ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
"Antonio Ruberti"
CONSIGLIO NAZIONALE DELLE RICERCHE

F. Conte, G. Fiscon

SWIM TOOL FOR STUDYING HUMAN PAPILLARY THYROID CARCINOMA

R. 7, 2018

Federica Conte - Institute for System Analysis and Computer Science "Antonio Ruberti" (IASI), CNR, Via dei Taurini 19, 00185 Rome, Italy. Email: federica.conte@iasi.cnr.it.

Giulia Fiscon - Institute for System Analysis and Computer Science "Antonio Ruberti" (IASI), CNR, Via dei Taurini 19, 00185 Rome, Italy. Email: giulia.fiscon@iasi.cnr.it.

ISSN: 1128-3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR

via dei Taurini 19, 00185 ROMA, Italy

tel. ++39-06-49937101/02
fax ++39-06-49937106
email: iasi@iasi.cnr.it
URL: http://www.iasi.cnr.it

Abstract

Thyroid cancer is the most frequent endocrine malignancy, and accounts for 1% of all tumours. Papillary thyroid carcinoma (PTC) and follicular thyroid carcinoma are the most frequent. PTC has a strong genetic component, since it displays the highest relative risks in first degree relatives. PTC has a social as well as economic impact that motivates investigation of novel methods and tools that can exploit the wealth of whole-genome expression data made available in recent years to improve cancer clinical practice. Such "big data" provide at the same time an opportunity and a challenge since devising algorithms and tools effectively harvesting the knowledge buried inside them is a major computational endeavor. Here, we study the PTC dataset from TCGA by running a recently developed software called SWIM, in order to extract a small pool of genes, called switch genes, crucially for the transition from physiological to pathological phenotype of the disease understudy. In particular, SWIM unveiled 131 switch genes out of 1718 differential expressed genes whose up-regulation was found strongly associated with p53 signaling pathway. Among the switch genes, we selected some promising candidate to be disease genes for thyroid carcinoma.

Key words: SWIM software, Papillary thyroid carcinoma, network analysis, disease genes

1. Introduction

Among endocrine tumors, thyroid cancer is the most common type that has rapidly increased in global incidence in recent decades [1]. The thyroid cancer is usually asymptomatic for long periods and thus, one of the main problems is that no clinical clue to its diagnosis are available. Approximately 50% of the malignant nodules are discovered during a routine physical examination or surgery for benign disease, while the other 50% are usually noticed by the patient, usually as asymptomatic nodules. Due to the typically indolent nature of this cancer, long delays in diagnosis are frequent that in turn may lead to a substantially worsening of disease course. Indeed, although the death rate of thyroid cancer is relatively low, the rate of persistence and recurrence is high, which is associated also with the increasing of cancer incurability and patient mortality [2]. Current treatments involve surgery, thyroid hormone, and radioactive iodine (RAI) therapy but they are often not effectively curative. The recent progress in understanding the molecular mechanisms underlying thyroid cancer represent a promising strategy for the development of more-effective treatments. The identification of genetic and epigenetic alterations of signaling pathways — such as the MAPK pathway and the PI3K/AKT pathway — is reshaping thyroid cancer medicine [3-4]. Numerous genetic alterations have been shown to play a fundamental role in the tumorigenesis of various thyroid tumors [5-7]. Papillary thyroid carcinoma (PTC) and follicular thyroid carcinoma are the most frequent thyroid tumors. PTC has a strong genetic component, since it displays the highest relative risks in first degree relatives.

To predict patient survival and decide best-suited treatment is central to oncology. Unfortunately, this is often challenged by vast tumor heterogeneity. High-throughput expression analysis yielded the emergence of several "gene expression signatures" over the last 15 years [8-12], or lists of genes whose expression is significantly associated with tum-or subtype, progression or likelihood of patient drug response (diagnostic, prognostic or predictive gene expression signature, respectively) [13-14]. However, gene signatures did not modified daily clinical practice and therapy selection, still firmly relies on immunohistochemical, histopathological and clinical features. Also, they often fail to consider important molecules (e.g. non-coding RNAs) and potential new drug targets as deriving from old techniques (e.g. microarrays). Hence, most cancer gene signatures risk to become obsolete even before any clinical potential evaluation.

Next-generation sequencing techniques produce masses of genomic data and are boosting the birth of international projects for storage, management and analysis. Among others, The Cancer Genome Atlas (TCGA) [15-16] is a resource project providing heavy patient data collections, including gene expression and clinical data, that is fueling many studies aimed at improve cancer.

In this work, we aim to capture specific gene subsets with key regulatory roles therein. To this end, we run SWIM software, which is able to highlight a small pool of genes (~100/20,000), called switch genes, with crucial roles in the cell phenotype changes. Our aim is to extract switch genes in the comparison between thyroid cancer and matched-normal samples. In particular, we analyzed papillary thyroid carcinoma (thca) dataset obtained from TCGA [15-16] (updated to December 2014).

2. Methods

2.1 Data retrieval

We download and analyzed RNA- and microRNA- (miRNA) sequencing assays of human thyroid carcinoma from TCGA repository, including: 1) 572 RNA-sequencing samples (of which, 513 are tumor and 59 normal samples) relative to 505 unique patients; 2) 573 miRNA-sequencing samples (of which, 514 are tumor and 59 normal samples) relative to 506 unique patients. Out of the whole set of patients, 59 have samples of cancer and matched normal tissues for both the RNA-sequencing (concerning protein-coding and non-coding RNAs abundance) and miRNA-sequencing patients.

We also downloaded and analyzed clinical data of 505 unique patients from TCGA. Patients are mainly female subjects (Fig.1A) and of white race (Fig.1B). It is well-known the generally indolent character of thyroid carcinoma with a tendency to chronicize and a favorable prognosis due to the low probability to show both clinical lymph node metastasis and distant metastasis. The analyzed dataset confirms this feature, where only 12 out of 505 patients have involved metastatic sites (Fig.1C) and the majority of them (57%) fall in the stage I of the thyroid cancer (Fig.1D).

