DIGITAL EPIDEMIOLOGY AND PRECISION MEDICINE

Differential Analysis of Gene Expression

Cruoglio Antonella   Mascolo Davide   Napoli Mario

GROUP 07

Abstract

The aim of this work was to identify hub genes closely related to the pathogenesis and prognosis of thyroid carcinoma (THCA) using gene expression analysis. In this study, we started from the expression of 18323 genes over 59 patients obtained from The Cancer Genome Atlas (TCGA) database and through differential gene expression analysis we found 669 differentially expressed genes (DEGs). Using these genes, we were able to build Co-expression networks and Differential Co-expressed Network and extract several hubs. Many of these genes are known in literature to be relevant for the development and the treatment of the disease.

1 Introduction

Thyroid carcinoma (THCA) is a common endocrine malignant cancer, with an incidence rate that is increasing over the years [1]. THCA is mainly divided into four categories, including papillary carcinoma (85%), follicular carcinoma (10–15%), medullary carcinoma (5–10%) and undifferentiated THCA (<5%) [2]. Although most tumors are benign, one of the main obstacles for the treatment is the capacity to generate metastasis in other parts of the body, in addition to the postoperative recurrence and the persistent disease [3]. The development of this tumor may depend by different factors among which a genetic change of the THCA; for this reason, it has become very important in recent years to study the genetic expression to find some characteristics that could be used to improve the therapies. Thanks to the huge amount of available data it's possible to use these techniques in order to identify different heterogeneous groups of patients, based on their characteristics, and use a specific therapy for each group. The goal of the study is to analyze the expression profile of 59 patients using RNA sequencing data in order to identify the differential expressed genes for the disease.

2 Materials and Methods

2.1 Data

The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/) is a landmark cancer genomics program, molecularly characterized over 20000 primary cancer and matched normal samples spanning 33 cancer types. THCA gene expression data were obtained from this platform and we have used two datasets: one about tumor tissue and the other about normal tissue. Each row of the datasets represents the expression of a gene registered for different patients that are represented by the columns. Then we did some pre-processing steps on these data. First of all we have considered only the common patients and common genes between the two conditions. We removed the genes that were not expressed with a significant level and we checked that there were no missing values.

2.2 Differentially Expressed Genes (DEGs)

In this step our goal was to identify the Differential Expression Genes and in order to perform this analysis we used the R package "DESeq2". Using the raw counts for both conditions, we obtain a dataset and we extract from it only the genes for which there are enough reads (at least 10 reads). With this data we compute the LFC and the adjusted p-value for the Wald's Test using FDR (Benjamini-Hochberg) adjustement for multiple testing. From this output, we considered only the results for which *p-value* is below 0.05 and $|LFC| \geq 1.2$ and we obtained 669 genes, of which 520 up-regulated genes and 149 down-regulated genes.

2.3 Co-expression network

Once we found the DEGs we used them to define genes co-expression network related to both conditions. We have two matrices composed by 669 genes for 59 patients that we used to compute the two correlation matrices using the Pearson correlation. From the obtained matrices we built the adjacency matrices putting equal to 1 only the pairs for which the adjusted *p-value* ≤ 0.05 and for which $|\rho|$ is higher than a certain threshold that we set to be 0.7. Through the adjacency matrices, we built the genes co-expression networks for both conditions. From them we computed the degree index in order to understand if the networks are scale free and to find the hubs, that are represented by the 5% of the nodes with the highest degree values. Then we compared the hubs sets related to the two conditions and identified the hubs characterizing only cancer tissue.

2.3.1 Soft-Thresholding

In the previous step we built the adjacency matrices using the hard-thresholding method, for which the correlation coefficient ρ_{ij} between gene i and gene j is assessed with respect to a threshold τ and both genes are connected by an edge only if $p_{ij} \geq \tau$. The drawbacks of this method are given by the loss of information and sensitivity to the choice of the threshold τ. Another way to build the adjacency matrix is to use the soft-thresholding method, in which the connection between each gene pair is weighted, this means that all the nodes of the network are ranked according to how strong their connection strength is with respect to the nodes under consideration. There are two types of adjacency function used:

$$A_{ij} = sigmoid(S_{ij}, \alpha, \tau_0) = \frac{1}{1 + e^{-\alpha(s_{ij} - \tau_0)}}$$
$$A_{ij} = power(s_{ij}, \beta) = |s_{ij}|^{\beta}$$

In this work we considered the power adjacency function. By using the R package "WGCNA", we tried to find the best power β to ensure that the resulting network has approximately a scale-free degree. For the cancer co-expression network we selected $\beta = 11$ with a scale free $R^2 = 0.92$. (Figure 1).

