Analyzing high throughput sequencing data for gene expression profiles in Thyroid Carcinoma to predict effective therapeutic targets

BY

Aqsa Qureshi

Fall-2018-MSTBI&00000274317

Supervised by

Dr. Rehan Zafar Paracha

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

in

Research Centre for Modeling & Simulation

September, 2020

Research Centre for Modeling and Simulation (RCMS)

National University of Sciences and Technology (NUST)

DECLARATION

I Aqsa Qureshi, hereby declared that work presented in this thesis is result of my own work except specific reference is made to the work of others wherever due. I also declared that the content presented in this thesis is original and have not been submitted in whole or in part to this university or to any other university for any other degree or qualification.

Aqsa Qureshi

September, 2020

Contents

Contents ii

List of Tables vi

List of Figures vii

ABSTARCT 1

1 INTRODUCTION 2
- 1.1 Thyroid Carcinoma . . . 2
  - 1.1.1 Molecular pathogenesis and mechanisms . . . 3
  - 1.1.2 Risk Factors . . . 5
  - 1.1.3 Prevalence . . . 6
- 1.2 Types of Thyroid Cancer . . . 7
  - 1.2.1 Papillary thyroid cancer (PTC) . . . 7
  - 1.2.2 Anaplastic thyroid cancer (ATC) . . . 8
  - 1.2.3 Follicular Thyroid Carcinoma (FTC) . . . 8
  - 1.2.4 Medelullary thyroid cancer (MTC) . . . 9
- 1.3 High Throughput Sequencing Techniques . . . 10
  - 1.3.1 Microarray Analysis . . . 11
  - 1.3.2 RNA sequencing . . . 12
- 1.4 Messenger RNA . . . 13
- 1.5 Micro RNA . . . 14
- 1.6 Pathway Analysis . . . 15

2 LITERATURE REVIEW 17
- 2.1 Role of miRNA, mRNA and lncRNA in Thyroid Carcinoma . . . 17
- 2.2 Role of miR-592 expression in thyroid carcinoma . . . 18
- 2.3 Role of miR-146b Expression in Thyroid Carcinoma . . . 19
- 2.4 Significant Differentially Expressed Genes . . . 20
- 2.5 Signalling Pathways and Networks . . . 21
- 2.6 Potential Therapeutic Targets and Biomarkers . . . 23
- 2.7 Study Rationale . . . 25
- 2.8 Problem Statement & Proposed Solution . . . 26
- 2.9 Objectives . . . 26

3 MATERIALS AND METHODS 27
- 3.1 Datasets Retrieval . . . 27
- 3.2 Microarray Analysis . . . 28
  - 3.2.1 maEndToEnd pipeline . . . 30
  - 3.2.2 Data Import . . . 30
  - 3.2.3 Quality Assessment . . . 31
  - 3.2.4 Preprocessing . . . 33
  - 3.2.5 Heatmap . . . 34
  - 3.2.6 Linear models . . . 34

3.2.7 Differential expression analysis 35
3.2.8 For Agilent platform 35
3.2.9 Volcano-plot 36
3.3 RNA sequencing Analysis 36
3.3.1 Galaxy Pipeline 37
3.3.2 Pathway Analysis 39

4 RESULTS 41
4.1 Results of Microarray Data Analysis 41
4.1.1 Microarray Dataset-1 (E-GEOD-65144) 41
4.1.2 Microarray Dataset 2 (E-GEOD-3467) 47
4.1.3 Microarray Dataset 3 (E-GEOD-40807) 52
4.2 Results of RNA-seq Data Analysis 58
4.2.1 RNA-seq Dataset 1 (E-GEOD-64912) 58
4.2.2 RNA-seq Dataset 2 (GSE57780) 61
4.3 Comparative Analysis of Microarray Datasets 70
4.4 Comparative Analysis of RNA-seq Datasets 70
4.5 Pathway Analysis 72
4.6 Microarray Datasets Pathway Analysis 72
4.7 RNA-seq Datasets Pathway Analysis 76

5 DISCUSSION 81

6 CONCLUSION AND FUTURE PERSPECTIVES 86

REFERENCES 87

Appendix A Source code of microarray analysis for Affymetrix 96

Appendix B Source Code for Ballgown-RNA-seq 104

Appendix C Source code of microarray analysis for Agilent platform 110

Nomenclature

Acronyms / Abbreviations

AA	Autunomous adenoma
ATC	Anaplastic Thyroid Carcinoma
BRAF	Serine-threonine kinase
CD4+T cells	Helper T cells
cDNA	Complementary Deoxyribonucleic acid
DEG	Differentially Expressed Genes
DTC	Distinct/Differentiated Thyroid Carcinoma
EGF	Epidermal growth factor
FA	Follicular adenoma
FDA	US Food and Drug Administration
FTC	Follicular Thyroid Carcinoma
GEO	Genome Expression Omnibus
GO	Gene Ontolgy
GSEA	Gene Set Enrichment Analysis
HTS	High Throughput Sequencing
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAPK	Mitogen activated protein kinase
MAPK	Mitogen-activated protein kinases
miRNA	Micro RNA
mRNA	Messenger RNA
MTC	Medullary Thyroid Carcinoma
P	Phosphorylation
PCA	Principal Component Analysis
PPI	Protein-Protein Interaction
PTC	Papillary Thyroid Carcinoma
RAF	Rapidly accelerated fibrosarcoma
RLE	Relative log expression
RNA	Ribonucleic Acid

