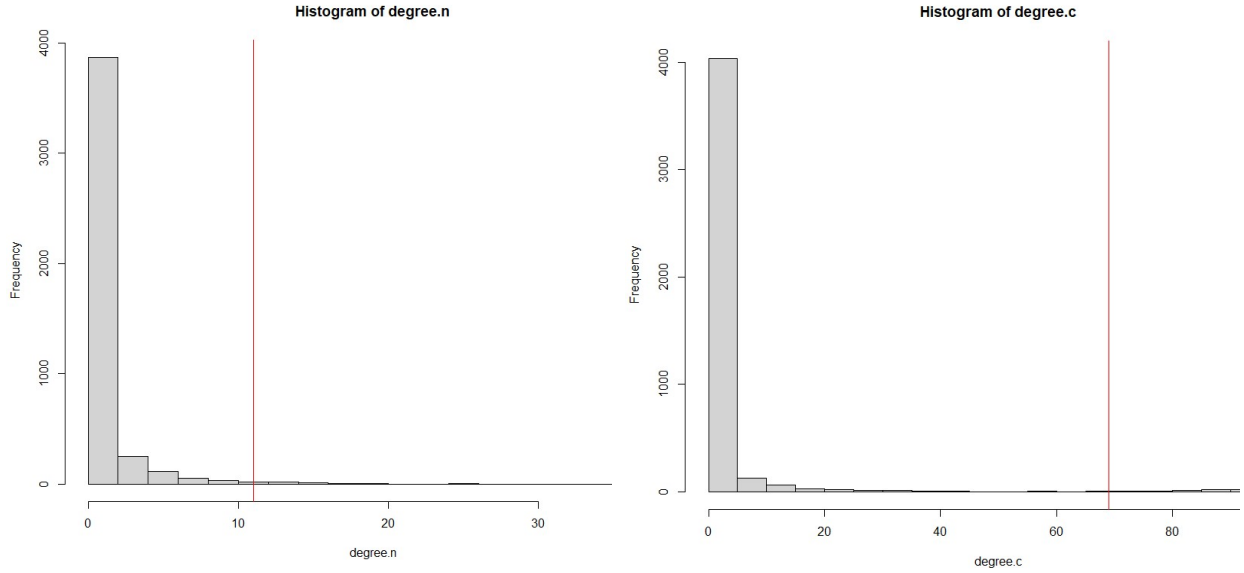


Figure 3 Co-expression networks

The identification of network hubs (nodes within the top 5% of degree values) revealed significant nodes in each network. In cancerous samples, nodes with a degree higher than 69, and in normal samples, nodes with a degree higher than 11 were categorized as hubs. The common hubs in both networks included genes such as AFM, PBLD, PRG4, DPYS, SLC2A2, and GSTA1.

The differential co-expression network analysis (Cancer vs. Normal) led to the construction of a binary adjacency matrix with aij=0 if |Z|<3. The histogram of the degrees follows a power-law distribution. Interestingly, only one gene, DPYS, emerged as a common hub in this comparison.



The hub subnetwork for cancer samples is illustrated in Figure 4, providing a focused view of the most interconnected genes within this condition.

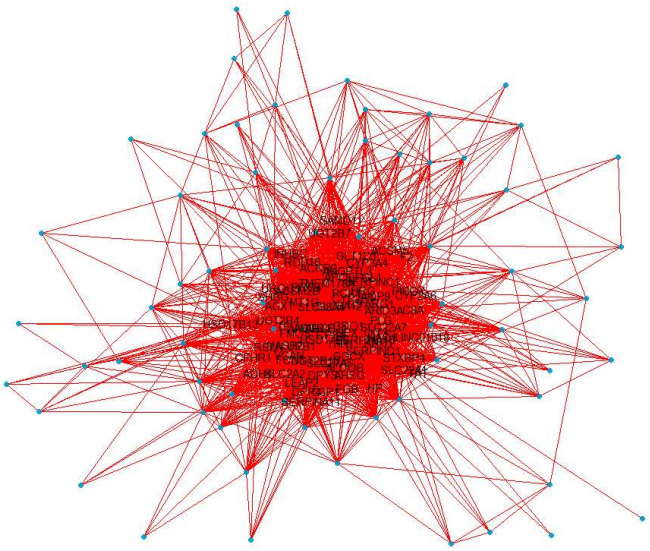


Figure 4 Hub subnetwork in Cancer Samples

Further, we constructed the Patient Similarity Network (PSN) using cancer gene expression profiles. Applying the Louvain algorithm, we detected 8 distinct communities within the network related to cancer samples, as shown in Figure 5.