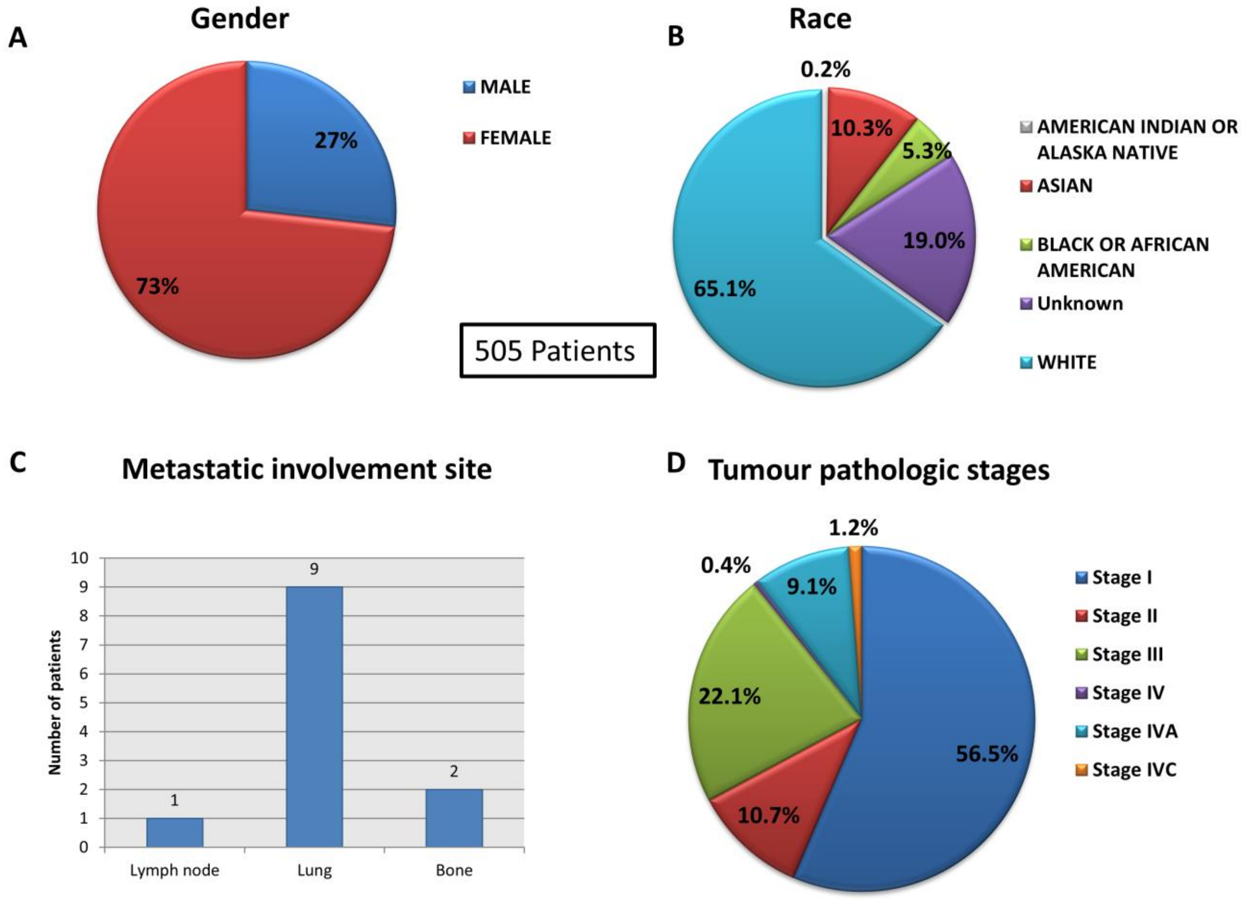


Figure 1 Overview of clinical data

The thyroid cancer mainly occurs in the ages ranging from 33 to 56 with a peak of the age at diagnosis between 33 and 38 years old (Fig.2).

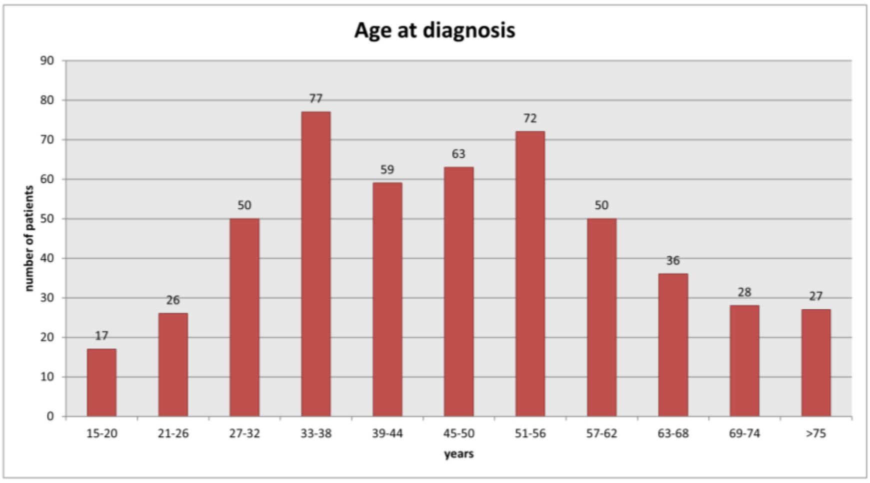


Figure 2 Overview of clinical data

The usual papillary thyroid carcinoma is the most common thyroid tumor, accounting for more than 70% of all thyroid tumors (357 patients out of 505) as shown in Fig.3.

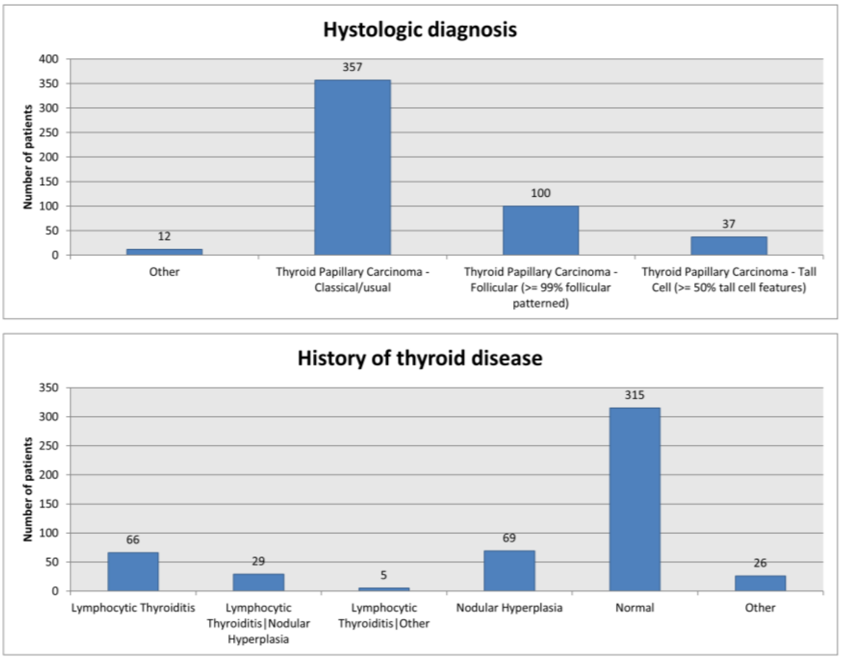


Figure 3 Overview of clinical data

The well-known slow growing of thyroid cancer explains also the relatively long survival percentage of patients (Fig 4A), whereas the reduction of survival of patients with metastatic involvement sites is more noticeable (Fig.4B).

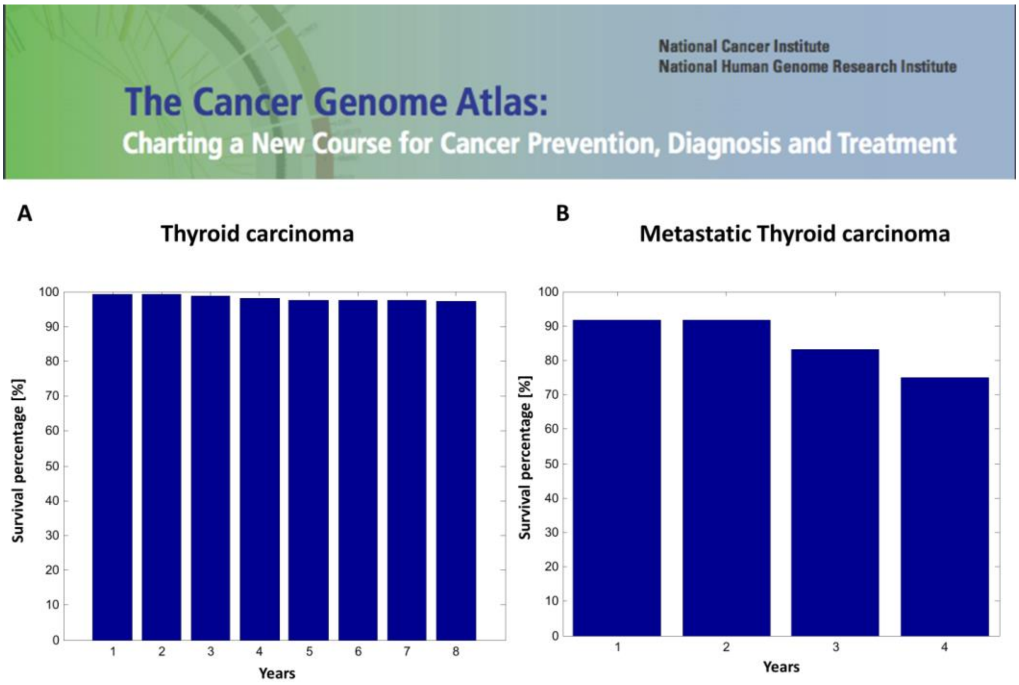


Figure 4 Thyroid cancer survival rate

2.2 SWIM algorithm

SWItchMiner (SWIM) [17] is a software able to extract information contained in complex networks, such as cancer gene networks. It combines topological properties of correlation networks with gene expression data combining genome-wide expression data and the topological properties of the correlation networks. The aim is to identify a small list of genes, called “switch genes”, with a key role in determining the transition between two interesting biological conditions such as physiological/pathological state. SWIM encompasses the steps briefly summarized in the following.

Differential Gene Expression Analysis

SWIM applied a pre-processing and filtering phase on data in order to remove genes whose expression did not change with statistical significance between the two conditions (matched normal and tumor). SWIM selected, for Thyroid Cancer of TCGA, 1682 RNAs and 36 miRNAs showing statistically significant differential expression based on False Discovery Rate (FDR) < 0.05.