RSeQC	Read duplication
RTKs	Receptor tyrosine kinases
SDRF	Sample and data relationship format
TC	Thyroid Carcinoma
Tg	Thyroglobulin
UTC	Undifferentiated Thyroid Carcinoma
WHO	World Health Organization

List of Tables

2.1 Genes implicated in thyroid tumorigenesis. 20

2.2 Drugs approved by FDA for thyroid cancer treatment. 25

3.1 Datasets for microarray analysis of thyroid carcinoma. 28

3.2 Sample information regarding RNA seq analysis. 28

4.1 Top 10 DEGs from dataset 1 (E-GEOD-65144). 47

4.2 DEGs of phenotype Normal vs PTC of E-GEOD-3746. 52

4.3 miRNAs of phenotype Normal vs MTC of E-GEOD-40807. 57

4.4 miRNAs targets of E-GEOD-40807 using miRWalk. 57

4.5 Top 10 DEGs of phenotype Normal vs Diseased of E-GEOD-64912. 61

4.6 Top 10 DEGs of phenotype Normal vs Tumor of GSE57780 (Normal vs Tumor). 65

4.7 miRNAs for Normal vs Tumor (GSE57780). 65

4.8 Top 10 DEGs of phenotype Normal vs Metastasis of GSE57780. . . 69

4.9 miRNAs for Normal vs Metastasis (GSE57780) 69

4.10 Pathways involved in DEGs obtained for dataset E-GEOD-65144. . 73

4.11 Pathways involved in DEGs obtained for dataset E-GEOD-3467. . . 74

4.12 Pathways involved in DEGs obtained for dataset E-GEOD-40807. . 75

4.13 Pathways in common DEGs for microarray. 76

4.14 Pathways involved in DEGs obtained for dataset E-GEOD-64912. . 77

4.15 Pathways involved in DEGs for dataset GSE57780 (Normal vs Tumor). 78

4.16 Pathways involved in DEGs for dataset GSE57780 (Normal vs Metastasis). 79

4.17 Pathways involved in common DEGs for RNA-seq analysis. 80

List of Figures

1.1 Molecular pathogenesis and thyroid cancer mechanisms. 3

1.2 Silencing of thyroid-specific genes in thyroid cells. 4

1.3 Radioactive iodine ablation and treatment. 4

1.4 Risk factors involved in thyroid carcinoma. 6

1.5 Overview of microarray technique. 11

1.6 RNA sequencing technique. 13

1.7 Biogenesis of messenger RNA's. 14

1.8 Biogenesis of micro RNA's. 15

3.1 General workflow of the study. 27

3.2 Workflow for microarray analysis. 30

3.3 Workflow for RNA-seq analysis. 36

4.1 PCA-plot of the raw expression data (E-GEOD-65144). 42

4.2 PCA-plot of the normalized data (E-GEOD-65144). 42

4.3 Boxplots for dataset 1 (E-GEOD-65144) 43

4.4 RLE plots of dataset 1 (E-GEOD-65144) 44

4.5 Heatmap of the summarised data (E-GEOD-65144). 45

4.6 Enhanced Volcano plot of the summarised data (E-GEOD-65144). . 46

4.7 PCA plots of dataset 2 (E-GEOD-3467) 48

4.8 Boxplots of dataset 2 (E-GEOD-3467) 48

4.9 RLE plots of dataset 2 (E-GEOD-3467) 49

4.10 Heatmap of the summarised data (E-GEOD-3467). 50

4.11 Enhanced Volcano Plot of the summarised data (E-GEOD-3467). . . 51

4.12 PCA plot of the summarised data (E-GEOD-40807). 53

4.13 RLE plot of the summarised data (E-GEOD-40807). 54

4.14 Boxplot of the summarised data (E-GEOD-40807). 55

4.15 Enhanced Volcano plot of the summarised data (E-GEOD-40807). . 56
4.16 RNA-seq dataset 1 (E-GEOD-64912) 59
4.17 Enhanced volcano plot of E-GEOD-64912. 60
4.18 RNA-seq dataset 2 GSE57780 (Normal vs Tumor) 63
4.19 Enhanced volcano plot of GSE57780 (Normal vs Tumor). 64
4.20 RNA-seq dataset 3 GSE57780 (normal Vs metastasis) 67
4.21 Enhanced volcano plot of GSE57780 (normal Vs metastasis). 68
4.22 Common DEGs among microarray datasets. 70
4.23 Common DEGs among RNA-seq datasets. 71
4.24 Common DEGs among RNA-seq and microarray datasets. 71

