Identifying Diagnostic and Prognostic targets for Papillary Thyroid Carcinoma through mining Gene Expression *BIG* Datasets using Adaptive Filtering and Advanced Bioinformatics Algorithms

Asma Almansoori [1 #], Poorna Manasa Bhamidimarri [1 #], Riyad Bendardaf [2,3], Rifat Hamoudi [1,2,4]

[1] Sharjah Institute for Medical Research, University of Sharjah, P.O. Box 27272, Sharjah, United Arab Emirates, [2] College of Medicine, University of Sharjah, P.O. Box 27272, Sharjah, United Arab Emirates, [3] Department of Oncology, University Hospital Sharjah, Sharjah, United Arab Emirates, [4] Division of Surgery and Interventional Science, University College London, London, United Kingdom.

Emails: U17104342@sharjah.ac.ae, pbhamidimarri@sharjah.ac.ae, Riyad.Bendardf@uhs.ae, rhamoudi@sharjah.ac.ae

# Authors contributed equally to this work

Abstract

Thyroid Cancer is the most common endocrine malignancy. Although the mortality rate of thyroid cancer is considered to be low, however the reoccurrence and persistence of the disease is still considered high. The most common type of thyroid cancer is papillary thyroid carcinoma consisting of >70% of all types of thyroid cancer. Thyroid cancer is heterogeneous and complex. *BIG* data in the form of publicly available gene expression (transcriptomics) datasets can provide valuable source to gain deeper understanding of complex diseases such as papillary thyroid carcinoma (PTC). In this study, we used a novel bioinformatics method based on adaptive filtering to reduce the number of genes expressed eliminating genes that are invariant across the various disease stages. In order to shed light on some of the mechanisms involved in PTC, the filtered genes were used in systematic pathway analysis searches across 20,500 annotated cellular pathways using modified Kolmogorov-Smirnov algorithm to identify the relevant differentially activated cellular pathways across the various stages of the disease. Our analysis from 95 PTC patient biopsies consisting of 41 normal, 28 non-aggressive and 26 metastatic papillary thyroid carcinoma revealed 2193 differential activated cellular pathways among non-aggressive samples and 1969 among metastatic samples compared to normal tissue. The key pathways for non-aggressive PTC includes calcium and potassium ion transport, hormone signaling pathways, protein tyrosine phosphatase activity and protein tyrosine kinase activity. The key pathways for metastatic PTC include growth, apoptosis, activation of MAPK activity and regulation of serine threonine kinase activity. The most frequent genes across the enriched pathways were KCNQ1, CACNA1D, KCNN4, BCL2, and PTK2B for non-aggressive PTC, and EGFR, PTK2B, KCNN4 and BCL2 for metastatic PTC. Survival analysis results showed that PTK2B, CACNA1D and BCL2 contributed to poor survival of PTC patients. The study identified insights into mechanisms of PTC.

Keywords: Microarray, BIG data analytics, absolute GSEA, adaptive filtering, Kolmogorov-Smirnov, transcriptomics, normalization

I. INTRODUCTION

Thyroid cancer was ranked as the most common endocrine malignancy [1] and the 9th most common cancer overall. Globally, the incidence has been on the rise over the past three decades. Between 2006 and 2012, the annual incidence rate was 6.5% in women and 5.4 in men [2,3]. In the United States, thyroid cancer incidence rate was the highest among all cancers between 2000 and 2009 [4]. The mortality rate of thyroid cancer is considered to be low at an estimated 44,000 deaths in both sexes combined, however the re-occurrence and persistence of the disease is considered high [5].

Morphologically, thyroid cancers are classified into different cellular subtypes such as papillary, follicular, medullary and anaplastic. Differentiated papillary thyroid carcinoma (PTC) form is the most common type and comprise more than 70% of all thyroid cases.

Multiple reasons might play role in hindering the understanding of the molecular mechanism of papillary thyroid cancer including the fact there are many environmental and physiological trigger that may work synergistically with the genetic profile of the individual

in initiating the cancer as well as the fact that response to therapy in each patient differ due to the intra-tumoural heterogeneity of thyroid cancer [6].