Network Analysis and Identification of switch genes

SWIM built a co-expression network of differentially expressed RNAs and miRNAs based on the Pearson correlation between expression profiles of gene pairs. In this network, two nodes are connected if the absolute value of the Pearson correlation for their expression profiles is greater than a given threshold (for this study the selected threshold is 0.70 or 92th percentile). The choice of this threshold should reflect a right balance between the number of edges and the number of connected components of the network: the number of edges should be as small as possible in order to have a manageable network (pointing towards a higher threshold) and the number of connected components should be as small as possible in order to preserve the integrity of the network (pointing towards a smaller threshold).

In order to detect the community structure of our network, SWIM used the k-means clustering algorithm [18], which partitions n objects (here network nodes) into a predefined number N of clusters. The quality of clustering was evaluated by minimizing the Sum of the Squared Error (SSE), depending on the distance of each object to its closest centroid. A reasonable choice of the number of clusters is suggested by the position of an elbow in the SSE plot computed as function of N. As distance measure, SWIM used dist(x,y) = 1 – ρ(x,y), where ρ(x,y) is the Pearson correlation between expression profiles of nodes x and y.

After finding the communities in the network, SWIM classified nodes by computing the Average Pearson Correlation Coefficients (APCCs) between the expression profiles of a hub (i.e., nodes with more than 5 connections [19]) and each of its nearest neighbors. This classification of hubs was set in [19] where the authors studied protein-protein interactions (PPI) networks in yeast. They found a general bimodal distribution of the APCC corresponding to two classes of hubs that they named "party hubs" (with very high positive values of APCC) and "date hubs" (with moderate positive values of APCC). By applying this definition to the different gene co-expression networks, SWIM almost always found a trimodal distribution of APCC [17], which allowed identifying an additional class of hubs, called "fight-club hubs", characterized by negative values of APCC.

Then, SWIM combined expression data with the topological properties of nodes by using the cartographic representation of modular networks presented in [20] that assign a role to each node based on their inter-cluster and intra-cluster connectivity.

This is reached by defining two statistics: the clusterphobic coefficient K_π and the within-module degree z_g. The clusterphobic coefficient measures the ``fear'' of being confined in a cluster, in analogy with the claustrophobic disorder. The global within-module degree $z_g$ measures how ``well-connected'' each node is to other nodes in its own community. According to K_π and z_g values, the plane is divided into seven regions (R1-R7), each defining a specific node role [20]. High z_g values correspond to nodes that are hubs within their module (local hubs), while high values of K_π identify nodes that interact mainly outside their community, i.e. having much more external than internal links.

Finally, SWIM colored each node in the plane identified by z_g and K_π according to its APCC value thus defining what we called a heat cartography map. SWIM identifies "switch genes" as a special subclass of fight-club hubs falling in R4 region that are no local hubs and mainly interact outside their community.

3. Results

An unsupervised clustering analysis correlation of the comprehensive human transcriptome, which represents the expression of 20531 RNAs, including messenger RNAs (mRNAs) and long non coding RNAs (lncRNA), and 1046 microRNAs (miRNAs) in 118 tissues of human thca, revealed a clear distinction between cancer and matched-normal tissues (Fig. 5A-B) suggesting a fundamental shift in the global gene expression during the transition from physiological to pathological condition.

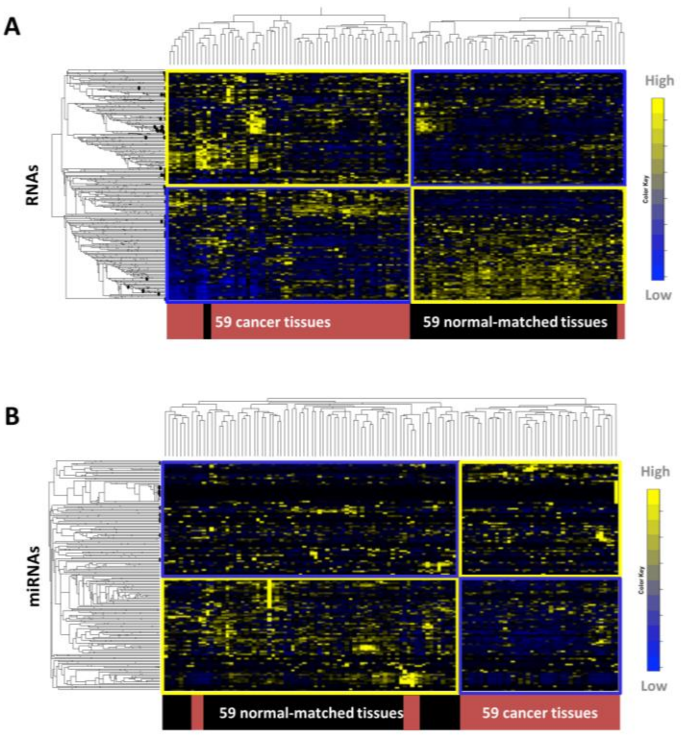


Figure 5 Heatmap of whole transcriptome

Restricting the whole transcriptome to the only differentially expressed mRNAs, lncRNAs, and miRNAs, the rewiring characterizing this transition appears more evident (Fig.6 A-B, Tables 1-2).

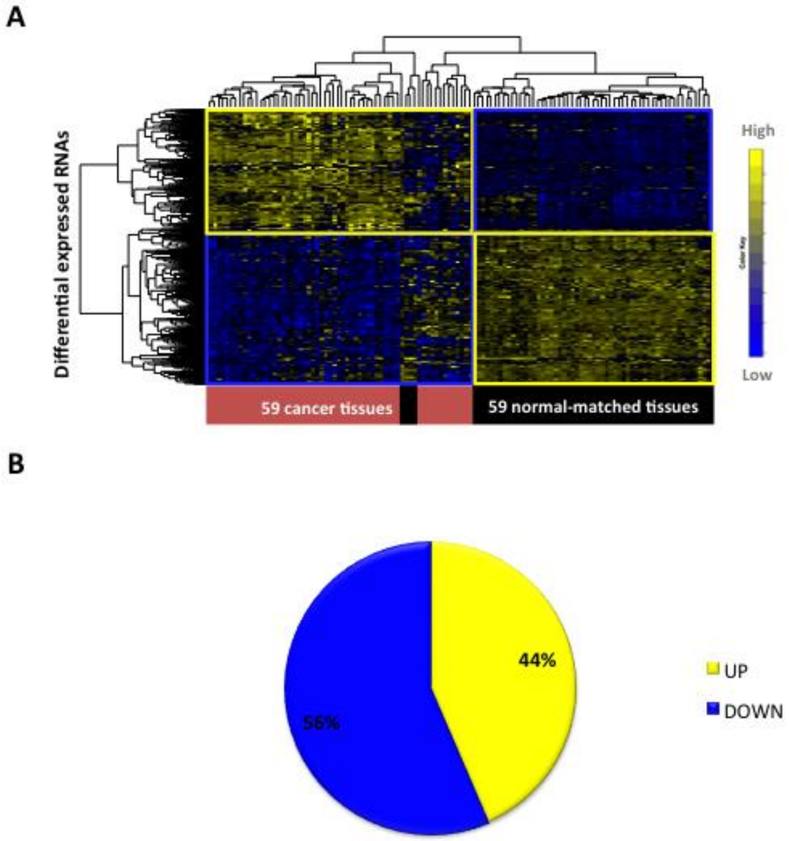


Figure 6 Heatmap of differential expressed genes

To shed light on the molecular mechanism underlying this rewiring, we run SWIM that unveiled 131 switch genes out of 1718 differential expressed genes candidate to be disease genes for thyroid carcinoma (Table 3, Fig.7A). Switch genes included 119 mRNAs and 8 lncRNAs and 4 miRNAs, and with a balanced number of significantly up-regulated (52%) and downregulated (48%) genes in tumor tissues (Fig.7B-C). Moreover, SWIM studied the effect of targeted removal of date/party/fight-club hubs and switch genes on the thca correlation network topology (Fig.7D). This analysis showed a critical contribution of switch genes and fight-club hubs in preserving the integrity of the network.