ABSTRACT

The most common endocrine tumor is thyroid carcinoma (TC). The clinical significance of thyroid carcinoma with respect to the recurrence of the disease state has been reviewed recently. Current therapy includes surgery (thyroidectomy) and radiotherapy that are not affordable and may indicate the risk of relapse. Although there are drugs available in the market; however, side-effects and drug-resistance limit their full potential to be used. Since the expression analysis identifies important cellular processes or metabolic pathways which are important during the phase of infection. Therefore, identifying effective therapeutic targets through microarray and high throughput sequencing technology might serve a purpose in the treatment of the thyroid carcinoma in its early stages. In order to achieve the objectives of the study, Microarray and RNA-seq data analysis have been performed. We analyzed different datasets of thyroid carcinoma induced in order to find similarities and differences between expression profiles. After identification of expression level of mRNAs and miRNAs, targets of miRNAs are also predicted. The data analysis has revealed 36 common differentially expressed genes (DEGs) for thyroid carcinoma. Out of these genes, only (Zinc finger and BTB domain containing protein 44) ZBTB44 is not considered a prognostic therapeutic target for thyroid cancer but for other carcinomas patients in literature, which needs further investigation to overcome the disease. While remaining differentially expressed genes are also validated through literature review. Pathway analysis is then performed on the all DEGs that shows their involvement in following pathways; Proteoglycans in cancer, Transcriptional misregulation in cancer, PI3K-AKT signaling pathway, WNT signalling pathway and MAPK signaling pathway. This study can provide the basis for further validation through systems biology approach and wet lab techniques.

Chapter 1

INTRODUCTION

1.1 Thyroid Carcinoma

Now the most commonly diagnosed tumor is thyroid cancer, which is on the upswing, presumably due to the massive use of prevalent imaging research. It provides a useful model for the other cancers since there are distinct histological features in its different forms. At the time of diagnosis, the treatment and the prognosis of thyroid carcinoma varies based on the nature of tumor and its stage (Nikiforov, 2008). The thyroid lesions may come through cells of parafollicular C or follicular (thyrocytes). Medullary thyroid carcinoma is the only tumor type that originates via parafollicular C cells. They constitute a very small portion (2-4 per cent) of all thyroid tumors. While, various types of thyroid malignancies originate through follicular cells. Every group has unique molecular, medical, and histopathological requirements. There are two distinct types of benign tumour: follicular adenoma (FA) and autonomous adenoma (AA). They are usually associated with production of thyroid hormones and high and low ability to take up iodide, respectively (Saiselet et al., 2016).

In the united states, thyroid cancer is the most prevalent cancer in women, and in 2015, an estimated 62,000 new cases occurred in both men and women (Cabanillas et al., 2016). Most clinicians would also meet a patient with that kind of illness at a certain stage in their professional life. While the prevalence is gradually rising, thyroid cancer mortality has minimally evolved across the previous 5 decades. The complexity confronted by doctors who diagnose thyroid carcinoma is to manage the treatment approach such that individuals with mild-risk disease or stable thyroid nodules are not mis-treated. At a certain period, individuals with more severe or high-risk illness need to be identified who require a more intensive care plan. Thyroid

carcinoma in most cases display a spectrum of clinical action abroad from indolent tumors with low mortality to very severe malignancies, such as anaplastic thyroid cancer. To tailor treatment appropriately, it is therefore crucial to undertake a proper diagnostic workup before treatment is started (Cabanillas et al., 2016).

1.1.1 Molecular pathogenesis and mechanisms

The molecular pathogenesis and thyroid cancer mechanisms arise when the extracellular mitogenic stimulus or growth factor activates the receptor tyrosine kinases (RTKs), resulting in the intracellular domain receptor dimerisation and activation by autophosphorylation of tyrosine residues shwn in Figure 1.1 (Xing, 2013). Through a sequence of adaptor proteins, the activated receptor contributes to stimulation of receptor tyrosine kinases (RAS) situated at the inside of the cell membrane. The stimulated RAS binds to the plasma membrane and employs the (rapidly accelerated fibrosarcoma) RAF proteins. The activated serine-threonine kinase (BRAF) is then phosphorylated which further activated the mitogen-activated protein kinase (MAPK/ERK) or (MEK). In nucleus, ERK is triggered through phosphorylation (P) where the tumor suppressor genes and tumor- promoter genes are upregulated and thyroid iodide driving genes are downregulated (Xing, 2013).

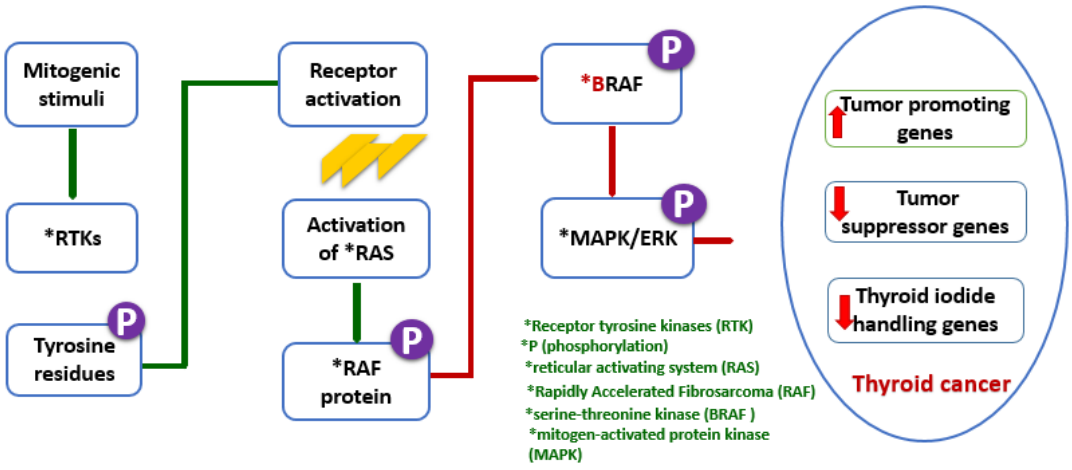


Figure 1.1. Molecular pathogenesis and thyroid cancer mechanisms.