Many OMICs data were carried out on thyroid cancer, however most of those studies tended to be genomics by nature focusing on mutational screening of thyroid cancer patients. Since the highest percentage of thyroid cancer subtype is PTC, few transcriptomic analyses were carried out on PTC identifying some of the pathways involved in its pathogenesis [7] . However, such studies were carried out on small number of patients using classical bioinformatics analysis that provided limited insights into the molecular basis of PTC and did not identify clear diagnostic, prognostic targets.

In this study, we carried out comprehensive in silico pathway analysis using adaptive filtering and advanced bioinformatics pipeline that has shown good ability to decouple the transcriptomic profiles between different stages of the same disease [8]. Most of the publicly available transcriptomics datasets were made using microarrays that can detect thousands of gene expression values simultaneously. Microarray technology presents the expression level of a gene as pixel intensity. In order to identify key genes, series of complex computational calculations needs to be carried out. Firstly, the scanned images is converted to pixel values and mapped to the specific genes using the coordinates on the microarray. Secondly, the gene expression values from all the samples undergoes global normalization algorithms in order to make them comparable to each other. Thirdly, in order to identify true differentially expressed genes (DEG) from the thousands of genes, set reduction is carried out using adaptive filtering techniques and finally the filtered dataset undergoes further reduction by applying systematically to well annotated pathways and for each pathway identify whether there is enrichment or not based on probabilistic techniques.

In this study, we applied this advanced bioinformatics analysis that was previously validated on different cancer dataset and showed remarkable performance in uniquely stratifying different stages from the same cancer [8] to filter the publicly available thyroid cancer datasets. We aimed to identify novel cellular pathways and putative diagnostic and prognostic targets that may shed light on some of the molecular mechanisms involved in the initiation and progression of papillary thyroid carcinoma.

I. HYPOTHESIS

Using a combination of adaptive filtering and advanced bioinformatics algorithms to search for significantly activated cellular pathways can shed light on explaining some of the mechanisms of PTC provide better identification of true differentially expressed genes between different stages of PTC.

A. Aim

The aim of this study is to attempt to identify the key molecular targets by identifying the transcriptomic signature patterns that drive non-aggressive and metastatic PTC. Such an approach can provide insights into some of the molecular mechanisms involved in PTC progression.

B. Objectives

- Collect raw image gene expression microarray data
- Carry out two different normalization routines on the microarray data
- Apply two different filtering techniques on the normalized data; variation filter and co-efficient of variation-based filter followed by intersection of the genes from both filters. This is used to reduce the dataset
- Identify significantly differentially activated cellular pathways using probabilistic model mining 20,500 well annotated cellular pathways
- Identify differentially expressed gene (DEG) from each pathway
- Validate the DEGs identified in larger cohort of PTC patients using survival analysis algorithms

II. METHODOLOGY AND RESULTS

A. Publicly available data sets for papillary thyroid carcinoma

Gene sets available in gene expression omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) in association with papillary thyroid carcinoma was retrieved. Datasets inclusive of patient matched normal thyroid tissue transcriptomics were considered for analysis. In order to analyze the data with similar Affymetrix array, data sets obtained from Affymetrix Human Genome U133 Plus 2.0 Array was considered. The three gene sets conforming to these criteria were GSE6004, GSE60542, and GSE3678. The raw CEL files corresponding to these gene sets were extracted and further processed for Gene Set Enrichment Analysis. The analysis for this study was approved by the Research Ethics Committee; the ethical approval number of the study is UHS-HERC-011-10062019.

B. Raw microarray analysis and normalization

The raw CEL files (n=95) obtained from the GEO for normal, non-aggressive and metastatic thyroid samples were processed for normalization using in house R script as described previously [8]. Two normalization algorithms; the Affymetrix microarray suite 5 (MAS5) and Gene Chip Robust Multiarray Averaging (GCRMA) implemented in Bioconductor using R software were applied to the transcriptomic data to normalize and remove the background noise. The non-variant probes were removed from the transcripts list, and adaptive filtering was performed to obtain the common set of

variant probes. The flow chart for the process of normalization and filtration is shown in figure 1.