For the normal co-expression networks we selected $\beta = 8$ with a scale free $R^2 = 0.91$ (Figure 2).

2.3.2 Different Centrality Index

In the step we computed different centrality indices and check the overlap between the 5% of the nodes with highest CI values and the degree-based hubs. The centraity measures that we considered are: *Betweenness centrality, closeness centrality, eigenvector centrality.*

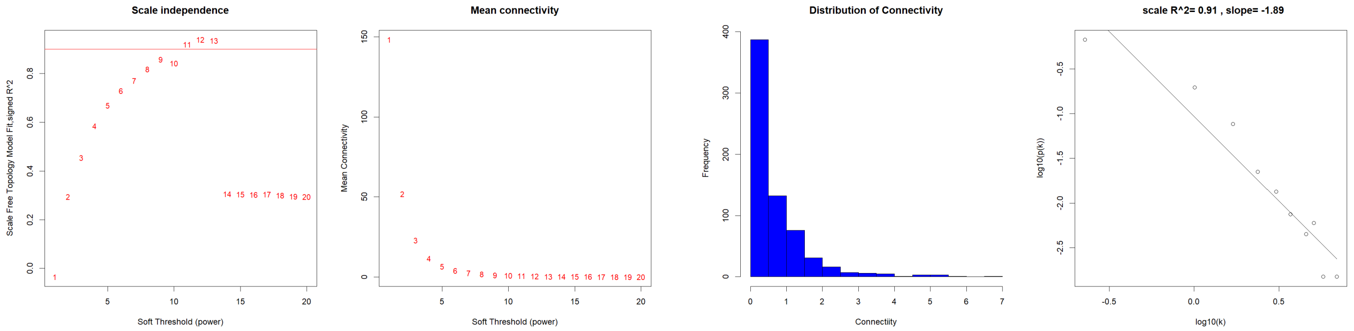


Figure 1: Analysis of cancer network topology for various soft-thresholding powers. The first panel shows the scale-free fit index as a function of the soft-thresholding power. The second panel shows the mean connectivity (degree) as a function of the soft-thresholding power. The third panel shows the histogram of connectivity for $\beta = 11$ (Cancer Network). The fourth panel shows log-scale free topology.

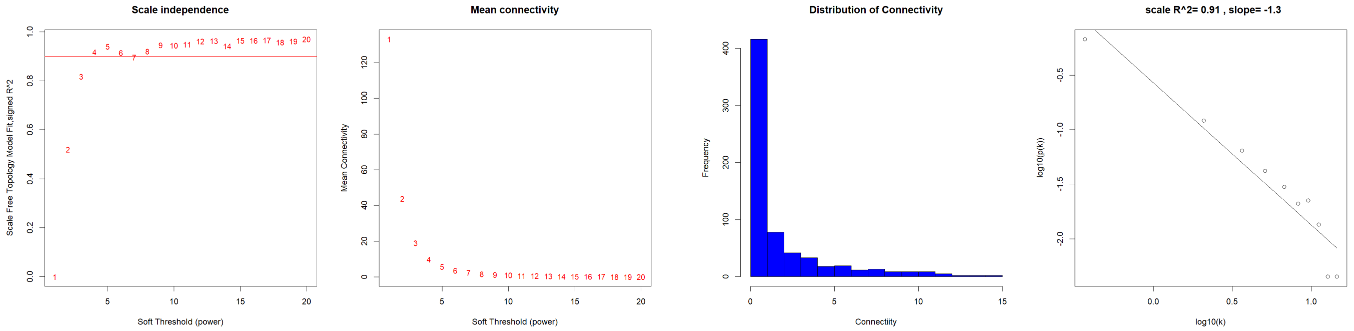


Figure 2: Analysis of normal network topology for various soft-thresholding powers. The first panel shows the histogram of connectivity for $\beta = 11$ (Normal Network). The second panel shows log-scale free topology. The third panel shows the scale-free fit index as a function of the soft-thresholding power. The fourth panel shows the mean connectivity (degree) as a function of the soft-thresholding power.