In Figure 1.2; The BRAF mutation activates the MAPK pathway, which is

highly prevalent in PTC. In comparison, RAS/PTEN mutation found in the remaining categories of thyroid cancer (FTC, ATC, MTC) triggers another pathway P13-ACT. Both pathways caused the thyroid –specific genes to be silenced and the iodide handling machinary turned off. Iodide uptake in the thyroid cells is therefore decreased, resulting in an increasing loss of radioiodine therapy (Xing, 2013).

Figure 1.2. Silencing of thyroid-specific genes in thyroid cells.

The treatment procedure involves the surgical removal of the entire thyroid gland (thyroidectomy) which removes all visible thyroid tissue. Still, following surgery, the re-occurrence of the state of the disease was noticed. Hence, radioactive iodine treatment has been suggested to cure the possible relapse. In this mechanism, iodide function by entering thyroid cells through the transporters of sodium iodide and emitting beta-rays of short wavelength causing acute cell death (Figure 1.3).

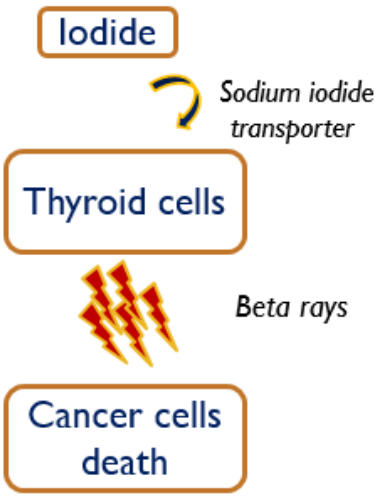


Figure 1.3. Radioactive iodine ablation and treatment.

The therapy has recently been approved for multi-kinase or tyrosine kinase

inhibitors. Though no overall survival impact has yet been seen. Since they also have severe health effects and must be restricted just for individuals with deadly illnesses (Nguyen et al., 2015)

1.1.2 Risk Factors

The risk factors include;

- Hereditary and epigenomic variations are the primary factors of thyroid cancer. Common sources of these mutations include the abnormalities in anaplastic lymphoma kinase (ALK), isocitrate dehydrogenase 1 (IDH1), TP53, β-catenin (CTNNB1), translocations (RET – PTC and paired box 8 (PAX8)–peroxisome proliferator-activated receptor-π (PPARG)), BRAF (BRAFV600E), RAS, PIK3CA, PTEN, and incorrect gene methylation (Xing, 2013).

- Molecular pathogenesis of thyroid carcinoma focuses on different signaling routes, such as MAPK, PI3K–AKT, RASSF1–mammalian STE20-like protein kinase 1 (MST1) forkhead box O3 (FOXO3), WNT–β-catenin, hypoxia-inducing factor 1α (HIF1α) and TSH-receptor (TSHR) (Xing, 2013).

- The progression of thyroid cancer is a mechanism for developing hereditary and epigenetic mutations with associated progressive aberration in the signaling pathway. All of these are accompanied by multiple secondary molecular changes that intensify and harmonize their effect on thyroid tumor origin, both in the cell and in the tumor microenvironment (Xing, 2013).

- Abnormal inhibition of thyroid iodide-driving genes and subsequent destruction of thyroid cancer radioactive iodine avidity induced by BRAF-V600E is a particular molecular pathology mechanism in thyroid carcinoma that induces rejection of radioiodine therapy (Xing, 2013).

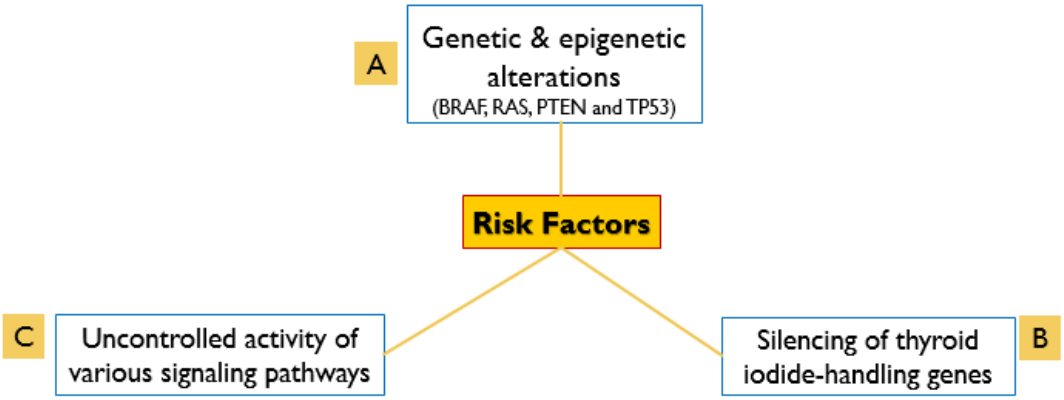


Figure 1.4. Risk factors involved in thyroid carcinoma.