C. Adaptive filtering of the microarray data

In order to reduce the gene set further, combination of variation filtering [9] based on raw values and variance filtering based on co-efficient of variation was applied to the normalized data. Probes with MAS5 value > 50 and coefficient of variation (CV) 10-100% in GCRMA among all the cases were passed and intersected to obtain probes with common variant set. The filtered and intersected transcripts from all the samples were genes using the Broad Institute software [10]. The probes with maximum expression for each gene were chosen excluding the housekeeping genes and the probes not assigned to any gene. From the total number of 54,675 probes in the Affymetrix Human Genome U133 Plus 2.0 Array, following MAS5 and GCRMA filter 15801 probes were extracted. These filtered probes were mapped to a total of 9394 genes which were used in absolute GSEA as mentioned next.

D. Absolute Gene Set Enrichment Analysis

Absolute Gene Set Enrichment Analysis (GSEA) is based on modification of the Kolmogorov-Smirnov algorithm [11] where the expression values for genes related to each pathways are placed in a list and tested using by taking one gene out a time and running the test to check if they are distributed randomly or there is significant enrichment in the genes at either end of the list. This is done across all genes in that pathway which will generate probability measure to see if the pathway is significantly activated across the different groups being test. In this study, the 9394 gene list obtained was further reduced by estimating the activated and enriched pathways in non-aggressive (NAG) and metastatic papillary thyroid carcinoma (PTC) samples in comparison to normal tissue. GSEA search was carried out using around 20,500 annotated cellular pathways obtained across seven gene ontology sets C1 to C7 from the database (https://www.gsea-msigdb.org). The significantly activated pathways in different types of PTC samples were selected based on $p < 0.05$ and FDR < 0.25 as previously described [8,12] The pathways selected were further processed to identify differentially enriched genes between the normal versus non-aggressive and normal versus metastatic PTC cases. This was followed by reducing the set of available genes by identifying the frequency of gene occurrence across different activated cellular pathways.

The three datasets normal, non-aggressive and metastatic papillary thyroid cancer samples were further processed for GSEA. The differentially activated significant pathways across the three different samples were identified by comparing the cancer samples with normal tissue. Based on the nominal $p < 0.05$ and false discovery rate (FDR) < 0.05, highly significant pathways were recorded. From the molecular functions and biological

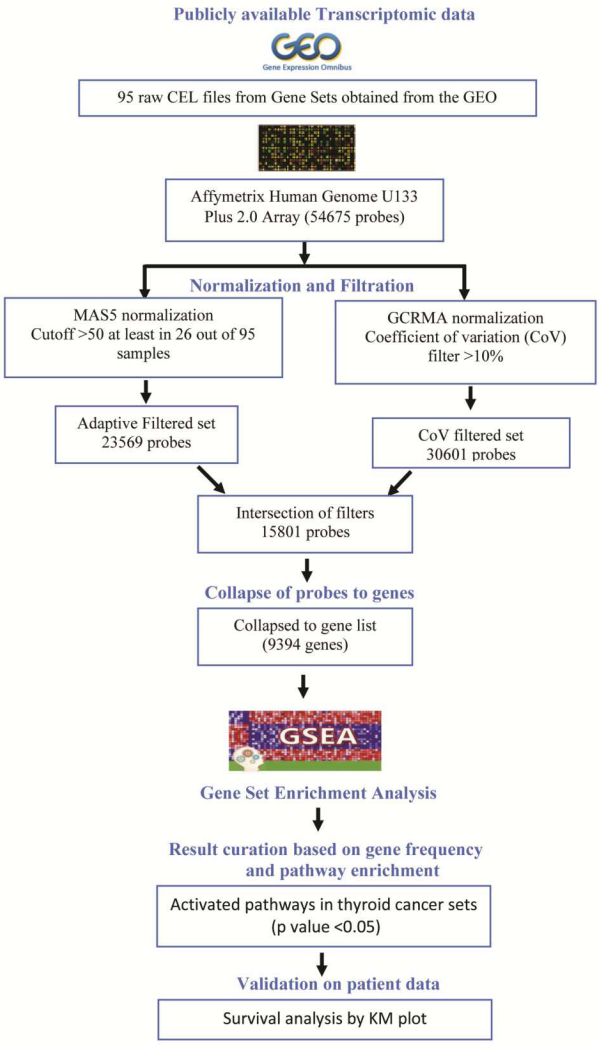

Figure 1. *Flow chart illustrating the method followed in the study*

processes ontology gene set, ~1795 significant pathways were identified. The most significantly enriched pathways include calcium and potassium ion transport pathways, protein tyrosine kinase and phosphatase activity in normal versus non-aggressive set as shown in table 1. Amongst the normal versus metastatic set, apoptosis, activation of MAPK activity, regulation of protein serine threonine kinase activity and transmembrane receptor protein tyrosine kinase signaling pathway were among the significantly enriched pathways as shown in table 1. Example representation of the output from the absolute GSEA for each data set is shown in figure 2 showing the graph for enrichment score and the heatmap of the expression values of the genes.

