Differential Analyses of Gene Expression

Elisa Pierini, Valeria Popolla

GROUP 3

Abstract

This study focuses on Colon Adenocarcinoma, a subtype of CRC (Colorectal cancer), aiming to find distinctive gene expression patterns through advanced bioinformatics analyses of TCGA-COAD project data. Differential gene expression analysis identified 802 Differentially Expressed Genes (DEGs). Co-expression networks unveiled intricate relationships among these genes, highlighting hub genes such as TRAP1 and DIXDC1. Patient similarity networks identified distinct communities corresponding to less common adenocarcinoma types. Centrality measures, including degree and betweenness, provided insights into network topology. Enrichment analysis using hub genes revealed significant associations with cancer-related pathways. Using both Pearson and Spearman correlations, robustness in findings was demonstrated. Our integrated approach provides a comprehensive molecular perspective on Colon Adenocarcinoma, offering potential biomarkers and therapeutic targets for further exploration.

1 Introduction

Colon Adenocarcinoma, a complex malignancy, necessitates a comprehensive exploration to understand its molecular structure. This study investigates the landscape of gene expression in Colon Adenocarcinoma employing advanced bioinformatics methodologies and network analyses. The scientific issue centers around identifying Differentially Expressed Genes (DEGs) and highlighting co-expression patterns contributing to a better understanding of the disease. Existing literature highlights the significance of DEGs hubs analysis. Principles of network analysis and centrality measures underscore the importance of identifying key players in the molecular landscape. Our study hypothesizes that DEGs and co-expression networks play pivotal roles in delineating molecular characteristics specific to Colon Adenocarcinoma. By studying patient similarity networks and alternative similarity measures, we aim to unveil hidden structures that may elucidate distinct patient subgroups. The outcomes of this research bear relevance in biomarker discovery and therapeutic targeting for Colon Adenocarcinoma. Understanding DEGs, co-expression networks, and patient similarities opens avenues for personalized medicine. In the following sections, we want to offer a detailed journey through materials, methods, results and discussions about Colon Adenocarcinoma's molecular landscape.

2 Materials and methods

2.1 Data retrieving and cleaning

In order to proceed with the analysis, we firstly needed to retrieve data of the TCGA-COAD project from the GDC Data Portal being careful to specify the data category: Transcriptome Profiling, data type: Gene Expression Quantification, workflow type: STAR - Counts and selecting only patients for which both data types: "Primary Tumor" and "Solid Tissue Normal" are available. We also needed to check for duplicates in both patients set, delete them if any and keep only common patients between the sets which left us with 38 patients.

After having verified the absence of null values, we created a unique dataframe, called full.data, made of 60660 reads (genes) and both conditions for the same patients (so 76 columns). We used DESeq2, a library of R, to perform size factor normalization, the size factors are calculated to adjust for differences in library sizes across samples. The normalization process involves dividing the raw counts for each gene in each sample by a size factor. This scaling helps to make the data comparable across samples, accounting for variations in sequencing depth. The size factors are estimated by DESeq2 to represent the median ratio of observed counts to expected counts for each sample. This accounts for the fact that, in RNA-seq data, highly expressed genes may have more reads, but the actual biological signal is proportional to the ratio of counts for a given gene relative to the total expression in that sample. In the end we kept only genes expressed at least as many times as 10 counts leaving us with 11646 genes.

2.2 Differentially Expressed Genes (DEGs)

Since the aim of this study is performing a Differetial Analyses of Gene Expression, we computed the Fold Change, which is a measure of how much the gene expression changes between two conditions (in our case cancer vs normal) and for each value of the Fold Change (set to 1.6) related to each gene we computed the p-value and the adjusted p-value (set to 0.05) to obtain the set of 802 Differentially Expressed Genes (DEGs).

2.3 Co-expression networks

The following analysis will be performed on the DEGs of both conditions only, which we selected from the list of all genes related to the two conditions. DEGs will be used to compute the Co-expression networks.
We started by performing the log-transformation of our data using $log2(x+1)$, then we computed the Pearson correlation matrices for the DEGs related to the two conditions, cancer and normal, and saved the correlation coefficients matrices (setting also their diagonal to zero) and respective p-values matrices. We computed the adjacency matrices of the DEGs of the two conditions taking into account only the correlations between DEGs which have the p-value ≤ 0.01 (for the cancer DEGs) and p-value ≤ 0.001 (for normal DEGs) so those for which observed difference in gene expression between the two sample groups is statistically significant or could be due to chance. So in the graph, if two genes have p-value greater than 0.01 there won't be a connection between them.
After these computations, we derived the Cancer and Normal networks. For both of them, we computed the degree index needed to draw the degree distribution, which represents the probability distribution of the degrees of vertices of the graph. Given their scale-free feature, we were able to find the hubs of each network. To do so, we needed to isolate the 5% of nodes with the highest degree index which, in practice, means to compute the 95^{th} percentile of the degree distribution and keep only nodes for which the degree index was greater than the value of the percentile. Comparison between the hubs related to the two conditions were made to check for common elements.