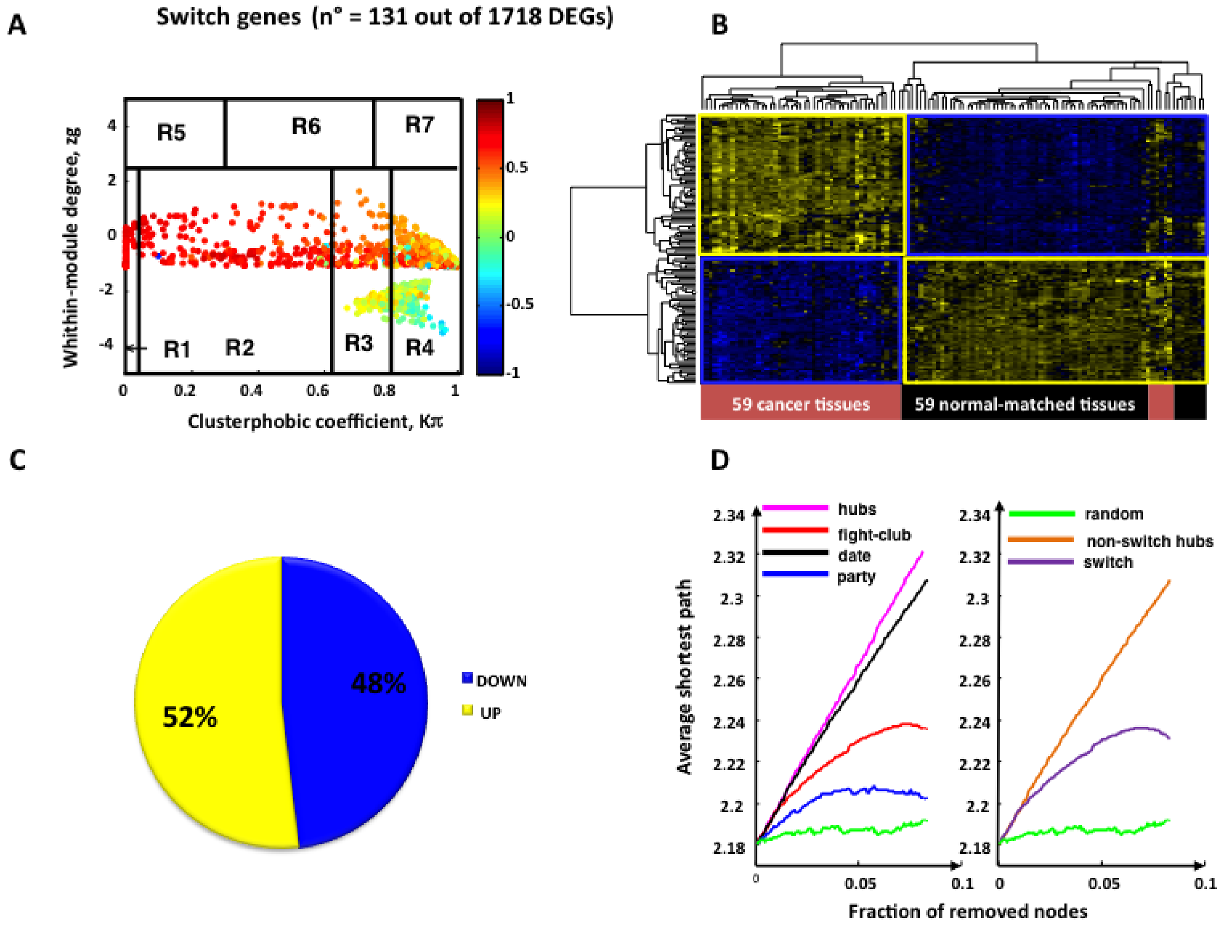
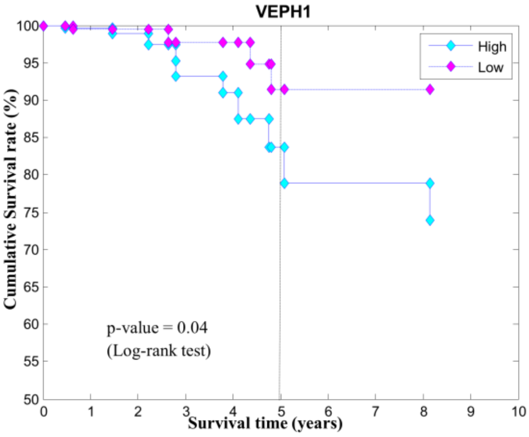
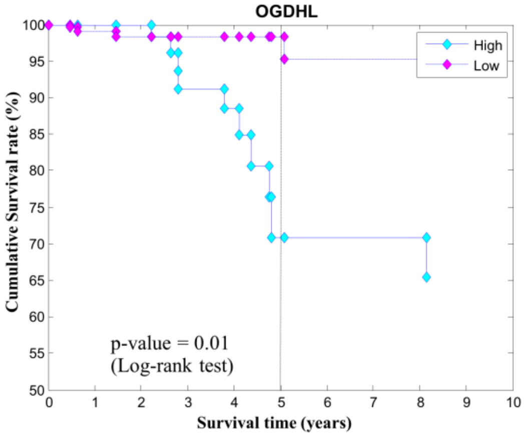
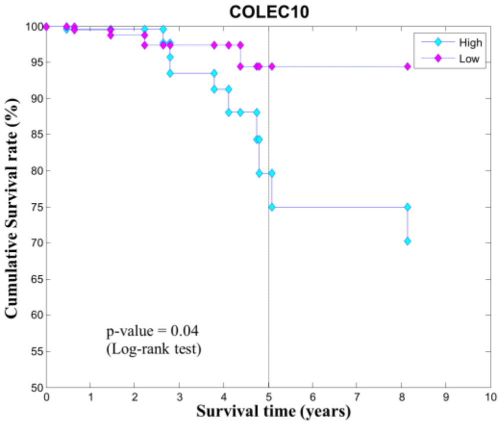


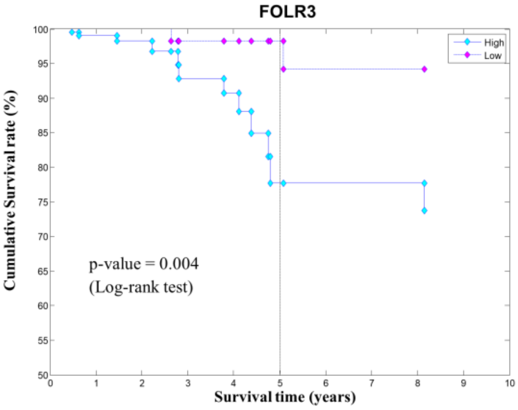
Figure 7 SWIM results on thca dataset

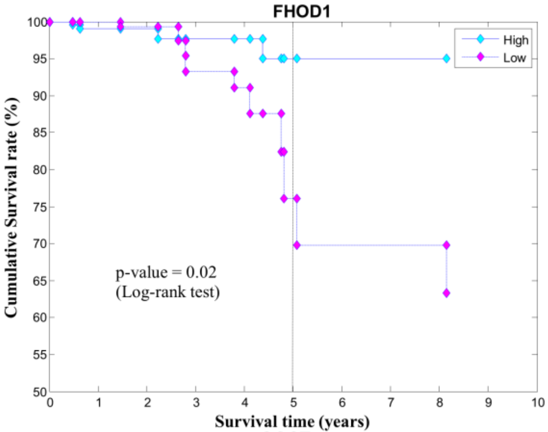
In order to evaluate the clinical relevance of the thca switch genes identified by SWIM, we investigated their prognostic value by performing a Kaplan-Meier survival analysis for each of them (Fig. 8). It is worth noting that the patients' survival rate increases of around 20 percentage related to the low/high expression of the selected switch genes. This is a significant result with respect to the already high rate of survival for patients affected by thyroid carcinoma.