Table 1. *List of the pathways activated in thyroid cancer samples in comparison to normal thyroid tissue analyzed by GSEA*

Gene Set	SIZE	ES	NES	NOM p-val	FDR q-val
Go_regulation_of_ion_transport	298	0.489	2.09	<0.0001	0.004
Go_regulation_of_membrane_potential	173	0.542	2.333	<0.0001	0.001
Go_intracellular_receptor_signaling_pathway	161	0.416	1.835	<0.0001	0.02
Go_hormone_transport	156	0.469	2.056	<0.0001	0.005
Go_positive_regulation_of_growth	146	0.481	1.963	<0.0001	0.01
Go_regulation_of_wnt_signaling_pathway	221	0.421	1.73	0.004	0.031
Go_transmembrane_receptor_protein_serine_threonine_kinase	198	0.471	1.846	0.004	0.018
Go_response_to_transforming_growth_factor_beta	161	0.437	1.72	0.006	0.032
Go_positive_regulation_of_map_kinase_activity	169	0.451	1.725	0.01	0.032
Go_cell_cycle_arrest	141	0.374	1.63	0.014	0.049
Go_regulation_of_apoptotic_signaling_pathway	256	0.375	1.605	0.018	0.054

Abbreviations: ES, enrichment score; NES, normalized ES; NOM, nominal; FDR, false discovery rate.

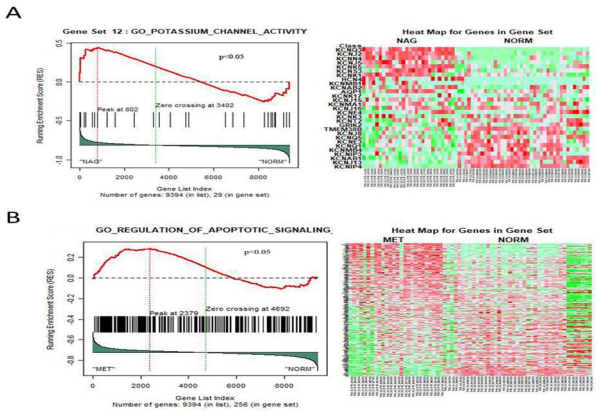

Figure 2. *GSEA output A. for Non-aggressive set B. for Metastatic set*

E. Differential gene expression in PTC samples compared to normal thyroid tissue

The differential gene expression analysis was addressed in two methods to obtain the information both based on pathway enrichment and microarray expression.

In the first approach, the significantly enriched pathways for each sample set were used to obtain genes occurring frequently in all the enriched pathways using R script as described previously [8] . Statistical analysis was applied for the gene frequency values obtained and a 95-percentile cut-off value was calculated in each sample set.

The enriched pathways from GSEA were subjected to gene frequency cutoff using the 95-percentile as a cut-off. Gene frequency can be defined as the number of times a gene occurs across all the enriched gene component from the significantly activated cellular pathways. This type of analysis showed the value for the frequency for non-aggressive (NAG) to be 13 and metastatic (MET) to be 10. Based on those frequency cutoff values, the number of genes with frequency higher than the cutoff in NAG was 355 and in MET was 280. The top 40 genes based on frequency cutoff were shown in figures 3A and 3B.

The frequency analysis identifying the occurrence of each gene in more than one pathway identified BCL2, CACNA1D, KCNQ1, KCNN4, EGFR and PTK2B as important in the progression of PTC and thus can be considered as putative molecular targets.