2.4 Differential Co-expressed Network

In the next paragraph, we will compute the differential Co-expressed network using DEGs, the significance of their change in co-expression and Fisher's z-transformation.
We needed to compute the transformed correlation values for normal and cancer tissues where the transformation is applied to these correlation values using the formula for Fisher's z-transform, we calculated the sample sizes for normal and cancer tissues and the denominator for the Z-score calculation, which involved the square root of the inverse of the degrees of freedom for normal and cancer tissues (n1-3 and n2-3, respectively). Then, we computed the Z-score for each pair of genes based on the transformed correlation values, which represents the standardized difference between the co-expression patterns in normal and cancer tissues. Finally, we set Z-scores to 0 for pairs of genes where the absolute Z-score was less than 3. It was a thresholding step, suggesting that only those gene pairs with a significant difference in co-expression would be considered.
After these computations, we computed the degree index of the Differential Co-expressed network and derived the respective degree distribution to check if it was scale-free. Since the answer was affirmative, we computed the hubs of this network and compared them with both the hubs of the normal and cancer networks.

2.5 Patient similarity (PSN)

The following paragraph is about the construction of the patient similarity network based on Pearson correlations between gene expression profiles and the visualization of the resulting network. The resulting graph represents the relationships between patients based on the similarity of their gene expression patterns.
Basically, we firstly computed the Pearson correlation matrix using cancer gene expression profile, then we excluded self-correlations setting the diagonal elements to zero and finally, by creating an undirected, weighted graph from the correlation matrix, we plotted the Patient Similarity Network.
To perform the community detection, we applied the Louvain algorithm to the PSN obtained before which partitions a patient similarity network based on gene expression correlations into communities, grouping together patients with similar expression patterns, thereby revealing modular structures in the network.

2.6 Compute a different centrality index (CI)

Instead of the Degree Index, we computed two others centrality measures: the Betweenness Centrality with a procedure similar to the one applied for the Degree Index. We also compared the hubs obtained for this new measure with those found before regarding the Degree Index.

2.7 Perform the study using a different similarity measure

We decided to perform the whole study we have described in the sections above with a different similarity measure: Spearman instead of Pearson. Comparisons of the results of the two studies will be presented in the Result section below.

2.8 Perform gene set enrichment analysis

The hubs of the Cancer Network were here used as a gene set to perform the enrichment analysis which is a bioinformatics method used to assess whether a particular set of genes shows a statistically significant over-representation of specific biological terms, such as gene ontology (GO) terms. The goal is to identify the biological processes, molecular functions, or pathways that are significantly associated with a given set of genes. It is widely used in transcriptomics, and other 'omics' studies to uncover the underlying biological processes associated with experimental results. Finally, we displayed and plotted the enrichment results, including top terms for both GO and KEGG pathways, ordered by p-value.

2.9 Perform task 5 using gene expression profiles related to normal condition and compare the community structure of the two conditions

We just changed the list of gene expression profiles we used to construct the PSN in task 5 with those related to normal condition. We also applied the Louvain Algorithm on this new PSN and obtained the communities to compare with those obtained using gene expression profiles of cancer network.

3 Results and discussion

Once we have retrieved and pre-processed data about the Colon Adenocarcinoma, we proceeded with identifying the Differntially Expressed Genes. As mentioned before, to do that, we set the Fold Change to 1.6 and the p-value to 0.05. Analyzing the data, we observe the presence of 367 genes with an FC score of -1.6 and a p-value of 0.05. Additionally, 435 genes are identified with an FC score of 1.6, also having a p-value of 0.05. We obtained 802 DEGs by selecting only the over and under expressed genes. The volcano plot in the image below represents with red points the under-expressed genes and with blue points the over-expressed genes. They both are our DEGs of interest.

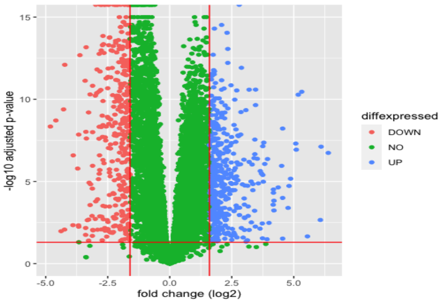

Figure 1: Volcano Plot

From this point on, we will consider only the DEGs of both conditions to compute the Co-expression networks using the Pearson's correlation as a similarity measure. We also applied the log-transformation to the DEGs, calculated the correlation and adjacency matrices in order to be able to compute and plot the Cancer Network and the Normal Network. Both networks composed of 802 nodes, have different densities: the cancer networks has 0.05169349, while the network with normal condition has 0.07028621. We can observe that the density in the cancer network is higher than the normal one, which means that in the first one there are more connections between genes. In the plots below, we show the degree distributions of both the networks.

