ANALYSIS AND REGULATION OF GENE EXPRESSION IN THYROID CANCER

Alfano Caterina 1746299
Cappelli Dario 1711562

November 2020

Contents

1 Abstract

This project analyzes gene expression in thyroid cancer while looking for the genes most affected by the carcinoma. To do this we compared 58 sample of healthy people with those of 508 patients with thyroid cancer, using co-expression gene networks and graph theory. Our results showed a significant increase of genes' correlation when the carcinoma is present; in particular the increase is due to a subset of genes that alter their number of correlations of even more than a hundred. Lastly, we did some research on how to focus on target genes for co-expression normalization practices through miRNAs, and plotted the amount of differences from the healthy condition against the number of normalized genes, while also highlighting the importance of ERBB3.

2 Introduction

Thyroid cancer is a rare type of cancer that affects the thyroid gland, a small gland at the base of the neck that produces hormones; it is diagnosed 1 % in the proportion of all tumors detected by clinical in developing countries. It is suggested that the endocrine system has the highest cancerization risk in the thyroid with an increased tendency each year. Thyroid cancer has multiple subtypes, including undifferentiated, medullary, follicular, and papillary thyroid cancer.[1]
The aim of this project is to analyze the changes in gene expression caused by thyroid cancer and identify the most affected genes. Given samples of gene expression over many healthy and ill patients, we looked for differentially expressed genes and used co-expression and differential co-expression gene networks to infer the main changes and alterations in the genes' behaviour. Running these tests we were able to notice that differentially expressed genes in thyroid cancer have a tendency to be more up-regulated than down-regulated and that the co-expression values among genes increases significantly. We highlighted the five genes that are most affected by this cancer and, for some of them, found scientific papers that were studying their relation to thyroid cancer.
Finally, we decided to try looking for the combination of five genes whose change in expression could be most beneficial in a patient, given that the others genes stay in their altered cancerous state. Doing this we spotted some trends and information that could be useful in future researches.

3 Materials and methods

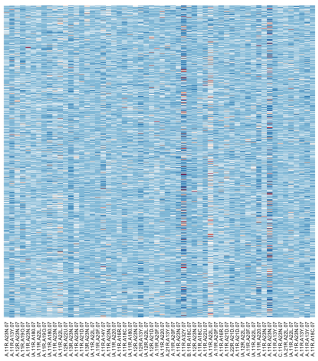
3.1 Data

The cancer we are focusing on is thyroid carcinoma (TCGA-THCA); using the script provided to us, we were able to download the tables containing the observations on which our project is based. For each of the two conditions we had samples for 56602 genes for many patients (502 ill ones and 58 healthy ones). Of course before starting our analysis we had to clean our data, removing rows with some null values and checking for NaNs. Finally, we just considered the genes that remained in each of the tables (i.e. that had complete information in both ill and healthy patients); we were thus left with 15965 genes among which to look for the differentially expressed ones.
Our project was carried out with R programming language and a variety of libraries (to plot, create networks. etc.)

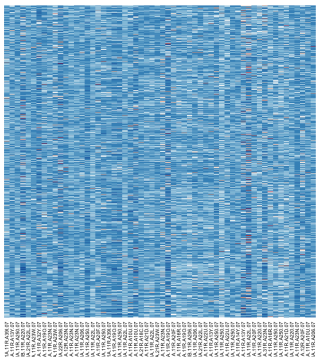
3.2 Differentially Expressed Genes (DEGs)

Differential gene expression (DGE) analysis requires that gene expression values be compared between sample group types, to find out which genes' behaviours were altered in the ill patients. To carry out this analysis we used the fold change value, which measures for each gene the ratio between the mean expression in the normal condition and the mean expression in the cancerous condition (in a log2 scale). Keeping in mind that not all our results are meaningful we performed the student test among samples of the same gene in the two conditions to obtain its p-value. Finally, we adjusted the p-values with a false discovery rate correction as a way to control the expected proportion of false "discoveries" obtained by a statistical test that performs multiple comparison. Applying a threshold of 1.2 and 0.05 on the absolute values of FC and p-values we were able to identify the 1094 DE-genes. Here you can visualize this procedure through a volcano plot that shows the up regulated genes in blue and the down regulated genes in red; meanwhile the horizontal line represents the p-value threshold.