1.1.3 Prevalence

Thyroid carcinoma (TC) accounts for around 1 per cent of all reported pathologies globally, by around 140,000 occurrences and 35,000 fatalities reported in 2002. Many thyroid cancers are avaricious malignant tumors with a median survival measured at approximately 85% in Europe and 95% in the United States for five years. The combined total likelihood of suffering cancer in women before 65 years of age was estimated at 0.2 per cent and in men at 0.1 per cent (Dal Maso et al., 2009). The American National Cancer Institute recorded that the median age at diagnosis for thyroid cancer in the years 2005–2009 was 50 years, with an estimated 56,460 new cases and 1,780 deaths from thyroid cancer in the United States in 2012 (Liu et al., 2013). Despite the positive prognosis of this disease, 15–20 percent of distinct cases of thyroid cancer (DTC) and most anaplastic cases remain resistant to specific therapeutic strategies such as radioactive iodine (RAI). In addition, nearly 30 per cent of cases of medullary thyroid cancer (MTC) demonstrate resistance after surgery. When classified as "advanced thyroid cancers," patients with these severe types have a 5-year survival rate of less than 50 percent, compared with the 5-year survival rate of around 98 percent for patients with iodine-sensitive DTC (Naoum et al., 2018).

1.2 Types of Thyroid Cancer

Its histological characteristics have identified four different forms of thyroid carcinoma;

- Papillary thyroid cancer (PTC)
- Anaplastic thyroid cancer (ATC)
- Follicular thyroid cancer (FTC)
- Medullary thyroid cancer (MTC)

1.2.1 Papillary thyroid cancer (PTC)

Papillary thyroid carcinoma (PTC) is the most common thyroid cancer histotype and constitutes around 1 per cent of human malignancies. Patients with PTC appear to have a good prognosis at the time of diagnosis, particularly in those younger than 45 years. Nonetheless, following initial diagnosis, approximately 5 percent of patients with PTC report a recurrence within 5 years. A tumor is located in the resected contralateral lobe in patients with local or distant recurrence following lobectomy in more than 60 per cent of the cases. In addition, PTC is often multifocal, with a recorded frequency varying significantly from 18% to 87%, 61% of which are bilateral (Abdullah et al., 2019).

Most papillary carcinomas are associated with the existence of > 1 biologically distinct target, identified in 18–87 percent of recorded cases of PTCs. Nevertheless, indeed there is a gap in knowledge on how this phenomenon occurs from various individual tumors or from the intrathyroid spread of a basic single tumor mass. Given the discrepancies between several studies in molecular techniques, the general conclusion was that PTC multiplicity may result either from multicentricity or intrathyroid spread (Chmielik et al., 2018).

1.2.2 Anaplastic thyroid cancer (ATC)

One of the most lethal illnesses remains anaplastic thyroid cancer (ATC). It accounts for 1.7 per cent of all U.S. thyroid cancers. Geographical prevalence, however, varies from 1.3% to 9.8%. Overall, the average lifespan of patients is 5 months, and fewer than 20 per cent withstand one year (Smallridge and Copland, 2010).
The American Joint Cancer Committee (AJCC) classifies ATC into Stage IVA, B & C. Stages IVA are intrathyroid tumors; the primary tumor has evidence of gross extrathyroid extension in Stage IVB, and distant metastases in Stage IVC patients. ATC reveals a diverse collection of genetic mutations from a greater incidence of genes involved in the signaling pathways to β-catenin, PI3 K and MAPK (mitogen-activated protein kinase) effectors. Mutations of the BRAF and RAS genes are extremely common in anaplastic carcinoma, reaching approximately fifty percent. Another extremely prevalent molecular phenomenon of anaplastic thyroid carcinoma is PIK3CA copy number gain, which often intertwines between several somatic alterations including BRAF mutations. A typical of ATC gene mutation is TP53, with a prevalence of about 70 percent in ATC cases, but it is not found in distinguished thyroid tumors (O'Neill and Shaha, 2013).

1.2.3 Follicular Thyroid Carcinoma (FTC)

FTC is the commonest form of thyroid carcinoma (TC) after papillary thyroid carcinoma (PTC). they have always been more aggressive compared with PTC, tend to become much more developed mostly at onset of disease, are far less prone to conventional treatment, and hence more frequently cause morbidity. Because of its resemblance to FAs, preoperative diagnosis of FTC is very difficult, and the distinction between the two tumors is generally based mostly on the existence of vascular or capsular invasion. Not only FA and FTC became identical in terms of histopathological characteristics but also have a shared genetic history. In both FTCs and FAs, PAX8 / PPARπ and RAS somatic mutations transformations were found, the primary FTC

abnormalities. This carcinoma displays significantly lower phenotypic diversity than papillary; about 80% of follicular cancer reports include trabecular or microfollicular formation, but also well-formed colloid follicles are located in about 20% of cases. If even a small number of individuals with repetitive follicular tumors with capsular intrusion had metastasis that might be clarified by molecular and cellular heterogeneity (Chmielik et al., 2018).

1.2.4 Medelullary thyroid cancer (MTC)

Medullary thyroid carcinoma (MTC) is produced by calcitonin (CT) forming parafollicular or C cells and accounts for 5-10 percent of all thyroid cancers. MTC is inherited for around 25 per cent of cases. The discovery of an MTC in a patient has many effects. Essentially, the severity of the disorder should be measured, pheochromocytoma and hyperparathyroidism should be tested for, and whether the MTC is sporadic or inherited should be determined by a direct RET proto-oncogene study (Leboulleux et al., 2004).