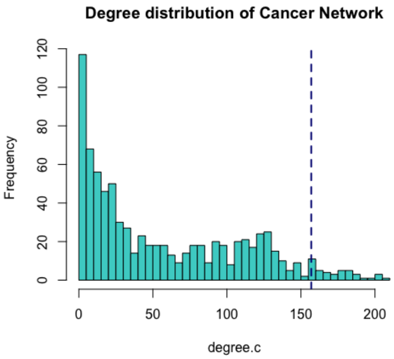

Figure 2: Cancer Network

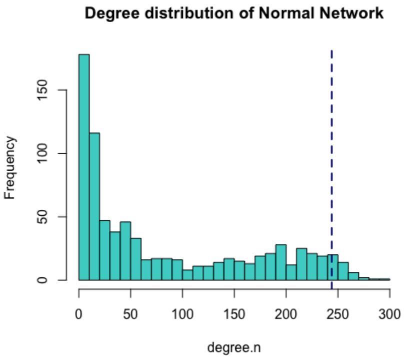

Figure 3: Normal Network

We can see that their shape seems to be scale-free which means we can also compute the hubs, which, as we said before, are the 5% of nodes with highest degree index, and analyze the meaning of some of them in our context. Two of the genes identified as hubs in the Cancer Network, TRAP1 and DIXDC1, are evidence that they are studied for the Colon Adenocarcinoma ([3], [1]). The novel marker tumor necrosis factor receptor-associated protein 1 (TRAP1) is a mitochondrial heat shock protein that has been related to drug resistance and protection from apoptosis in colorectal cancer. TRAP1 is largely studied in our context since it may be a key regulator of the CRC cell response to hypoxia ([4]). Since we also found a common hub between cancer and normal networks, we investigated its features. The biological role of DIXDC1 in the formation and progression of human colon cancer remains largely unknown. This might explain why it can be found in both networks. In the study which reference is reported below, it is provided the evidence that DIXDC1 overexpression could promote colon cancer cell proliferation both in vitro and in vivo through facilitated G1/S phase transition, which is a very common phenotype in cancer cells ([2]).
Using the procedure already described in the paragraphs above, we computed the Differential Co-expression network and measured its degree index which distribution turned out to be scale-free as shown in the figure below. In this context, we computed and plotted the sub-network of the hubs of the Differential Co-expression network.

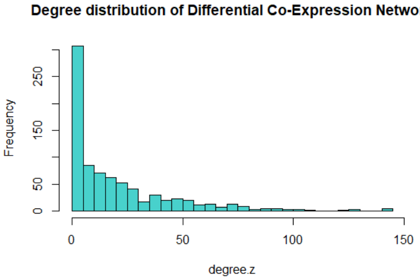

Figure 4: Differential Network

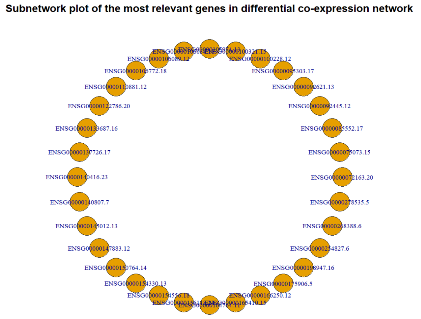

Figure 5: Subnetwork

We also computed the hubs of this network. One of them is PHGDH which study found that both PHGDH mRNA and protein was highly expressed in tumor tissues in comparison with matched adjacent non-tumor tissues ([5]). We compared the hubs of the Diffrential Co-expression Network with the ones of the Co-expression networks and we found 7 hubs between the Cancer Network and the Differential one and 2 hubs between the Normal Network and Differential one.

We computed the Patient Similarity Network where nodes represent patients and edges represent the similarity between patients calculated using their gene expression patterns. Starting from it, we performed community detection, which procedure and aim was described before, and we found two groups of patients which correspond to the two less common types of Colon Adenocarcinoma: Mucinous adenocarcinoma and Signet ring cell adenocarcinoma ([6]). In the plot below, the communities are highlighted in different colors.