CI	Common Genes
Betweenness	CLIP3, NECTIN4, PDLIM4, SLC4A4, LGALS3, RIN1
Closeness	ITGA3, ENTPD2
Eigenvector	MIR31HG, CLIP3, MDK, NPAS1, VAC14-AS1, PROC, CCDC33, TMEM58L, DOK7, CASC15, ENTPD2, SPOCD1, LINC02981, PIANP, RIN1, C1QTNF12

Table 1: Overlap for cancer condition

CI	Common Genes
Betweenness	KCNN4, SLC1A5, STING1, RUNX1, HES6, CDKN2A, PDZK1IP1, DUSP4, STRA6, SYTL1
Closeness	NFE2L3, ALOX5, BUD, RASGRF1
Eigenvector	NFE2L3, KCNN4, TMC6, SLC1A5, STING1, ALOX5, CTSH, BID, WNT10A, RASGRF1, RUNX1, SPOCK2, DUSP4, STRA6, SYTL1, TMEM163

Table 2: Overlap for normal condition

2.3.3 Different Similarity Measure

In this section we tried to perform the study using a different correlation measure. We chose to use the *Spearman Correlation* to build the adjacency matrix for both conditions. After to obtain the networks, we verified if they are scale-free networks and we extracted the hubs.

Condition	Common Genes
Cancer	KRT19, TMPRSS4, ITGA3, EVA1A, NECTIN4, PERP, RUNX1, SLC4A4
Normal	NFE2L3, KCNN4, TMC6, STING1, ALOX5, RUNX1, DUSP4

Table 3: Intersection between Pearson and Spearman Hubs

The hubs characterizing only the cancer network are: TMPRSS4, FN1, CD55, ELF3, EVA1A, GRB7, AHNAK2, KRT19, CRYBG2, MUC1, TMPRSS6, ITGA3, MRO, PERP, B3GNT3, MPPED2, SERPINA1, NECTIN4, SLC4A4, MET, ERBB3.

3 Differential Co-Expressed Network

Using only DEGs we computed the differential co-expressed network for both conditions. First of all we applied the *Fisher Z-transformation* on the correlation coefficients in each condition ($z1$ and $z2$) and from these two z-score matrices we obtained an overall Z-score computed as follow:

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}, \quad where\ n_i = sample\ size\ for\ the\ condition\ i.$$

Therefore, we used this values to compute the adjacency matrix. Setting a threshold equal to 3, we assigned the value 0 if the $|Z| \leq t$, otherwise 1 to build the differential co-expression network.

We did the same operation of the previous step. First of all we verified if the network is scale free and then we made a comparison with the hubs computed in the step 3.

Condition	Common Genes
Cancer	ITGA3, PERP, EVA1A, NPAS1
Normal	ETV4

Table 4: Intersection between Differential Expression Hubs

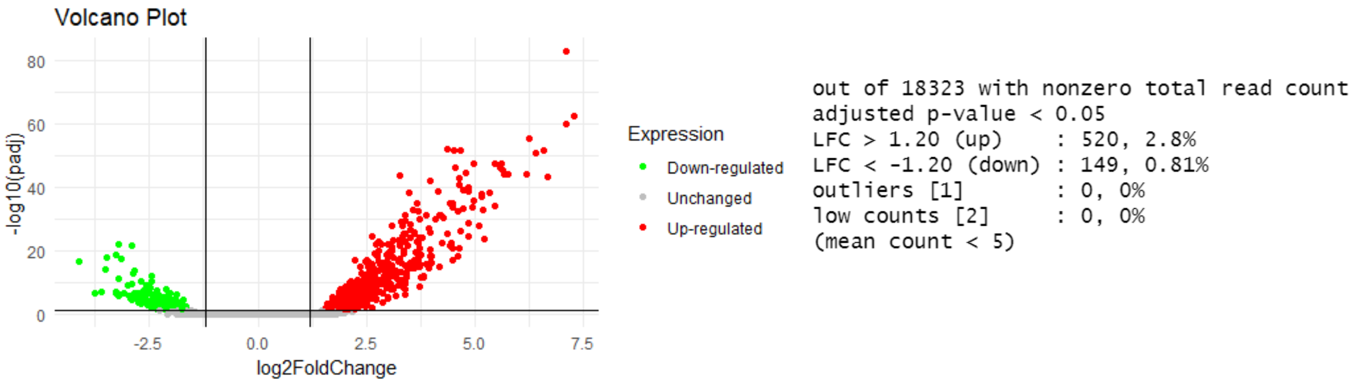
4 Patient Similarity Network

Using cancer gene expression profile, we computed the patient similarity network. We chose to measure patient similarity transforming the *Euclidean Distance* into a similarity measure. We used this matrix to compute the network on which was applied the Louvain algorithm to make community detection. This method is useful to find high modularity partitions of large networks in short time. Using the techniques defined above, we computed the patient similarity network also for the normal condition in order to identify different subgroups. The network are shown in the Figure 6.