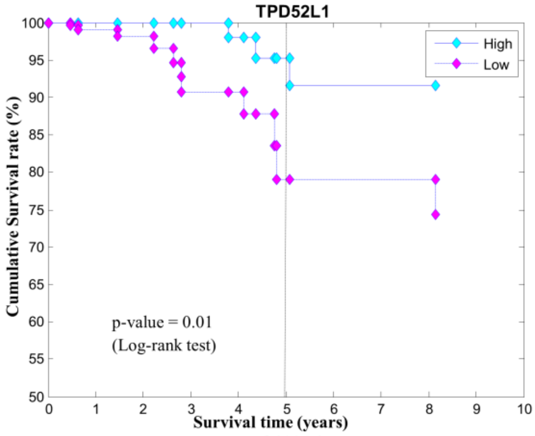

VEPH1
High
Low
Cumulative Survival rate (%)
100
95
90
85
80
75
70
65
60
55
50
0
1
2
3
4
5
6
7
8
9
10
Survival time (years)
p-value = 0.04
(Log-rank test)

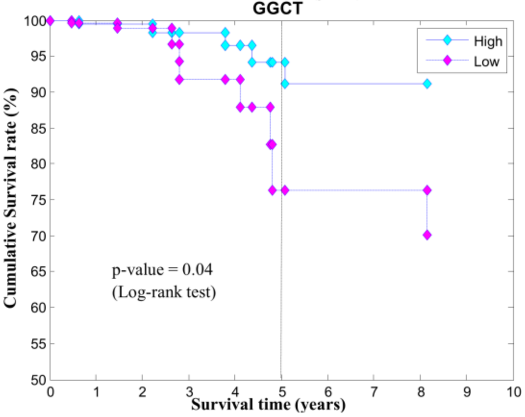

OGDHL
High
Low
Cumulative Survival rate (%)
100
95
90
85
80
75
70
65
60
55
50
0
1
2
3
4
5
6
7
8
9
10
Survival time (years)
p-value = 0.01
(Log-rank test)

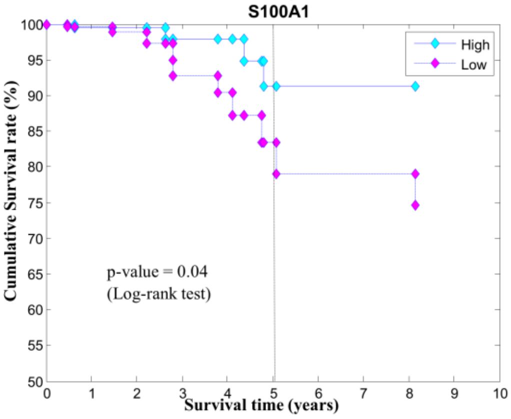

COLEC10
High
Low
Cumulative Survival rate (%)
100
95
90
85
80
75
70
65
60
55
50
0
1
2
3
4
5
6
7
8
9
10
Survival time (years)
p-value = 0.04
(Log-rank test)


FOLR3
High
Low
Cumulative Survival rate (%)
100
95
90
85
80
75
70
65
60
55
50
0
1
2
3
4
5
6
7
8
9
10
Survival time (years)
p-value = 0.004
(Log-rank test)


FHOD1
High
Low
Cumulative Survival rate (%)
100
95
90
85
80
75
70
65
60
55
50
0
1
2
3
4
5
6
7
8
9
10
Survival time (years)
p-value = 0.02
(Log-rank test)


TPD52L1
High
Low
Cumulative Survival rate (%)
100
95
90
85
80
75
70
65
60
55
50
0
1
2
3
4
5
6
7
8
9
10
Survival time (years)
p-value = 0.01
(Log-rank test)


GGCT
High
Low
Cumulative Survival rate (%)
100
95
90
85
80
75
70
65
60
55
50
0
1
2
3
4
5
6
7
8
9
10
Survival time (years)
p-value = 0.04
(Log-rank test)


S100A1
High
Low
Cumulative Survival rate (%)
100
95
90
85
80
75
70
65
60
55
50
0
1
2
3
4
5
6
7
8
9
10
Survival time (years)
p-value = 0.04
(Log-rank test)

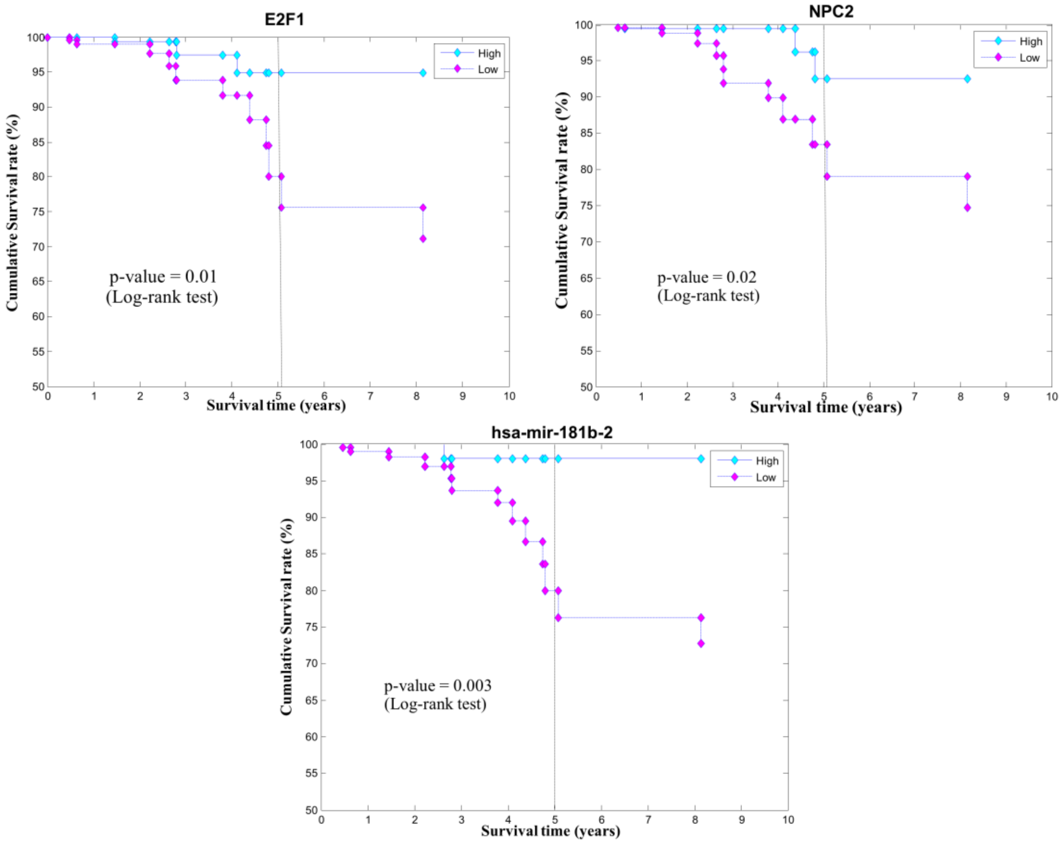


Figure 8 SWIM results on thca dataset

We investigated switch genes function by performing an enrichment analysis in KEGG pathway [21] by using the FIDEA server [22]. This analysis pointed out a strong association of up-regulated switch genes in the p53 signaling pathway (Fig.9).

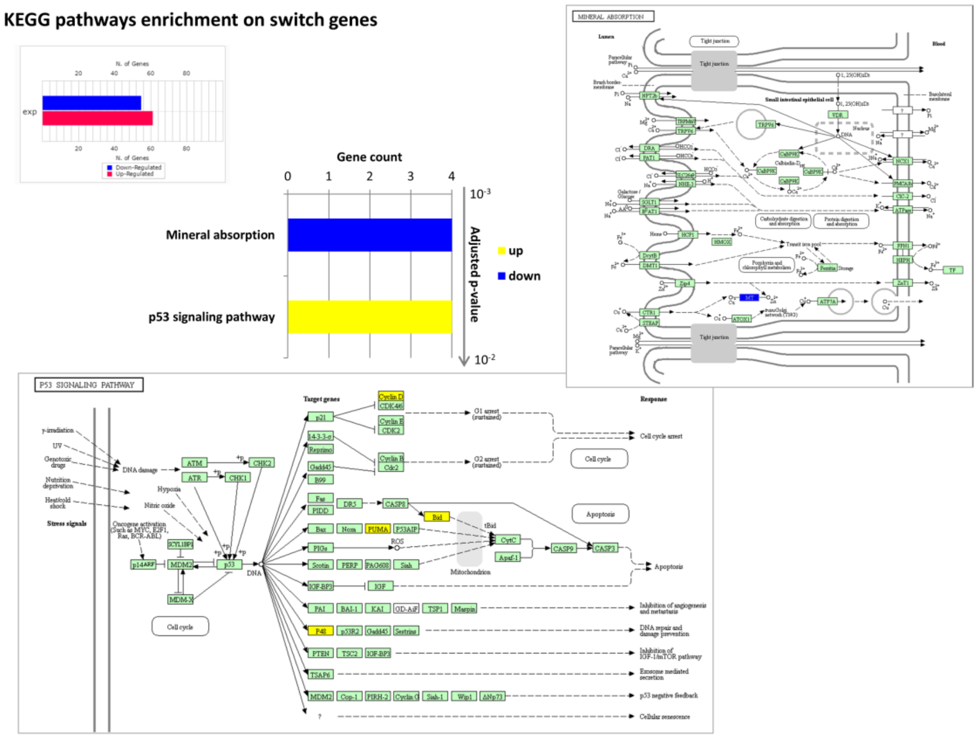


Figure 9 Functional enrichment analysis

A motif analysis of promoter regions of switch genes performed by using Pscan software [23] revealed the presence of an enrichment in DNA motifs belonging to EGR, SP, E2F transcription factor families and TP53 (Fig. 10A).

