DIGITAL EPIDEMIOLOGY AND PRECISION MEDICINE

Project - 1

Differential Analyses of Gene Expression

A. Alessia Ciarlo, B. Alessandro Garbetta
Group 02

December 21, 2022

Abstract

Colorectal cancer (CRC) is one of the leading causes of cancer-related mortality in the world, in which colon adenocarcinoma (COAD) is the most common histological subtype of CRC. In this study, our aim is to identify gene modules and representative candidate biomarkers for clinical prognosis of patients with COAD, and help to predict prognosis and reveal the mechanisms of cancer progression. Gene co-expression network analysis and Differential co-expression analysis was performed to construct some networks and identify gene modules correlated in COAD patients. A total of top 38 hub genes was identified, in which 25 of the hub genes show a significant up/down-regulation in COAD as compared to normal tissue, including TNS1, CCDC80, DNAJB5, DCLK1, SPARCL1, FHL1, JAM2, MGP, and CAVIN2.

1 Introduction

TA gene is the basic physical and functional unit of heredity of living organisms' cells.
The gene expression is the process by which information from each gene is used in the synthesis of a functional gene product (e.g. proteins). Although gene expression is a highly regulated process (by hormones and biological mechanisms), genes in a cell can become abnormal, due to gene mutations, causing the cell growth and its division out of control (genetic disease, meaning uncontrolled growth). This leads to the presence of cancer.
We use Gene expression to provide a qualitative assessment of the presence of transcripts (mRNA molecules) or of proteins encoded by that gene. Knowing which genes are expressed in a cell allows the comparison of expression profiles between samples.
Differential gene expression (DGE) analysis is one of the most common applications of RNA-sequencing (RNA-seq) data. This process allows for the elucidation of differentially expressed genes across two or more conditions (e.g. normal-cancer) and is widely used in many applications of RNA-seq data analysis.
By sequencing the DNA e RNA of cancer cells and comparing the sequences to normal tissue, scientists can identify genetic differences that may cause cancer. In this way they can also measure the activity of genes encoded in our DNA to understand which proteins are abnormally active or silenced in cancer cells.

The goal/aim of this study is to:

- Identify Differentially expressed genes that can be candidate biomarkers for clinical prognosis of patients with COAD;

- Build Co-expression networks to find gene hubs related to Normal and Cancer conditions;
- Identify Patient Similarity Network to find the patients' clusters that share some characteristics in the gene expression profile.

In the next paragraphs we will talk about the main results of scientific literature and the steps of our study starting from the data acquisition and the methods used to achieve the goals. Then we will explore some (optional) experimental steps to make a comparison between the results obtained and, at the end, we will discuss the results showing some interesting images.

1.1 Main results of scientific literature

Thanks to lots of scientific researches, quite a number of genes have been found to play important roles in Adenocarcinoma Of Colon, such as BCL2, CASP3, CDKN1A, CEACAM5, CEACAM7, TNS1, EGFR, IFNG, PDZD4, IL2, KRT20, KRT7, MAPK1, ADAMTSL3, PSG2, PTGS2, TNF, TP53, TYMS, VEGFA, VIP. These and other hub genes were identified in many scientific papers (some References at the end, referred also to the relationship between our results and scientific literature) and they show a significant down/up-regulation in COAD as compared to normal tissue.

2 Materials and methods

In this section we will describe the resources, the experimental procedures, data analysis procedures and statistical methods used for the study.
We have used R programming language to simplify all the statistical computations with some libraries also for visual support techniques to visualize the results of our analysis like graphs, networks, hinstograms etc.

2.1 Data

We use gene expression data of Colon Adenocarcinoma cancer cells from the National Cancer Institute GDC Cancer Portal. We are interested in the Gene Expression Quantification of the Transcriptome Profiling in STAR - Counts file format.
From the data, we extract only patients with clinical data and for whom cancer and normal tissue files are available and we do a data pre-processing to remove all the genes where at least one value is zero. So we obtain the value of 15488 genes for 38 patients.

2.2 Differentially Expressed Genes (DEGs)

A gene is Differentially Expressed if an observed difference or change in expression levels between two experimental conditions (normal/cancer) is statistically significant. To identify this kind of genes it is important to find statistical distributional properties of the data.
For this first task we have used the DESeq2 package that provides methods to perform accurate DGE analysis. Giving our data (slightly preprocessed through a merge between the normal and cancer matrix) to this model, we have obtained a DESeqDataSet object and a result structure containing lots of useful informations from the comparison between normal and cancer genes: baseMean, log2FoldChange, lfcSE, stat, pvalue, padj. The choice of this model lies in the fact that its output is complete and also it internally normalizes the count data.
Then we have identified DEGs using a threshold of 1.2 and we have obtained a subset of 1200 genes.