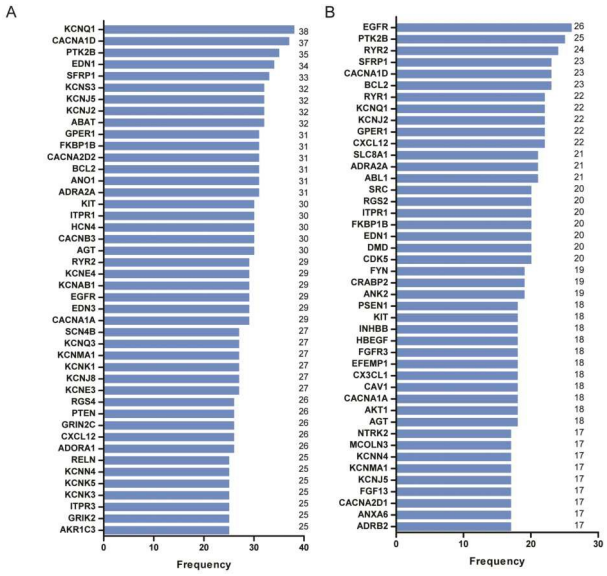

Figure 3. *Gene frequency histogram in A) non-aggressive group and B) metastatic group*

F. Validation of the pathways and genes identified by GSEA in independent cohort

Metascape Analysis

To validate the pathways identified by GSEA, the high frequent genes from non-aggressive and metastatic samples were considered. The commonly occurring high frequent genes among both the sample types were input in the Metascape [13].The most frequent genes across the enriched pathways identified using the absolute GSEA were used to validate the cellular pathways activated between NAG and MET samples in comparison to normal samples. The validation was carried out using Metascape to search for well pathways annotated according to the Gene Ontology format [14,15]. The analysis revealed key cellular pathways such as calcium ion transport, positive regulation of protein phosphorylation and signaling by receptor tyrosine kinase were enriched in non-aggressive PTC as shown in figure 4A. In the metastatic PTC, in addition to the pathways identified in NAG group, other pathways related to caner hallmark are identified those include apoptosis and growth signaling pathways as shown in figure 4B.

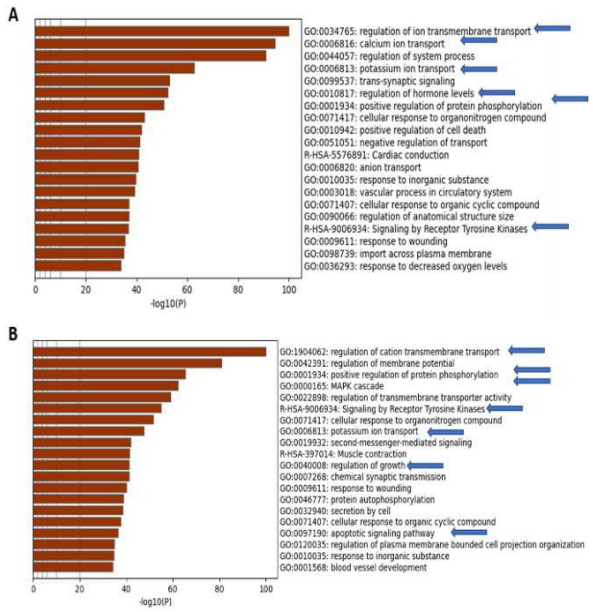

Figure 4. *Metascape analysis for the high frequent genes from A. normal and non-aggressive set and B. normal versus metastatic set*

Survival analysis of the identified genes

The survival analysis for the differentially expressed genes identified in the PTC samples was performed was carried out using the Kaplan-Meier algorithm via KMplotter [16] where an independent cohort of 502 thyroid cancer patients' data was used. Survival analysis of the differentially enriched genes showed that PTK2B, CACNA1D and BCL2 contribute to poor survival in an independent cohort of 502 thyroid cancer patients as shown in figure 5.

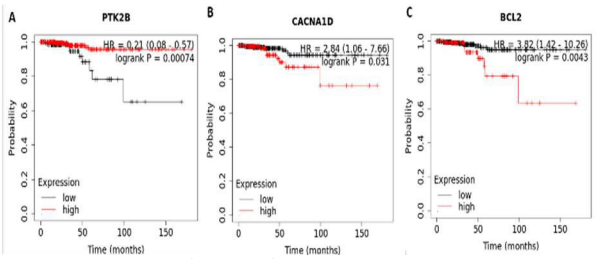

Figure 5. *KM Plot for high frequent genes*

Ion transport and tyrosine kinase and protein phostphatase pathways are involved in PTC progression