Moreover we can visualize the expression of the genes in both conditions with a heatmap:



(a) healthy gene expr



(b) cancerous gene expr

3.3 Co-expression networks

A gene co-expression network (GCN) is an undirected graph, where each node corresponds to a gene, and a pair of nodes is connected with an edge if there is a significant co-expression relationship between them. It is constructed by looking for pairs of genes which show a similar expression pattern across samples. Starting from the expression data shown above, we were able to compute the Pearson correlation coefficient to state the correlation rate of two genes. With this information we were able to construct 2 networks (one for each condition) in which two nodes/genes are connected with an edge if their Pearson coefficient is greater than 0.7 (common threshold for strong correlations).

The two graphs are quite different, not in the overall structure (even though the normal graph seems too have more high density regions), but in the sense that the most 5% connected nodes are almost completely different: only 7 elements are hubs in both networks. In fact, a lot of genes alter their degree (n° of genes they are correlated with) by big factors. More precisely, 48 genes change their degree of at least 100. We selected them and reported the ones who went through the biggest changes in the table below.
To obtain the genes' names from the ensembl ID we used the Biomart library to extract the information from the ensembl human database.

Ensemble ID	Gene Name	Degree Can	Degree Norm	Difference
ENSG00000131435	PDLIM4	187	4	183
ENSG00000137648	TMPRSS4	166	2	164
ENSG00000173227	SYT12	162	3	159
ENSG00000124145	SDC4	168	11	157
ENSG00000132334	PTPRE	185	28	157
ENSG00000171812	COL8A2	157	0	157
ENSG00000115414	FN1	157	6	151
ENSG00000184156	KCNQ3	148	0	148
ENSG00000197249	SERPINA1	179	32	147
ENSG00000184292	TACSTD2	151	10	141