P-value threshold was setted to be less than or equal to 0.05 and Fold Change (FC) threshold: $|FC| \geq 1.2$[1].

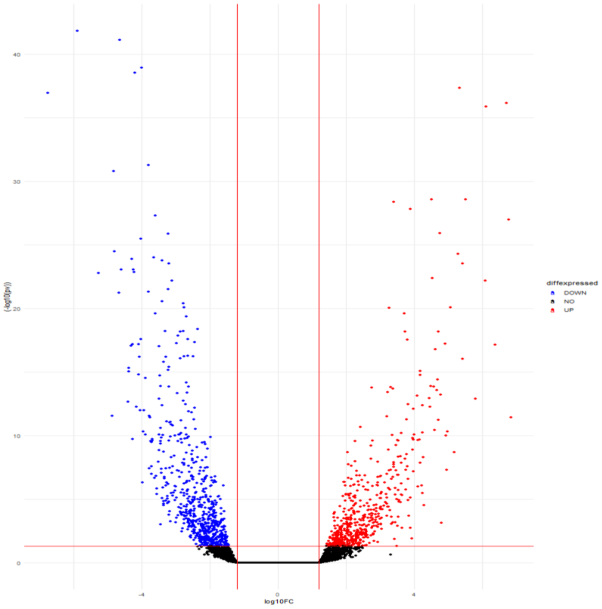

Figure 1: Volcano plot

2.3 Co-expression networks

Gene co-expression network analysis is a systematic biological method that delineates correlations between genes and clinical traits. It identifies highly correlated genes to investigate potential biological functions.
Using only DEGs found in the previous part, we have computed the gene co-expression networks related to the 2 conditions (cancer, normal) using:

- Pearson's correlation;
- Binary adjacency matrix where $a_{ij} = 0$ if $|\rho| \geq 0.7$

Starting from the density, the largest component and the degree index computation, we have extracted all the information written in Table 1 for both cancer and normal networks.

	nodes	density	nodes largest component	un-connected nodes	nodes hub
Network cancer	1200	0.015675	709	435	38
Network normal	1200	0.051705	1008	177	52

Table 1: Summary information on Cancer/Normal Network

The result networks are scale free (see Figure 2, Figure 3) and this is fundamental to find the hubs, so the 5% of the nodes with highest degree values.

[1] Using $|FC| \geq 1.2$ and a p-value ≤ 0.05 we obtain a more scale free networks; with greater treshold value the number of node with intermediate degree is greater, so the net is not totally scale free although the numer of genes is about hundreds; with lower treshold, instead, the network is more scale free, but the number of nodes is much greater (¿2000)

In cancer network we have found 38 hubs, corresponding to the most connected genes, that are: TNS1, AMOTL1, DNAAF9, MEIS1, C14orf132, AOC3, MSRB3, CCDC80, NEGR1, RNF150, DNAJB5, DCLK1, ANK2, CLIP3, SPARCL1, DACT3, FHL1, JAM2, SALL2, ADAMTSL3, TP73-AS1, ARMCX1, MYLK, SHISAL1, DDR2, MGP, SPART, CRYAB, LRCH2, HSPB8, CLMP, PRELP, MAB21L2, CAVIN2, PDZD4, SYNPO2, GNAO1, LMOD1.

In normal network we have found, instead, 52 hubs.

Comparing the hubs related to the two condition (cancer, normal) we can see that the genes characterizing only cancer tissue are: TNS1, C14orf132, AOC3, MSRB3, CCDC80, RNF150, DNAJB5, DCLK1, CLIP3, SPARCL1, DACT3, FHL1, JAM2, SALL2, ARMCX1, MYLK, SHISAL1, MGP, CRYAB, HSPB8, PRELP, MAB21L2, CAVIN2, SYNPO2, LMOD1.