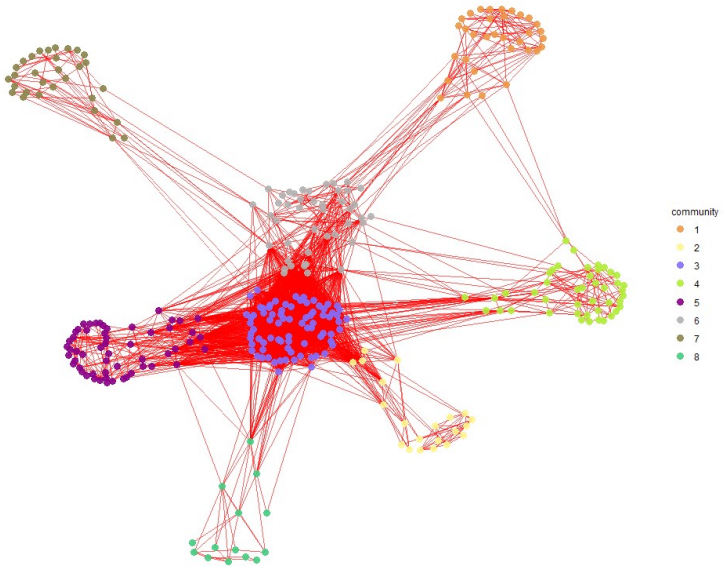
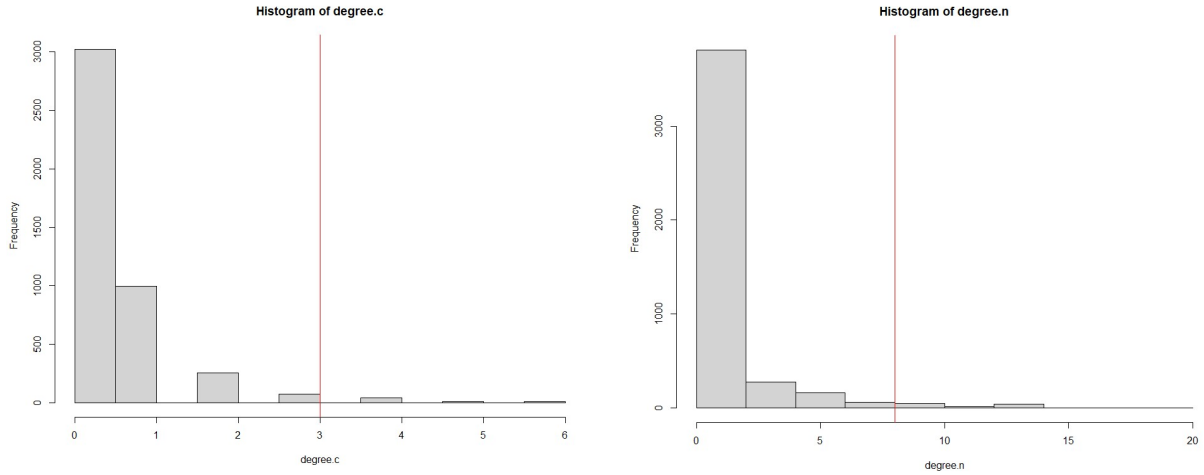


Figure 5 Communities related to the cancer samples

Bonus Parts:

1- We extract the nodes using another centrality index. (in this case we used betweenness CI) Then we labeled the top 5% of the nodes as hub and calculated the intersect of the hubs in the cancer and normal samples. The hubs are: SEC62, AFM, ALDH8A1,PODXL,DMGDH,INHBE,FMNL3, MEG3.(only the AFM is common with the previous ones)

2- We performed the study using a different similarity measure. (Spearman correlation) you can see the degree distributions in the plot below



Also, you can see the networks graph which are sparser than the previous ones that we computed with Pearson correlation.