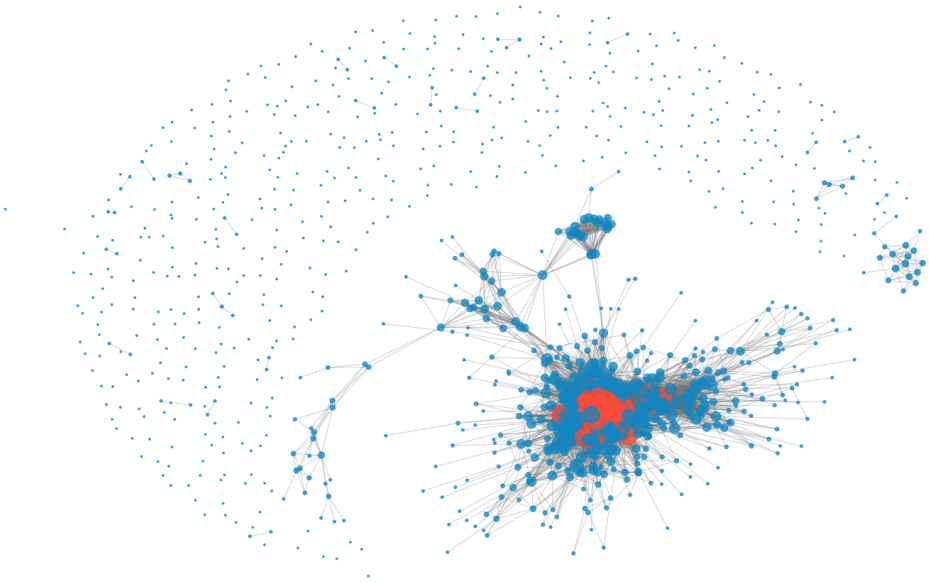


Figure 1: Graph of cancerous co-expression with highlighted hubs

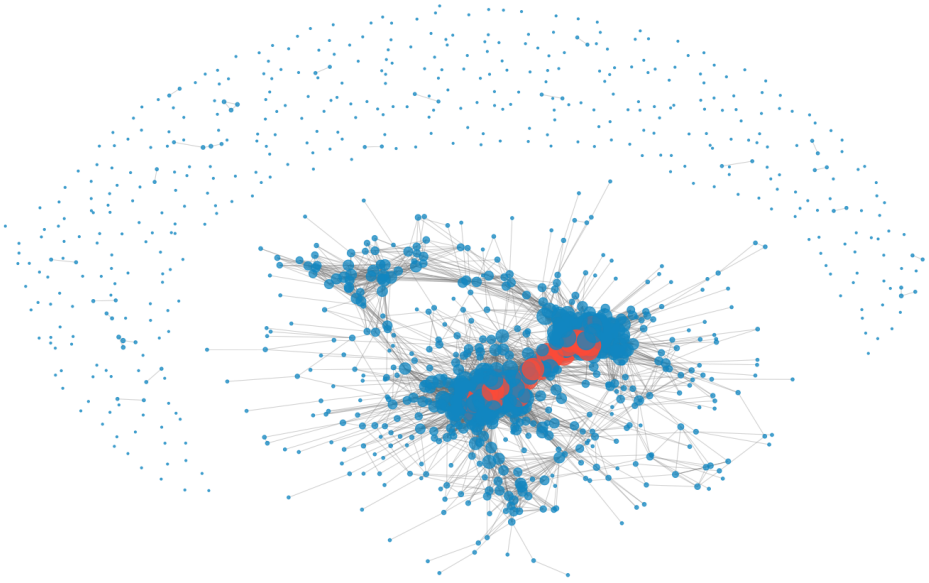


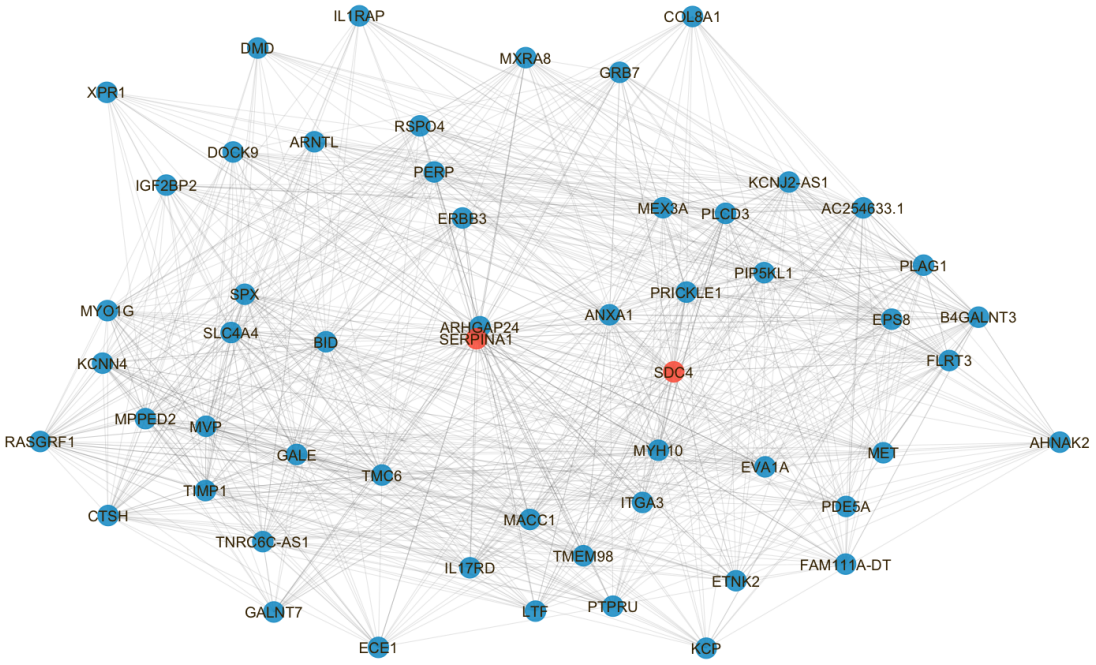
Figure 2: Graph of normal co-expression with highlighted hubs

3.4 Differential Co-expressed Network

In a normal co-expression network nodes are genes and they are connected if they have a significant pairwise expression profile association. As we already showed their co-expression changes in the two conditions; so, instead of establishing that the co-expression is significant in one condition and not in the other, we can test directly if the change in co-expression is significant. This is done with the differential co-expression network: if two genes are connected it means that their co-expression changed significantly from one condition to the other. This allows us to find the nodes who had more changes in the cancerous samples, because they will be the ones with more edges in this differential graph.
To obtain it we started from the Pearson correlation matrices and applied the Fisher z-transformation to stabilize the variance of sample correlation coefficients in each condition and allow the resulting values to be compared for statistical significance through the z-score method. In fact, once each z value has been determined, statistical significance can be assessed (and encoded with an edge in the network) by checking if the observed value is greater than the critical value; we chose 7 as critical value to avoid having a graph that was too connected.

Analyzing the 5% most connected nodes of this network we saw that their average degree is 83, but with an uneven distribution, since the degree of the biggest and smallest hubs differ of 117. We also checked if the 10 genes we focused on before (table on page 2), because of their sudden changes in co-expression, are also hubs of this graph. Only two of them are are: SDC4 and SERPINA1. This shows how important the transformation to the Pearson correlation and the z-scores are to infer meaningful results.

In the following figure we showed a sub-network of the differential co-expression network representing its hub and highlighting our two genes from the previous table.