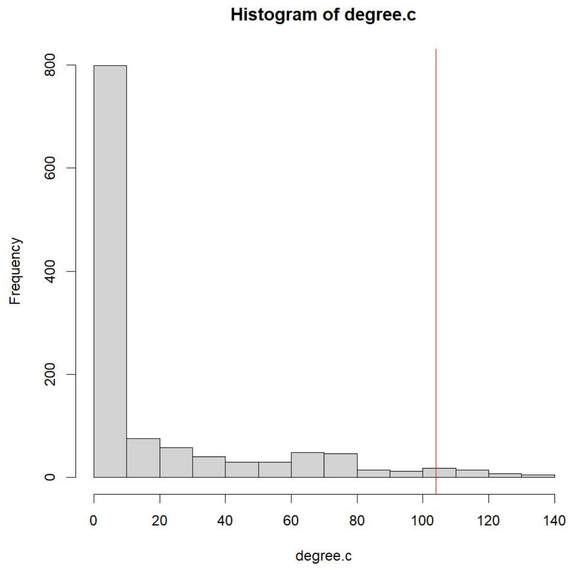

Figure 2: Histogram of Cancer Network

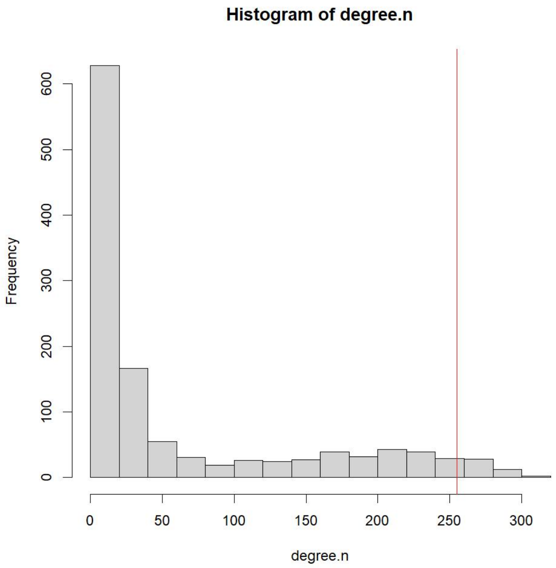

Figure 3: Histogram of Normal Network

2.4 Differential Co-expressed Network

Differential co-expression (DCE) analysis helps to look into conditio specific changes of co-expression networks. In other words, if a set of co-expressed genes behave in a certain way or respond in a particular fashion to biological changes, it is termed as differentially co-expressed. Using only DEGs, we have computed the differential co-expression network comparing the 2 previous condition neworks (Cancer vs. Normal). For doing this, we have applied Fisher transformation on the node correlation matrices and then, using the results, we have computed the Z_{score}, where z_1 and z_2 are the Fisher Transformation matrices and n_1 and n_2 are the sample size for each of the conditions.

$$Z_{score} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

Then binary adjacency matrix is built with aij=0 if $|FC| < 0.9$. With this value of Z we have found the degree index and tested that the network is scale free (Figure 4).

The 32 hubs obtained here are: TNFRSF12A, TMPRSS3, SALL4, LINC02983, ELFN1-AS1, MILIP, PRELID3A, CBX4, CCDC78, CXCL3, CITED2, GFRA2, KLK15, IGHV3-53, CDC25B, NMU, CRYAB, EIF2S2P4, TRIM29, PGGHG, LMTK3, PID1, GRHL3, PDPN, FOLR2, TNXB, MYEOV, UGT2B15, SLC28A3, SNORD17, ID2-AS1, RPL23AP65.

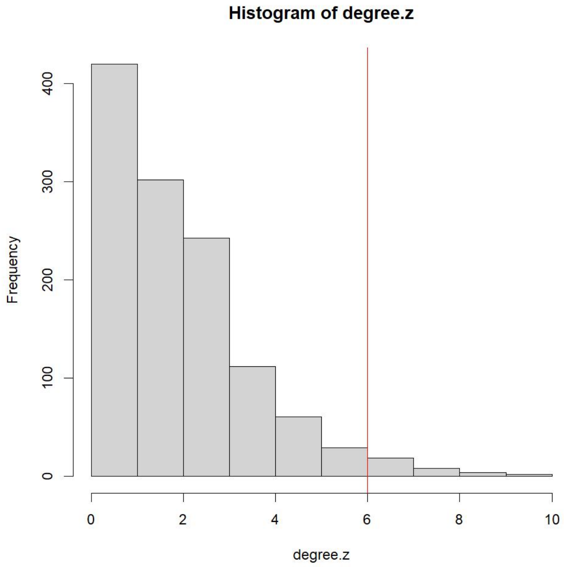

Figure 4: Histogram of Differential Co-Expressed Network

2.5 Patient Similarity Network (PSN)

In a patient similarity network, each node is an individual patient and an edge between two patients corresponds to pairwise similarity for a given feature. Each input patient data feature (e.g. age, sex, mutation status - in our case gene expression) is represented as a network of pairwise patient similarities.

Each feature is represented as a different “view” of patient similarity that can be integrated with all the other views to identify patient subgroups or predict outcome.

First of all, we have computed the Patient Similarity Network using cancer gene expression profile through cosine similarity and pearson correlation. The 2 resulting networks are totally ugual.

Then we have applied the Louvain algorithm to the PSN to perform the community detection based on gene expression profile. The Louvain method is useful to extract communities from large networks and it’s a greedy optimization method that appears to run in time $O(n \cdot \log n)$ where n is the number of nodes in the network. It maximizes a modularity score for each community and it quantifies the quality of an assignment of nodes to communities.

As result we have found 5 clusters that are shown in Figure 5.

2.6 OPTIONAL TASKS - other experimental trials

(1) We have performed all the study using also Spearman as similarity measure instead of Pearson. The results obtained were almost identical both as network structures, as hubs and clusters of patients in PSN.