To elucidate the cascade of events that could lead to the pathological condition, we investigated whether the switch genes list resulted significantly enriched in targets of thyroid-specific miRNAs and in which processes are involved their first neighbors.

From the miRNAs enrichment analysis only the miR-149-5p stands out (Fig. 10B) and among its targets we found three targets of TP53 [24]. Interestingly, the miR-145-5p resulted to be down-regulated, while its targets up-regulated in thyroid cancer tissues.

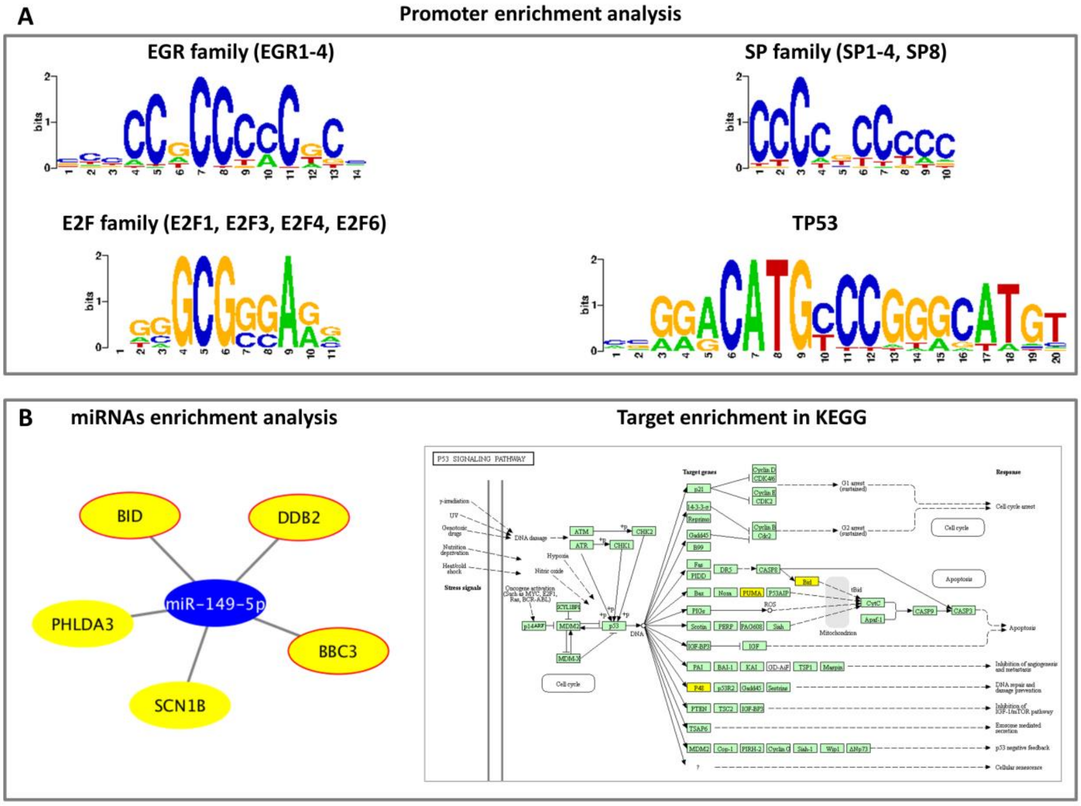

A
Promoter enrichment analysis
EGR family (EGR1-4)
SP family (SP1-4, SP8)
E2F family (E2F1, E2F3, E2F4, E2F6)
TP53
bits
B
miRNAs enrichment analysis
Target enrichment in KEGG
BID
DDB2
PHLDA3
miR-149-5p
BBC3
SCN1B
P53 SIGNALING PATHWAY

Figure 10 Promoter and miRNAs enrichment analysis

By investigating the negatively correlated neighbors of switch genes (Table 4), we found a set of six down-regulated genes (DIO1, DIO2, TG, IYD, SLC5A5, TPO) having a role in thyroid related processes such as thyroid metabolic process and hormone generation (Fig.11-12). Interestingly, by analyzing the positively correlated neighbors of switch genes (Table 5), we found the same set of six down-regulated genes (DIO1, DIO2, TG, IYD, SLC5A5, TPO) that are in turn correlated to other switch genes, this time down-regulated in thyroid cancer (Fig.13).