After this analysis we focused on a new set of genes: the top 5 most connected nodes inside this network, which, given their importance in the graph (computed ranking all nodes' eigenvector centralities), we believe to have a critical role in this cancer.

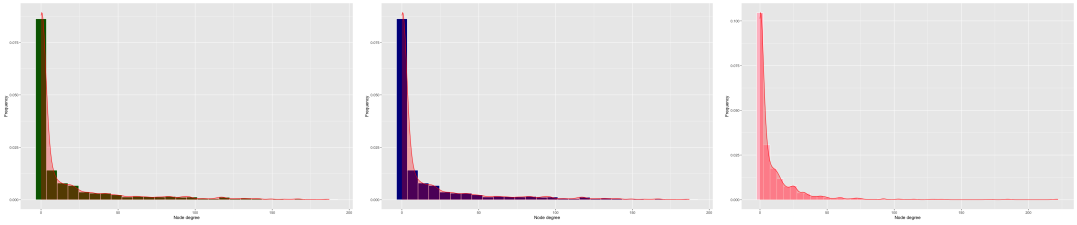
Ensemble ID	Gene Name	Degree	Importance
ENSG00000005884	ITGA3	184	3
ENSG00000065361	ERBB3	222	2
ENSG00000133026	MYH10	171	4
ENSG00000138735	PDE5A	220	1
ENSG00000143845	ETNK2	185	5

4 Results and discussion

4.1 Final results and interpretation

The reason why it's common to study the nodes' degree in the network is because genes that have a strongly altered connectivity are thought to play an important role in the disease phenotype. To better visualize the degrees of the nodes we can study the degree distribution through an histogram. In particular, we wanted to check that all of our networks were scale-free, i.e. they followed a power law distribution. For a network this means that there is a small number of highly connected nodes (the hubs or tail of the distribution) and many poorly connected ones.

We can see from the histograms that the networks are indeed scale free, but to be sure we also used the 'fit-power-law' function from the igraph library. This function compares our distribution to a power law one with the Kolmogorov–Smirnov test; the resulting KS.stat value denotes better fits with smaller values (in a range from 0 to 1). The KS.stat values for the normal network, cancer network and differential network were respectively: 0.08, 0.11, 0.04.



(a) Degree distr. - Normal (b) Degree distr. - Cancer (c) Degree distr. - Differential

Besides the degree distribution, the other main thing to study in such graphs are the hubs, i.e. the most connected nodes. Generally speaking, highly connected subgraphs in gene co-expression networks correspond to clusters of genes that have a similar function or are involved in a common biological process. Hubs often have a key role inside a highly connected subgraph because of their high degree (their importance can be quantified with centrality measures). Checking their behaviour is really important, since genes that have a strongly altered connectivity are thought to play an important role in the disease phenotype.
Here you can see a summary of our main findings about the hubs of the three networks:

Network	Mean degree	Max degree	Min degree	Degree Span	Up-reg	Down-reg
Normal	110	137	98	39	16	39
Cancerous	126	187	93	94	52	3
Differential	83	222	45	177	50	5

After the computation and analysis of these co-expression networks we can infer that the presence of a thyroid carcinoma is related to the up-regulation of genes (687) more than to their down-regulation (407); this change is also visible from how the amount of up/down regulated genes changes in the hubs of the networks. The presence of the carcinoma also seems to lead to a general increase of the co-expression of genes, since the average degree and max degree of a normal hub are smaller than the ones of a cancerous one. This is more intuitively seen looking at the heatmaps on page 3: the one corresponding to the cancerous samples has much darker colors, meaning stronger co-expressions. Moreover, looking at the last row of the table we can see that, on average, amongst the hubs of the differential network a gene had 83 co-expressions values altered.
It is also interesting to state that 20 genes that were hubs in the cancerous network are also hubs in the differential one.

At the end of our project we were able to select the 5 genes which we believe to have a key role in this cancer, because of their co-expression changes. Most of them are actually studied in the context of thyroid cancer (supplementary information can be found in the references). We decided to focus mainly on ITGA3 and found out that it has been approved as one of the direct targets of miR-524-5p.
MicroRNAs are short noncoding RNAs that act as fine-tuners of the expression of protein-coding or noncoding genes; they have both oncogenic and tumor-suppressive functions. They are beneficial for cancer therapy as they can simultaneously downregulate multiple targets involved in diverse biological pathways related to tumor development. miR-524-5p could inhibit papillary thyroid

cancer cell viability, migration, invasion, and apoptosis through targeting ITGA3, as targeting ITGA3 prevents thyroid cancer progression through different pathways including cell cycling and autophagy.[1] Using miRNA with this gene could therefore be helpful to the research field focused on normalizing genes expression.