(2) In PSN we have tried to make Community Detection also using gene expression profiles related to normal condition. As before, we have used both Cosine and Pearson for similarity measure and, in

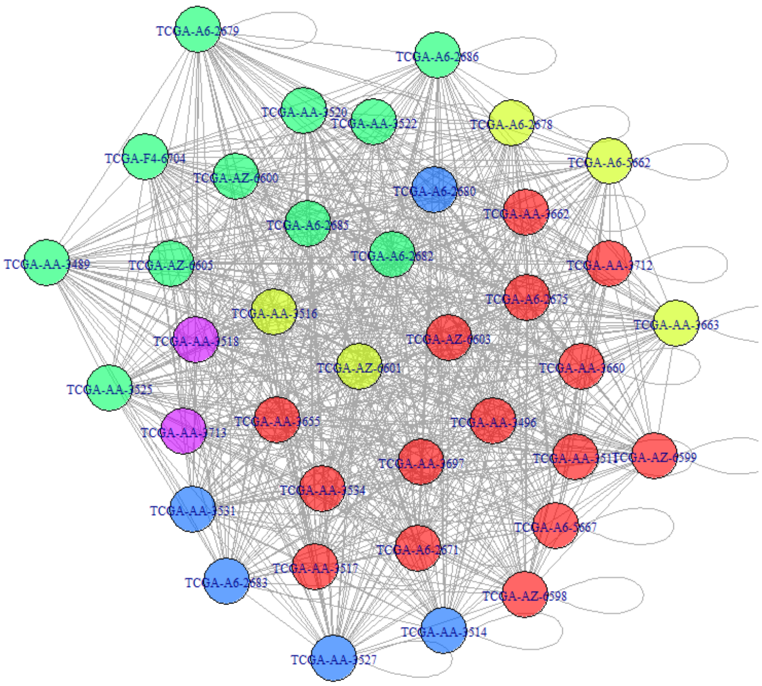

Figure 5: PSN Cancer Graph

this case, we have noticed a difference in the number of Pearson in the extracted communities. The number of communities found is 11, as shown in Figure 17 and 18.

	nodes	density	nodes largest component	un-connected nodes	nodes hub
Network cancer	1200	0.01236725	671	455	39
Network normal	1200	0.03116486	937	234	49

Table 2: Summary information on Cancer/Normal Network with Spearman

3 RESULTS AND DISCUSSION

All the obtained results, described above, are shown in the attached images. List of figures linked to this report:

- Figure 1a: log2FC
- Figure 6: Co-expressed cancer network
- Figure 7: Subnet of Co-exp cancer network
- Figure 8: Co-expressed normal network
- Figure 9: Subnet of Co-exp normal network
- Figure 10: Differencial Co-expressed network
- Figure 11: Subnet of Diff-coexp network
- Figure 12: Matrix PSN Cancer Condition

- Figure 13: Graph PSN Cancer
- Figure 14: Cluster PSN Cancer
- Figure 15: Matrix PSN Normal Condition
- Figure 16: Graph PSN Normal
- Figure 17: Cluster PSN Normal with Cosine similiraty
- Figure 18: Cluster PSN Normal with Pearson correlation
- Figure 19: Co-expressed cancer network SPEARMAN
- Figure 20: Subnet of Co-exp cancer network SPEARMAN

Seeing our results we found a strong similarity, especially for some genes, with other scientific papers concerning colon and bowel cancers.

Working with this type of data was not so easy because, choosing (through a threshold value) a subset of a few hundred genes, led to a faster computation time, but with non-scale-free networks; While, working with many genes, the networks were perfectly scale-free, but the computation took a lot of time. For these reasons we have come to the compromise of working with a subset of 1200 genes.

Similar speech also for the part of the Differential Co-Expessed network. Here we have chosen the threshold of Z score equal to 0.9 because, for greater value, the network was not scale free, but it was more similar to a small-world network (there was an increasing first part and, after a maximum value in the middle part, a decreasing one).

Relations with our results and other pubblications:

- TNS1: Elevated transgelin/TNS1 expression is a potential biomarker in human colorectal cancer (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5787423/)
- C14orf132: Down-regulated C4orf19 confers poor prognosis in colon adenocarcinoma identified by gene co-expression network (https://pubmed.ncbi.nlm.nih.gov/35281862/)
- CCDC80: Dro1/Ccdc80 inactivation promotes AOM/DSS-induced colorectal carcinogenesis and aggravates colitis by DSS in mice (https://pubmed.ncbi.nlm.nih.gov/29901779/: :text=Downregulated)
- CCDC80: DRO1/CCDC80: a Novel Tumor Suppressor of Colorectal Carcinogenesis (https://link.springer.com/article/10.1007/s11888-015-0276-3)