Using the transcriptomics *BIG* data mining, this study identified using advanced bioinforamtics techniques the unique cellular pathways in non-aggressive and metastatic PTC. Interestingly, many of the genes and pathways overlapped between the two entities, these include calcium and potassium ion transport and tyrosine kinase and protein phosphatase pathways. The NAG group showed more unique association with regulation of hormone levels and cell signalling related to hormones whereas the study identified more impact of MAPK activation as well as activation of other cancer hallmark pathways such as regulation of apoptosis and cell growth in the metastatic pathways. Overall, pathway analysis indicated that PTC is highly complex disease with high level intra-tumoral heterogeneity. DEG analysis results identified the following target genes linked to PTC initiation and progression: BCL2, CACNA1D, KCNQ1, KCNN4, EGFR and PTK2B.

BCL2

B cell Lymphoma-2 (BCL2) is well know anti-apoptotic protein responsible for inhibiting programmed cell death or apoptosis [17]. Aksoy, M., et al. found that lower BCL2 expression in thyroid cancer supports the formation of oncoytic neoplasms in early thyroid cancer stages by inhibiting apoptosis of tumor cells [18]. This finding from this study supports the results obtained in our study where BCL2 overexpression infers poor survival for thyroid cancer patients ($p < 0.01$), and the frequency search showed that BCL2 is present in both the NAG and metastatic groups, therefore BCL2 may be putative marker for early PTC.

CACNA1D

One of the recurrent activated pathway identified from this study is related to ion transport and more specifically calcium and potassium transport. Many genes related to calcium and potassium transcrport were identified. CACNA1D gene is responsible for regulating positively charged calcium chanels (CaV1.3) across cell membranes and specifically adrenal gland to form alpha-1 subunit. These subunits are involved in the regulation of adrenal hormones production such as aldosterone which maintains blood pressure and fluid balance in the body [19,20]. Cancer cells can undergo oncogenic switch by transforming apopotosis inducing Ca influx pathway to proliferative calcium influx which in turn can promote growth and apoptosis resistance in cancerous cells [21]. This was also confirmed by the fact that pathway analysis showed the activation of calcium ion transport pathways in both NAG and metastatic PTC. The results from this study, showed that CACNA1D overexpression infers poor survival in thyroid cancer patients ($p < 0.05$) suggesting that it may be a putative prognostic marker for PTC progression.

PTK2B

Another protein identified is Protein tyrosine kinase 2 beta (PTK2B). This has multiple functions including regulator of cell growth, survival, proliferation and invasion [21]. It encodes a cytoplasmic protein tyrosine kinase that is involved in calcium-induced regulation of ion channels and activation of the MAP kinase signaling pathway. Methylated PTK2B favouring overexpression is linked to c-Src activation, development of Pyk2/c-Src complex and the activation of ERK/MAPK signaling pathway [22]. The survival analysis in the current study showed that PTK2B gene silencing lead to poor survival in thyroid cancer patients ($p < 0.0001$). Few studies have shown some links between EGFR and PTK2B. The

current study showed metastatic samples were enriched with EGFR and PTK2B genes and the combination might be effective in treating metastatic PTC. Therefore, since PTK2B is linked to EGFR, MAP kinase activation and calcium ion transport, it is probably an attractive theraputic target and since it is linked with poor survival it can be a good prognostic biomarker. This study shows the value of using computational approaches based on systems engineering in generating global solutions to reduce the genetic noisy dataset to identify key targets associated with PTC. The methodology in this study can be used for other complex diseases providing deeper insights into their mechanisms.

IV. CONCLUSIONS

Taken together, the adaptive filtering following by absolute GSEA managed to reduce the set of genes based on search of cellular pathways. The differentially activated cellular pathways and genes from this study showed the involvement of ion transport as well as other cancer related pathways including tyrosine kinase and protein phosphatase in the initiation of PTC during the non-aggressive phase and further progression to PTC metastatic phase. Understanding of differentially activated pathways during carcinogenesis, invasion and metastasis can have significant clinical outcome in developing better prognostic assays and molecular inhibitors that can replace classic generalized PTC treatments. DEG analysis of transcirptomics data identified putative diagnostic and prognostic target genes including EGFR, PTK2B, KCNQ1, KCNN4, BCL2 and CACNA1D which might be involved in key mechanisms of thyroid cancer pathogenesis. The survival analysis showed that BCL2, CACNA1D and PTK2B infer poor survival and are thus putative diagnostic and prognostic targets for PTC.