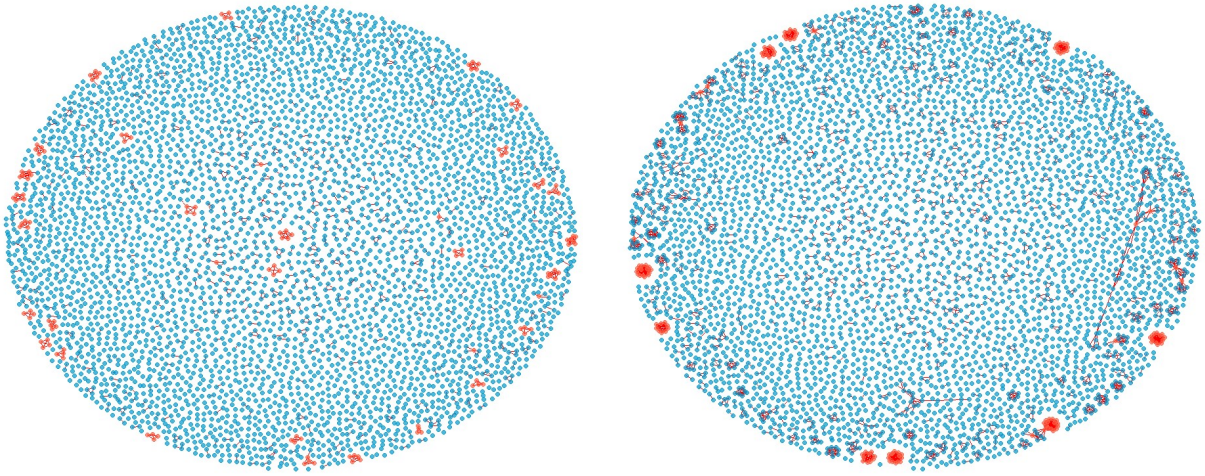
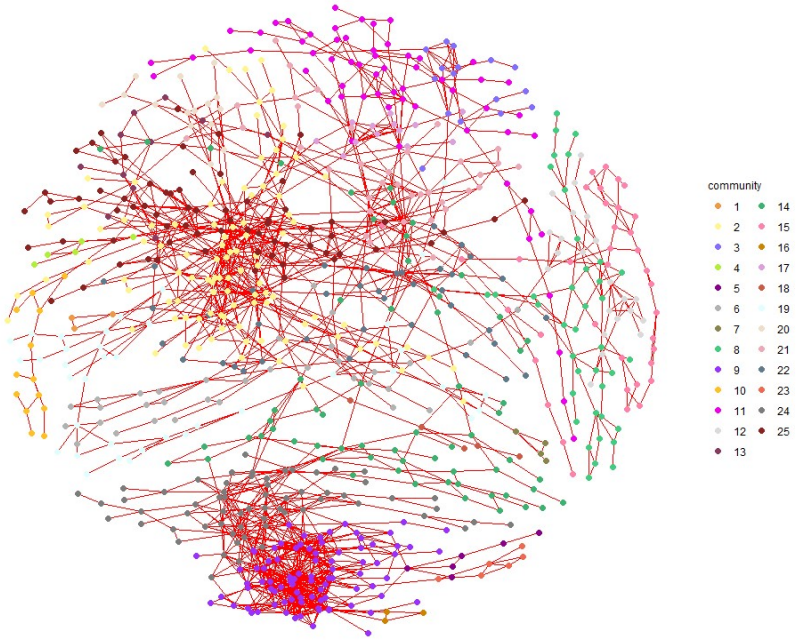


Figure 6 Cancer and normal samples (calculated with the Spearman Correlation)

3- Perform community detection using gene expression profiles related to normal condition You can see the plot of communities for normal samples. Which we have 25.



Discussion

The findings of this study offer a comprehensive view of the genomic landscape of Cholangiocarcinoma. The differential expression of 4,412 genes, with distinct sets being up-regulated and down-regulated, underscores the complexity of the disease's molecular profile. The identification of common hubs in both cancerous and normal samples suggests potential key regulators in the pathogenesis of Cholangiocarcinoma.

Conclusions and Future Directions

This study significantly contributes to the understanding of Cholangiocarcinoma at a molecular level. The identification of differentially expressed genes and the analysis of gene co-expression networks furnish valuable insights into the disease's pathology and potential therapeutic targets.