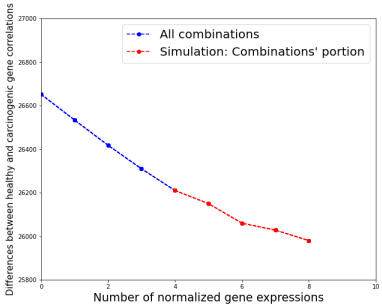
(a) ITGA3 expression in normal condition (b) ITGA3 expression in cancerous condition

4.2 Gene expression regulation

At the end of our project we decided to do some additional research into how difficult it would be to find a right amount of genes to target for a therapy based on miRnas and genes' expression normalization. The idea was to try and simulate a different normalized expression for some genes and find out how closer to the normal condition of co-expression we could get. Since we were not interested in the precise correlation value between two genes we decided to measure the difference between co-expressions by comparing the resulting adjacency matrices, i.e. considering only differences that were big enough to change the relation (edge or no edge) between two genes.
In our trials we decided to only consider the hubs of the differential co-expression network, since they are the more altered genes.

Initially we tried by changing the expression of just one gene, creating a vector with mean equal to the mean expression of that same gene in a normal condition. Then, since the results were not as good as expected we increased the number of genes to alter at the same time (or rather to target with miRna), and looked for the best combination. We tried out all possible combinations of 2, 3 and 4 genes at a time. Unfortunately due to the factor of time we could not try all the combination of 5 genes (3 millions and a half), but trying out half a million of them we managed to find the best result thus far with genes ERBB3 SDC4, GRB7, COL8A1, MEX3A. Normalizing their expression allowed us to reduce the difference of the normal adjacency matrix and the cancerous one from 26652 to 26150 (these numbers are the sum of how many differences are present when comparing the two matrices element-wise).

Knowing that we lacked the means to carry out an exhaustive research we decided to look for a trend in the improvements of the results in function of the number of genes. Doing some additional non-exhaustive tests on sets of 6 and 7 genes too we traced the following function. As we can see when we add more then 5 genes the speed of the improvement rate decreases, because the selection of genes will also include less influential ones. To see in details the code we used for our simulation you can check out this github repository.



One last thing we believe to be interesting to point out is that in all the best combination of five genes we encountered the gene ERBB3, the second most important hub of the differential co-expression network. As well ITGA-3 analysis, miRNA could be helpful also for the ERBB3 targeting: an inverse correlation between miR-143 and miR-145 levels and ERBB3 ones has in fact been identified . It has also been demonstrated that the repression of ERBB3 by miR-143/145 suppressed the proliferation and invasion of cancer cells, and that miR-143/145 showed an anti-tumor effect by negatively regulating ERBB3. [2]
With greater means and more time to keep the simulation going it would be interesting to find out the real best combination with a maximum of 10 genes and find out empirically and not only through, the previous linear function, how helpful that could be in fighting thyroid cancer.

References

[1] Hui Liu, Xi Chen, Ting Lin, Xingsheng Chen, Jiqi Yan, Shan Jiang, *MicroRNA-524-5p suppresses the progression of papillary thyroid carcinoma cells via targeting on FOXE1 and ITGA3 in cell autophagy and cycling pathways*, 2019

[2] Xin Yan, Xi Chen, Hongwei Liang, Ting Deng, Weixu Chen, Suyang Zhang, Minghui Liu, Xiujuan Gao, Yanqing Liu, Chihao Zhao, Xueliang Wang, Nan Wang, Jialu Li, Rui Liu, Ke Zen, Chen-Yu Zhang, Baorui Liu Yi Ba, *miR-143 and miR-145 synergistically regulate ERBB3 to suppress cell proliferation and invasion in breast cancer* 2014

[3] Marialuisa Sponziello 1, Antonella Verrienti 1, Francesca Rosignolo 1, Roberta Francesca De Rose 2, Valeria Pecce 1, Valentina Maggisano 2, Cosimo Durante 1, Stefania Bulotta 2, Giuseppe Damante 3, Laura Giacomelli 4, Cira Rosaria Tiziana Di Gioia 5, Sebastiano Filetti 1, Diego Russo 6, Marilena Celano 2, *PDE5 expression in human thyroid tumors and effects of PDE5 inhibitors on growth and migration of cancer cells* 2015

[4] https://github.com/cat-erina/GeneExpression