MTC is usually quite well-defined and unencapsulated. MTCs are generally unilateral in occasional cases, while inherited tumors are normally multi, and bilateral. MTC's heterogeneity occurs primarily as simultaneous differentiation between follicular and parafollicular cell lines. PTC and MTC may occur as overlapping and synchronous tumors that are anatomically distinct. They create a so-called collision tumor, when MTC and PTC combine. All components are closely intermixed in mixed medullary-follicular (papillary) carcinoma (MMFC), and display anatomical and histopathological characteristics from both forms within a certain tumour. The MMFCs are exceedingly small, representing 0.15 per cent among all thyroid cancers (Chmielik et al., 2018).

1.3 High Throughput Sequencing Techniques

The compilation of the human genome project enhances the understanding of primary genome sequences. It leads the researchers towards a deep understanding of the biological cause of the disease. Starting from a rough draft of the sequenced genome to human diversity and disease and enter into a nascent era of personalized medicine. This is because of all the significant advancements in DNA sequencing techniques over the past few years. Sequencing has progressed far beyond in such a way that after the analysis of DNA sequences, the other biological components such as RNA and protein sequences have also been analyzed. It enables us to show their interactions in complex biological networks. Medical application of sequences has been made accessible due to increasing throughput and decreasing cost (Soon et al., 2013). The very first draft of the sequence of human genomes was compiled in 2001. After that, the sequence of many organisms was sequenced. Classical methods of sequencing, i.e., Sanger DNA sequencing, had low throughput with high cost (Reuter et al., 2015).

Through high throughput sequencing (HTS) techniques, one can sequence thousands and millions of molecules in a single run. Next-generation sequencing methods are now becoming popular. Different popular platforms have been used for sequencing. The pyrosequencing method was developed by 454 Life Sciences (also called Roche), which uses luciferase enzyme to readout nucleotide signals and then added to DNA templates. Illumina is another platform of sequencing that uses reversible dye terminator techniques. It adds a single nucleotide to the DNA template in each cycle. SOLiD sequencing by Life Technologies adopts two base sequencing methods through the process of ligation by using an enzyme. Increased throughput with increased and easy accessibility and lower cost broad the spectrum of research and enlightens the way of developing a rich catalog of HTS applications. DNA microarray technology was one of the original methods of sequencing based on intensities of probes labeled with

different dyes for different phenotypes (Soon et al., 2013).

1.3.1 Microarray Analysis

The microarray technology is used to identify the differential expression of genes. Microarray also permits the analysis of DNA sequence variation, proteomics levels, tissue level, and cell-level studies. Microarray relies on the hybridization cycle. In hybridization, cDNA is hybridized to probes to form a complementary sequence. The microarray technique is used for the analysis of multiple genes expressed efficiently. It allows the scientific community to understand the pathology of genetic causes occurs in the human body. Recent studies showed the advancement of microarray technology. Microarray analysis allows the mapping of chromosomal aberrations by using different platforms such as Affymetrix, Illumina, Agilent, and nano string (Hegde et al., 2000). The process is shown in Figure 1.5

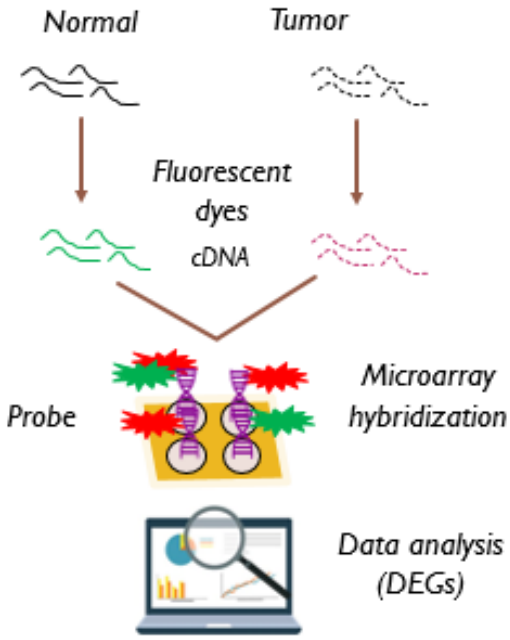


Figure 1.5. In the first step, mRNA is extracted from normal and tumor samples. These mRNAs are reverse transcribed into cDNA through the process of reverse transcription. Then add a fluorescent dye to label it. Finally, these cDNAs are hybridized with probes that are present on the microarray chip.