5 Results and Discussion

5.1 Differential Expressed Genes (DEGs)

In order to visualize the DEGs extracted in the section 2.2, we used the *Volcano Plot* that is a scatterplot that shows the adjusted p-value vs fold change to identify genes with a large FC that are also statistically significant.



```
out of 18323 with nonzero total read count
adjusted p-value < 0.05
LFC > 1.20 (up)     : 520, 2.8%
LFC < -1.20 (down)  : 149, 0.81%
outliers [1]        : 0, 0%
low counts [2]      : 0, 0%
(mean count < 5)
```

Figure 3: Volcano plot and output of DEGs.

5.2 Degree Distribution Networks

Once we computed the degree index for each node of the networks, we can plot the degree distributions and we can see in the figure 7 that both networks are scale-free.

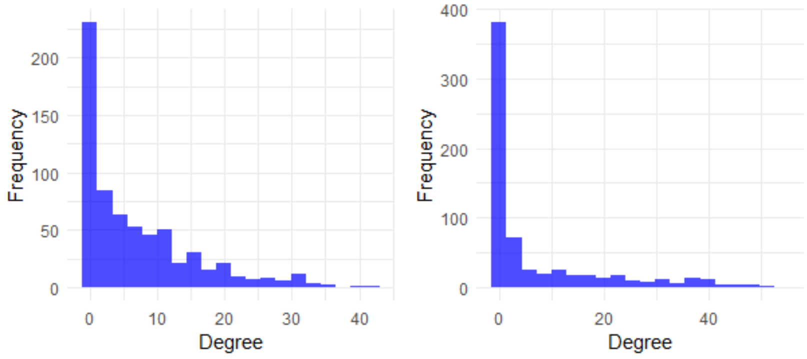


Figure 4: Degree distribution for Co-expression Networks (Cancer and Normal)

From the previous distribution we extracted the 5% of the nodes with the highest degree. We have 26 hubs for the cancer network and 23 hubs for the normal network and they have only one hub in common, that is: RUNX1. The genes that characterized the cancer networks are:

Condition	Hubs Genes
Cancer	MIR31HG, CLIP3, MDK, NPAS1, KRT19, VAC1-AS1, PROC, TMPRSS4, CCDC33, TMEM59L, ITGA3, DOK7, CASC15, ENTPD2, EVA1A, SPOCD1, NECTIN4, LINC02981, PERP, PDLIM4, PIANP, RUNX1, SLC4A4, LGALS3, RIN1, C1QTNF12

Table 5: Hubs characterized Cancer condition

5.3 Differential Co-Expressed Network

Using the approach defined in the section 3, we obtained the differential co-expressed network and we report the following sub-graph that represents the most connected hub and its neighbours. The most connected hub is: PRSS22, that is a gene that encodes a member of the trypsin family of serine proteases.

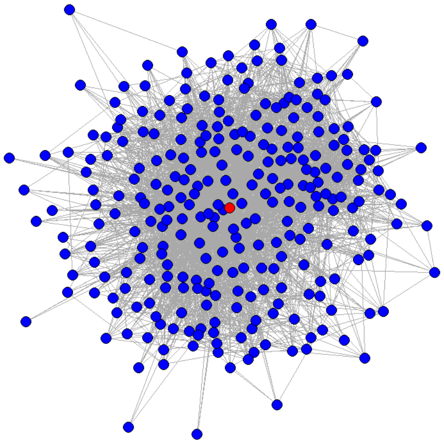


Figure 5: Sub-graph of most connected hub and its neighbours

These are the hubs in the differential co-expressed network: PRSS22, ITGA3, PSD, ERBB3, RSPO4, MET, GALNT7, PERP, EVA1A, ECE1, TGFBI, APOE, NPAS1, ARNTL, C5AR2, ENTPD1, PDE5A, GRB7, XPR1, MXRA8, BNIPL, LMOD1, SPINT1, SCG5, PIP5KL1, ZFPM2, KCNS3, ETV4, H2AW, DMD, NPTXR, ELFN1, MEX3A, KCNJ2-AS1.