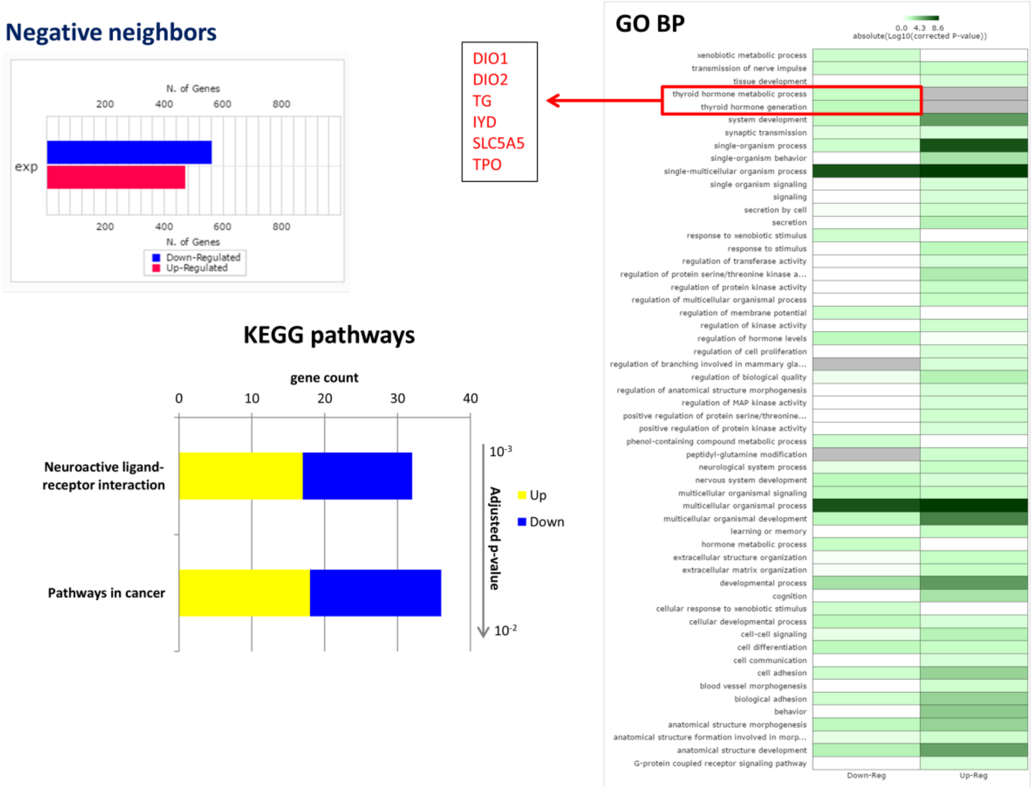


Figure 11 Negative neighbors functional enrichment analysis

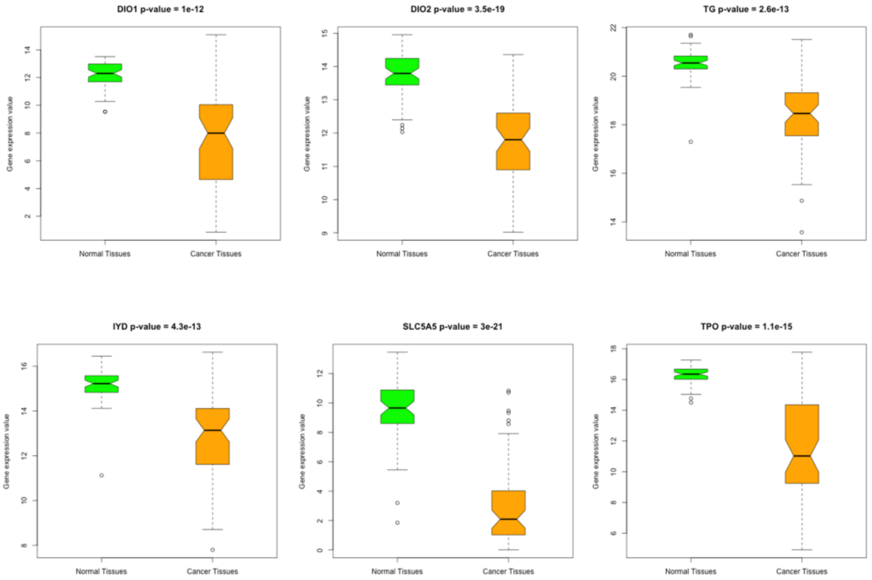


Figure 12 Set of thyroid related genes

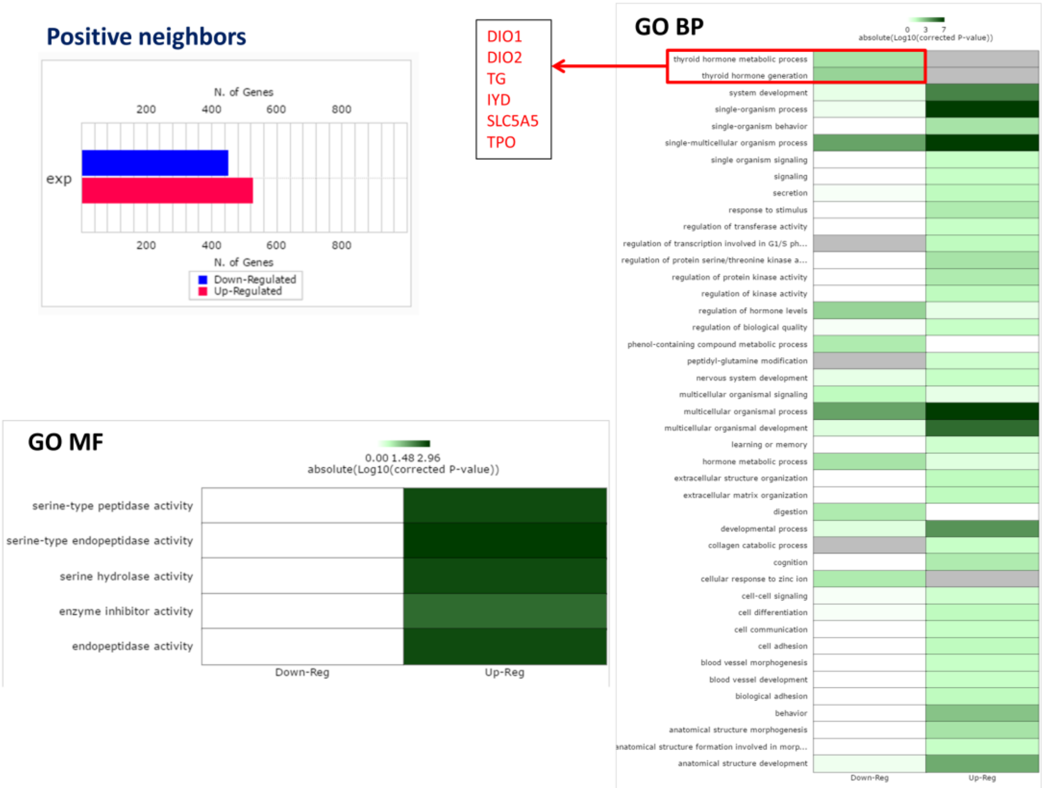


Figure 13 Positive neighbors functional enrichment analysis

We extract the switch genes negatively and positively correlated whit the set of the above-mentioned six genes. We found that the six thyroid related genes shared 5 negatively correlated switch genes (Fig. 14) and 5 different positively correlated switch genes (Fig. 15).

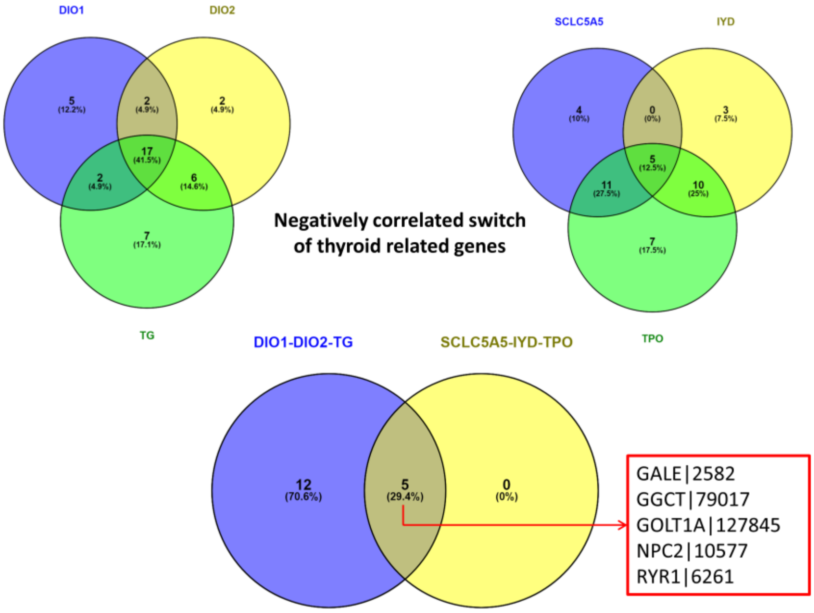


Figure 14 negatively corelated switch of thyroid related genes

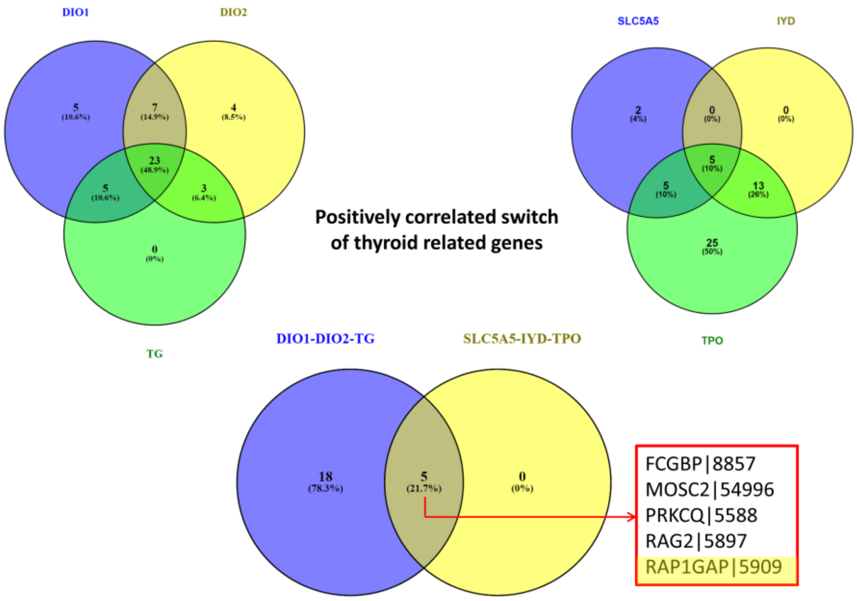


Figure 15 positively corelated switch of thyroid related genes

Among the positively correlated switch genes, we selected RAP1GAP, down-regulated in the analyzed cancer tissues (Fig.16), that is mainly expressed in human thyroid related tissues (Fig. 17A) both for concerning RNA and protein expression levels (Fig. 17B).

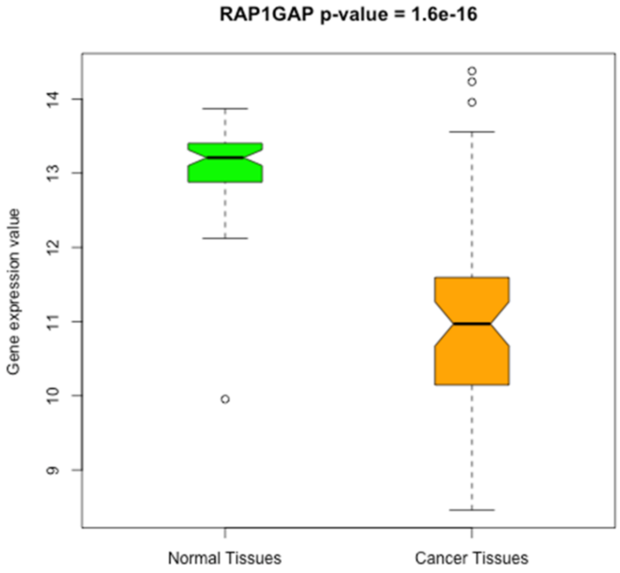


Figure 16 Boxplot of RAP1GAP in normal and cancer tissues.