1.3.2 RNA sequencing

Next-generation sequencing technologies enlightens the path towards the era of genome sequencing by introducing advanced sequence-based technologies along with the advantages of high-throughput, sensitivity, and speed. The RNA-sequencing technique is now being used widely for transcriptome analysis to facilitate the researchers to find solutions to biological problems (Han et al., 2015). RNA-seq allows the sequencing of transcripts by high-throughput sequencing technologies. For whole-genome transcriptome profiling, RNA-seq becomes a useful approach as compared to the microarray. It can be used for transcriptome analysis, such as the detection transcripts counts, allelic expression, and splice junctions. There is no need for prior probe selection in this method and is free from all the biases that one has faced during the hybridization of microarray. RNA-seq technology is beneficial for gene and transcript based analysis. It is comprised of a few steps. Firstly, small complementary DNA sequences (cDNA) are formed through fragments of RNA samples, and then these fragments are subjected to high throughput sequencing machine. Second, the small generated sequences are mapped to the reference genome. Third, gene and transcript counts have been estimated. Fourth, then data are normalized, and by using statistical and methods of machine learning for the identification of genes that are differentially expressed (DEGs) are identified. In the end, the produced data can help solve the biological problem. As the demand for RNA-seq has been increased, different software, pipelines, and tools have developed for differential expression analysis (Costa-Silva et al., 2017). The whole mechanism has been shown in Figure 1.6

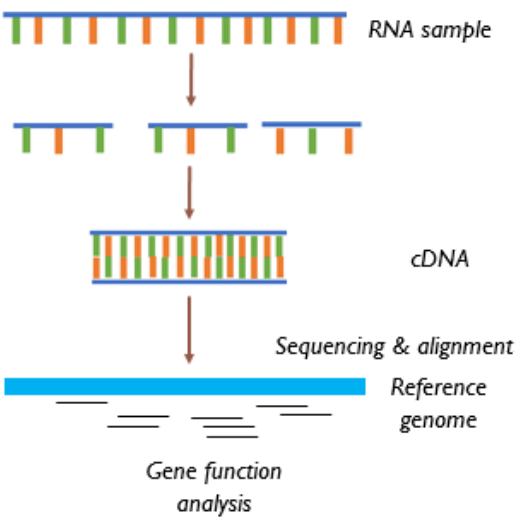


Figure 1.6. Firstly, RNA is isolated from a sample than RNA is converted to cDNA fragments through the process of reverse transcription. Reads are aligned to the reference genome. The reads which are mapped can be used to measure expression levels of genes or transcripts.

1.4 Messenger RNA

Gene expression means the formation of its corresponding proteins, which constitutes of two significant steps. The process of transcription started in the nucleus. The genetic information from DNA is towards messenger RNA (mRNA) molecule. The sequence of DNA is served as a template for complementary base-pairing with the help of enzyme RNA polymerase II for the formation of pre-mRNA. It is further processed to form mature mRNA. The mRNA then moved out of the nucleus into the cytoplasm. It attached to the ribosomes and translated into protein (Tomari and Zamore, 2005). The whole process has been shown in the Figure 1.7

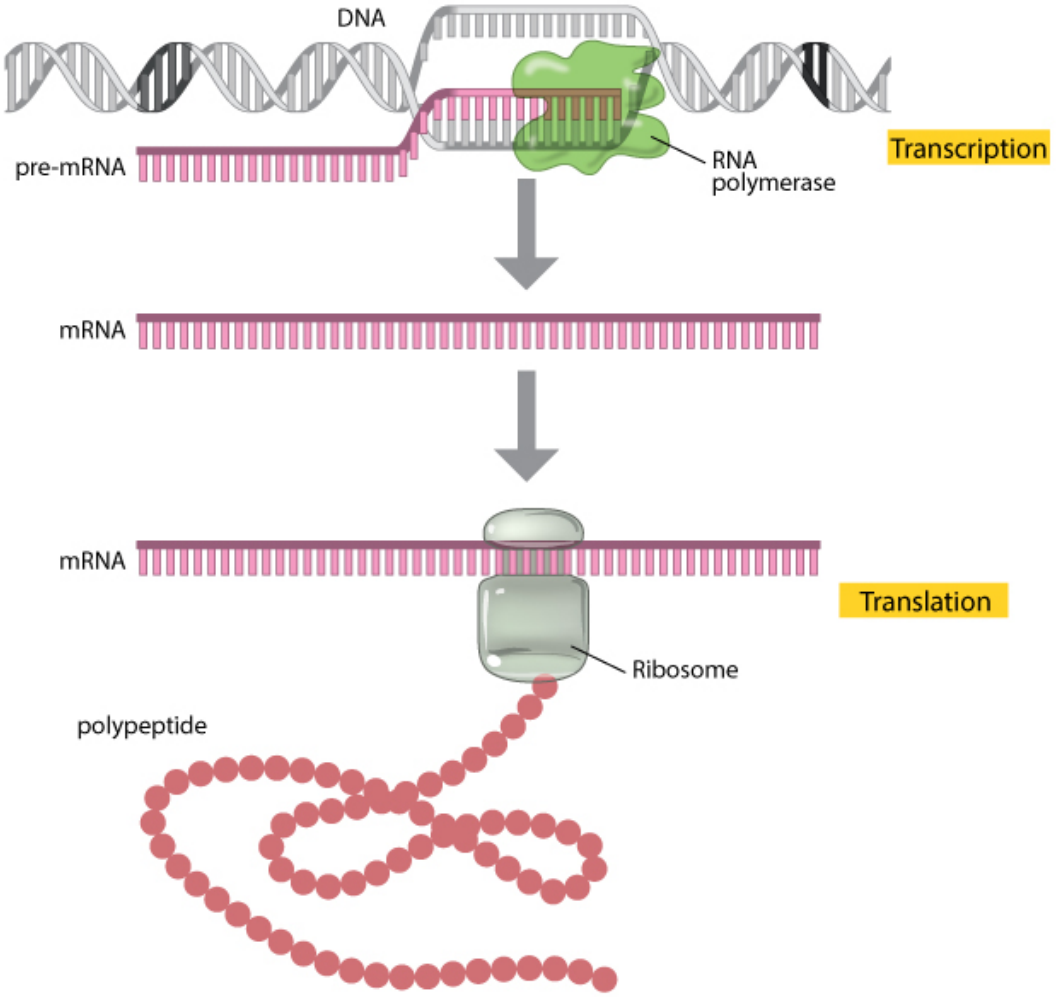


Figure 1.7. DNA is transcribed into RNA which further translated into protein. Process of reverse transcription convert RNA into double stranded DNA. The Figure has been adapted from (Clancy and Brown, 2008).