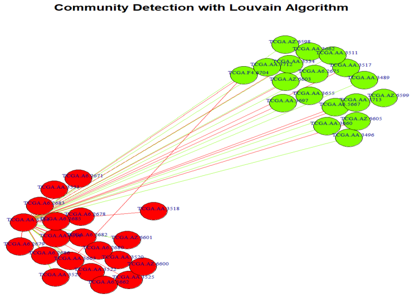

Figure 6: Communities of patients

As mentioned before, we also computed the Betweenness as centrality measure for the Cancer Network to compare the results (in particular the hubs) with the ones obtained using the Degree Index. Three hubs turned out to be common which strengthen our findings.

In the sections above, the analysis was performed using the Pearson's correlation as a similarity measure. We wanted to see what would have changed if we have used Spearman's correlation as a similarity measure. The common hub, in this case, is COL8A1 which was identified and proved to be correlated with the progression and prognosis of human colon adenocarcinoma ([7]). Louvain algorithm showed two communities, as before, but as we can see from the picture below, they are grouped differently, one of the two communities is greater than the other.

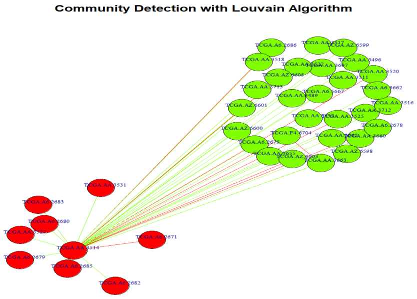

Figure 7: Communities of patients (Spearman)

We proceeded with the Enrichment Analysis, which aim was explained in the sections above, and the results are shown in the plots below. Terms that are significantly enriched (usually those with lower p-values) represent biological processes or pathways associated with the input gene set (which is the hubs set of the Cancer Network). The plots help quickly identify the most relevant and enriched terms, providing insights into the potential functions or pathways related to the genes of interest.

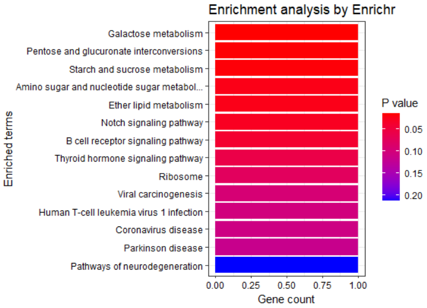

Figure 8: Enrichment on GO dbs

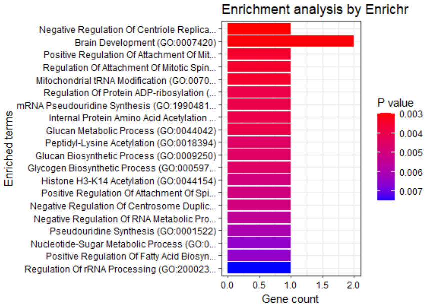

Figure 9: Enrichment on KEGG dbs

Finally, we wanted to compare the communities of the Cancer Network with the ones of the Normal Network. The plot below represents the result of the community detection performed with the Louvain algorithm on the Normal Network and differs from the one about the Cancer network because of the different division of the communities.

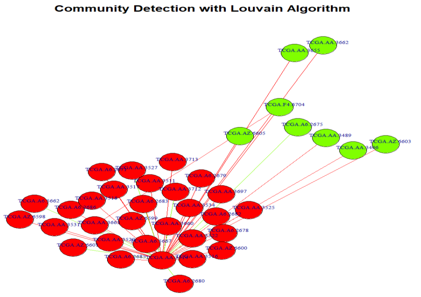

Figure 10: Communities of patients (Normal Network)

References

[1] Evidence of DIXDC1 (ENSG00000150764) in colon cancer. Link: https://pubmed.ncbi.nlm.nih.gov/19572978/

[2] Wang L, Cao XX, Chen Q, Zhu TF, Zhu HG, Zheng L. DIXDC1 targets p21 and cyclin D1 via PI3 K pathway activation to promote colon cancer cell proliferation. Cancer Sci. 2009;100(10):1801–8. Link: https://pubmed.ncbi.nlm.nih.gov/19572978/

[3] Clinicopathologic significance of TRAP1 expression in colorectal cancer: a large scale study of human colorectal adenocarcinoma tissues. Link: https://diagnosticpathology.biomedcentral.com/articles/10.1186/s13000-017-0598-3

[4] TRAP1 regulates the response of colorectal cancer cells to hypoxia and inhibits ribosome biogenesis under conditions of oxygen deprivation. Link: https://pubmed.ncbi.nlm.nih.gov/35543151/

[5] Increased Expression of PHGDH and Prognostic Significance in Colorectal Cancer. Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4907894/

[6] Colorectal cancer types. Link: https://www.cancercenter.com/cancer-types/colorectal-cancer/types

[7] Co-expression Network Analysis Identified COL8A1 Is Associated with the Progression and Prognosis in Human Colon Adenocarcinoma. Link: https://pubmed.ncbi.nlm.nih.gov/29497907/