V. ACKNOWLEDGEMENTS

We would like to acknowledge the University of Sharjah for supporting this work.

VI. REFERENCES

1. Xing M (2013) Molecular pathogenesis and mechanisms of thyroid cancer. Nature reviews Cancer 13: 184-199.
2. Miranda-Filho A, Lortet-Tieulent J, Bray F, Cao B, Franceschi S, et al. (2021) Thyroid cancer incidence trends by histology in 25 countries: a population-based study. The Lancet Diabetes & Endocrinology 9: 225-234.
3. Cramer JD, Fu P, Harth KC, Margevicius S, Wilhelm SM (2010) Analysis of the rising incidence of thyroid cancer using the Surveillance, Epidemiology and End Results national cancer data registry. Surgery 148: 1147-1152; discussion 1152-1143.
4. Noone A-M, Cronin KA, Altekruse SF, Howlader N, Lewis DR, et al. (2017) Cancer Incidence and Survival Trends by Subtype Using Data from the Surveillance Epidemiology and End Results Program, 1992-2013. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 26: 632-641.
5. Tuttle RM, Ball DW, Byrd D, Dilawari RA, Doherty GM, et al. (2010) Thyroid carcinoma. J Natl Compr Canc Netw 8: 1228-1274.
6. Fugazzola L, Muzza M, Pogliaghi G, Vitale M (2020) Intratumoral Genetic Heterogeneity in Papillary Thyroid Cancer: Occurrence and Clinical Significance. Cancers (Basel) 12.
7. Teng H, Mao F, Liang J, Xue M, Wei W, et al. (2018) Transcriptomic signature associated with carcinogenesis and aggressiveness of papillary thyroid carcinoma. Theranostics 8: 4345-4358.
8. Hamoudi RA, Appert A, Ye H, Ruskone-Fourmestraux A, Streubel B, et al. (2010) Differential expression of NF-kappaB target genes in MALT lymphoma with and without chromosome translocation: insights into molecular mechanism. Leukemia 24: 1487-1497.
9. Sun J, Zhang Q, Wang F, Zhao X (2008) On the Generalization of Digital Total Variation Filter. IEEE Congress on Image and Signal Processing 2008: IEEE. pp. 5.
10. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102: 15545-15550.
11. Charmpi K, Ycart B (2015) Weighted Kolmogorov Smirnov testing: an alternative for Gene Set Enrichment Analysis. Stat Appl Genet Mol Biol 14: 279-293.
12. Hachim MY, Hachim IY, Elemam NM, Hamoudi RA (2019) Toxicogenomic analysis of publicly available transcriptomic data can predict food, drugs, and chemical-induced asthma. Pharmgenomics Pers Med 12: 181-199.
13. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, et al. (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 10: 1523.
14. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25-29.
15. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res 47: D419-D426.
16. Győrffy B (2021) Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. Computational and Structural Biotechnology Journal 19: 4101-4109.
17. Youle RJ, Strasser A (2008) The BCL-2 protein family: opposing activities that mediate cell death. Nature Reviews Molecular Cell Biology 9: 47-59.
18. Aksoy M, Giles Y, Kapran Y, Terzioglu T, Tezelman S (2005) Expression of bcl-2 in Papillary Thyroid Cancers and its Prognostic Value. Acta chirurgica Belgica 105: 644-648.
19. Azizan EA, Poulsen H, Tuluc P, Zhou J, Clausen MV, et al. (2013) Somatic mutations in ATP1A1 and CACNA1D underlie a common subtype of adrenal hypertension. Nat Genet 45: 1055-1060.
20. Scholl UI, Goh G, Stölting G, de Oliveira RC, Choi M, et al. (2013) Somatic and germline CACNA1D calcium channel mutations in aldosterone-producing adenomas and primary aldosteronism. Nat Genet 45: 1050-1054.
21. Monteith GR (2014) Prostate cancer cells alter the nature of their calcium influx to promote growth and acquire apoptotic resistance. Cancer Cell 26: 1-2.
22. Yoon S, Seger R (2006) The extracellular signal-regulated kinase: multiple substrates regulate diverse cellular functions. Growth Factors 24: 21-44.