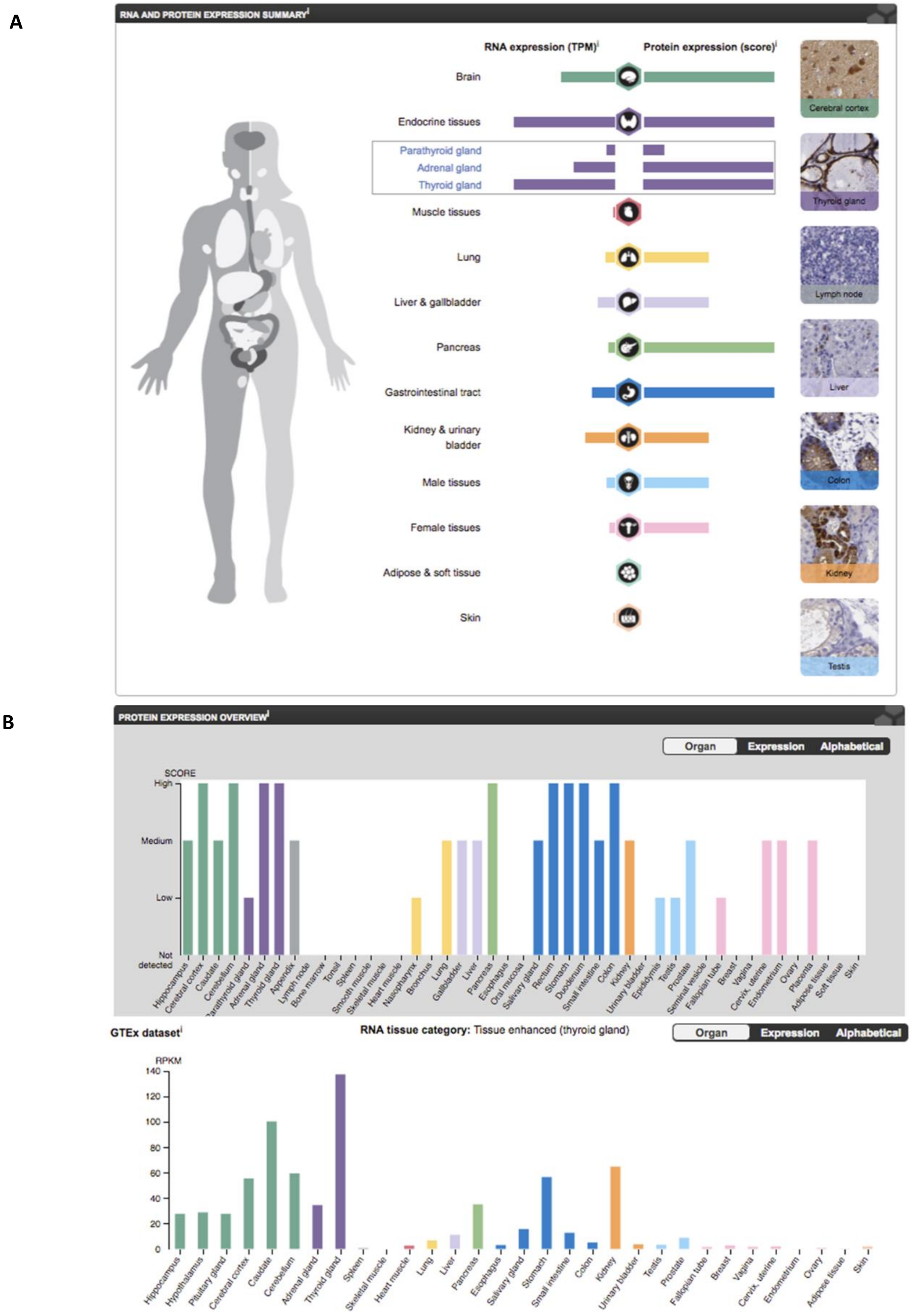


Figure 17 Expression of RAP1GAP in human tissues.

References

[1] Cancer Statistics Review, 1975-2014 - SEER Statistics. Available at: https://seer.cancer.gov/csr/1975_2014/. (Accessed: 18th September 2017)

[2] Tuttle, R. M. *et al.* Thyroid carcinoma. *J. Natl. Compr. Cancer Netw. JNCCN* **8,** 1228–1274 (2010).

[3] Xing, M. Molecular pathogenesis and mechanisms of thyroid cancer. *Nat. Rev. Cancer* **13,** 184–199 (2013).

[4] Agrawal, N. *et al.* Integrated Genomic Characterization of Papillary Thyroid Carcinoma. *Cell* **159,** 676–690 (2014).

[5] Cohen, Y. *et al.* BRAF mutation in papillary thyroid carcinoma. *J. Natl. Cancer Inst.* **95,** 625–627 (2003).

[6] Garcia-Rostan, G. *et al.* RAS mutations are associated with aggressive tumor phenotypes and poor prognosis in thyroid cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **21,** 3226–3235 (2003).

[7] Xing, M. Genetic Alterations in the Phosphatidylinositol-3 Kinase/Akt Pathway in Thyroid Cancer. *Thyroid* **20,** 697–706 (2010).

[8] Gene expression profiling predicts clinical outcome of breast cancer : Article : Nature. Available at: http://www.nature.com/nature/journal/v415/n6871/full/415530a.html?foxtrotcallback=true. (Accessed: 25th September 2017)

[9] Jarzab, B. *et al.* Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications. *Cancer Res.* **65,** 1587–1597 (2005).

[10] Vasko, V. *et al.* Gene expression and functional evidence of epithelial-to-mesenchymal transition in papillary thyroid carcinoma invasion. *Proc. Natl. Acad. Sci.* **104,** 2803–2808 (2007).

[11] Huang, Y. *et al.* Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. *Proc. Natl. Acad. Sci.* **98,** 15044–15049 (2001).

[12] Brennan, K. *et al.* Development of prognostic signatures for intermediate-risk papillary thyroid cancer. *BMC Cancer* **16,** 736 (2016).

[13] Ramaswamy, S. *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 15149–15154 (2001).

[14] Chibon, F. Cancer gene expression signatures - the rise and fall? *Eur. J. Cancer Oxf. Engl. 1990* **49,** 2000–2009 (2013).

[15] Weinstein, John N., et al. “The cancer genome atlas pan-cancer analysis project”. Nature genetics, 2013. 45(10): p. 1113-20.

[16] Tomczak, Katarzyna, Patrycja Czerwińska, and Maciej Wiznerowicz. “The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge”. Contemporary oncology,2015. 19(1A):p. A68-A77.

[17] Paola Paci, Colombo T, Fiscon G, Gurtner A, Pavesi G, Farina L. “SWIM: a computational tool to unveiling crucial nodes in complex biological net- works”. Scientific Reports (2017), 7, Article number: 44797

[18] Hartigan JA, Wong MA “Algorithm AS 136: A K-Means Clustering Algorithm” JR Stat Soc Ser C Appl Stat 1979, 28:pp.100–108.

[19] Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M “Evidence for dynamically organized modularity in the yeast protein-protein interaction network”. Nature 2004, 430:pp.88-93.

[20] Guimera R, Amaral LAN “Functional cartography of complex metabolic networks”. Nature 2005, 433:pp.895.

[21] Kanehisa, M, Sato Y, Kawashima M, Furumichi M, Tanabe M. “KEGG as a reference resource for gene and protein annotation”. Nucleic Acids Res 2016, 44: pp. 457-462.

[22] D’Andrea D, Grassi L, Mazzapioda MG, Tramontano A “FIDEA: a server for the functional interpretation of differential expression analysis” Nucleic Acids Res 2013.

[23] Zambelli, F., Pesole, G, Pavesi, G. “Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes”. Nucleic Acids Res 37, W247–W252 (2009).

[24] Fischer, M. “Census and evaluation of p53 target genes”. Oncogene (2017). doi:10.1038/onc.2016.502.