5.4 Relevant Genes and Reference Papers

In this section we provided some reference papers where the genes that we identified are studied in the context of thyroid cancer. We found the references for the following genes:

- MIR31HG: is the most connected hub for the co-expression cancer network and it is over-expressed in the human thyroid cancer. [4]

- MDK: also this gene is over-expressed in the human thyroid cancer and is the third most connected hub for the co-expression cancer network. [5]

- SERPINA1, FN1: these two genes characterize the co-expression cancer network obtained with the Spearman correlation. [6][7]

- ITGA3: this is a very interesting gene and it is central to the development of the disease. It characterize the co-expression cancer network using both correlations (Pearson and Spearman) and also using the Closeness centrality. Furthermore, it is also present in the differential co-expressed hubs. [8]

- LGALS3: this gene is one of the most connected hub in the co-expression cancer network and with higher betweenness. [7]

- MET: this gene is present in the co-expression cancer hubs using Spearman correlation and in the differential co-expression network hubs. [7]

- SLC4A4: this gene is down-regulated and it is present in the co-expression cancer network using both correlations and is also one of the genes with highest betweenness centrality measure. [9]

5.5 Patient Similarity Network

In the following plots, is shown the community detection obtained with Louvain algorithm for both conditions. We obtained three sub-groups for both conditions and a low modularity (0.12 for cancer and 0.082 for normal). Maybe this is due by the fact that we used only one "view" of the patient similarity that is given by the gene expression data, but we don't have additional informations that can be useful. Another possible explanation could be the fact that this pathology has homogeneous characteristics among patients.

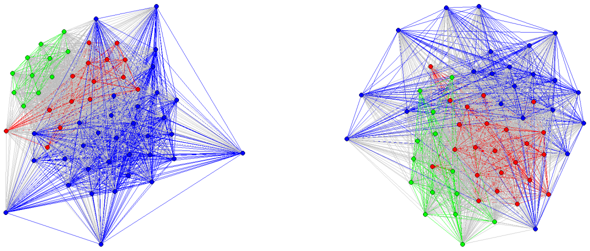


Figure 6: Patient Similarity Network for Cancer (left) and Normal (right) condition

6 References

[1] Kim K, Jeon S, Kim TM, Jung CK. Immune Gene Signature Delineates a Subclass of Papillary Thyroid Cancer with Unfavorable Clinical Outcomes. Cancers (Basel), 2018 Dec 5.

[2] Liu R, Cao Z, Pan M, Wu M, Li X, Yuan H, Liu Z. A novel prognostic model for papillary thyroid cancer based on epithelial-mesenchymal transition-related genes. Cancer Med, 2022 May 24.

[3] Wang Y, Huang H, Hu F, Li J, Zhang L, Pang H. CITED1 contributes to the progression of papillary thyroid carcinoma via the Wnt/β-catenin signaling pathway. Onco Targets Ther, 2019 Aug 21.

[4] Peng S, Chen L, Yuan Z, Duan S. Suppression of MIR31HG affects the functional properties of thyroid cancer cells depending on the miR-761/MAPK1 axis. BMC Endocr Disord, 2022 Apr 20.

[5] Jee YH, Celi FS, Sampson M, Sacks DB, Remaley AT, Kebebew E, Baron J. Midkine concentrations in fine-needle aspiration of benign and malignant thyroid nodules. Clin Endocrinol (Oxf), 2015 Dec.

[6] Vierlinger K, Mansfeld MH, Koperek O, Nöhammer C, Kaserer K, Leisch F. Identification of SERPINA1 as single marker for papillary thyroid carcinoma through microarray meta analysis and quantification of its discriminatory power in independent validation. BMC Med Genomics, 2011 Apr 6.

[7] Huang, Ying, et al. "Gene Expression in Papillary Thyroid Carcinoma Reveals Highly Consistent Profiles." Proceedings of the National Academy of Sciences of the United States of America, vol. 98, no. 26, 2001.

[8] Zhang G, Li B, Lin Y. Evaluation of ITGA3 as a Biomarker of Progression and Recurrence in Papillary Thyroid Carcinoma. Front Oncol, 2022 Jan 31.

[9] Huang Y, Ling J, Chang A, Ye H, Zhao H, Zhuo X. Identification of an immune-related key gene, PPARGC1A, in the development of anaplastic thyroid carcinoma: in-silico study and in-vitro evaluation. Minerva Endocrinol (Torino), 2022 Jun.