1.5 Micro RNA

MicroRNAs are small, double-stranded, non-coding RNAs that intervene gene expression at the post-transcriptional level. It regulates gene expression by deregulating the mRNAs. MicroRNAs are about 22-25 nucleotides in length and bind to 3' untranslated region of mRNA. MicroRNAs are transcribed through RNA polymerase II. After the process of transcription, primary miRNA is converted into precursor miRNA by RNAse III endonuclease within the nucleus. Finally, Exportin-5 transports precursor miRNA to the cytoplasm where Dicer (another RNAse III) transforms

it into functional miRNA (Peng and Croce, 2016). This process is shown in Figure 1.8

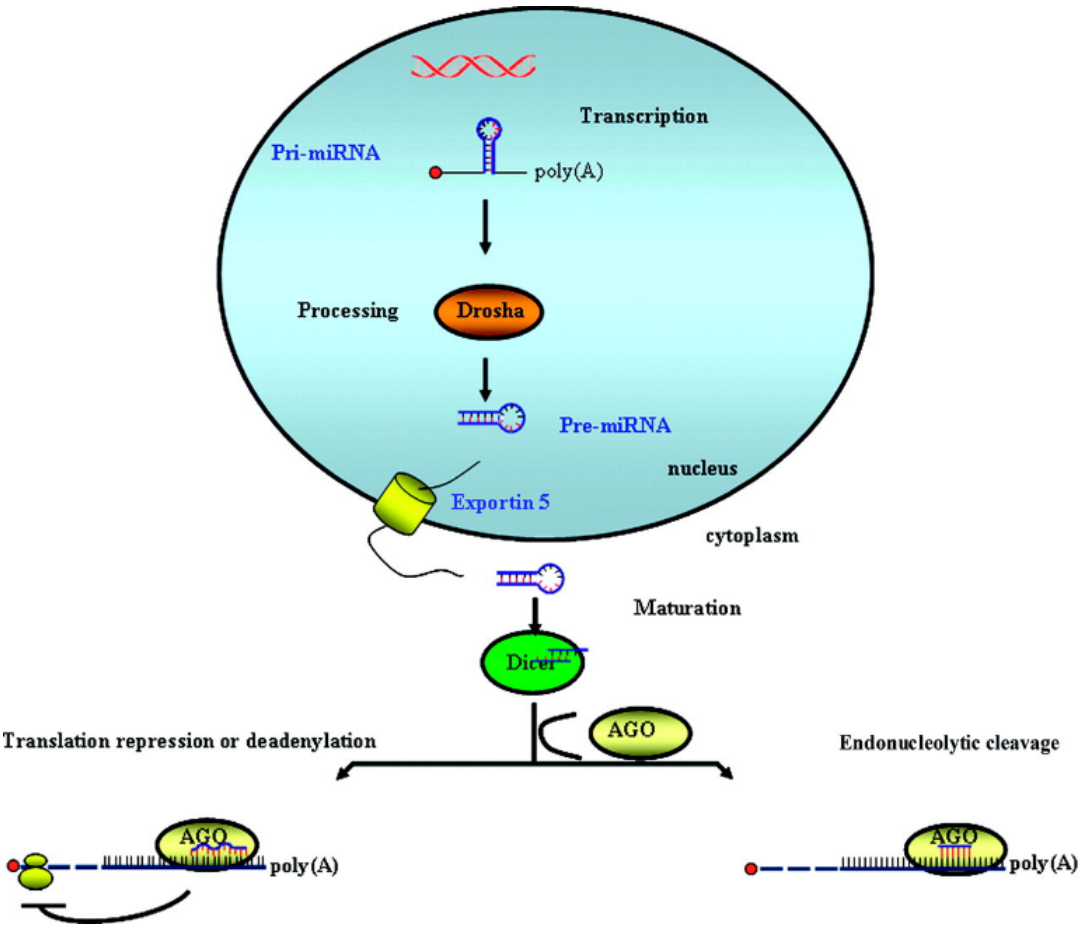


Figure 1.8. MicroRNAs are transcribed through RNA polymerase II. After that primary miRNA is converted into precursor miRNA by RNAse III endonuclease in nucleus. Finally, Exportin-5 transports pre-cursor miRNA to the cytoplasm where Dicer transforms it into functional miRNA. The Figure has been adapted from (Peng and Croce, 2016).

1.6 Pathway Analysis

In natural sciences research, pathway analysis is a leading tool meant to provide the research community with holistic information about the relation of the molecules and their impact on a specific signaling event. Besides analyzing the data, it also helps in intersecting the biological samples by finding genes that are somehow functionally related to each other and grouping them accordingly. Interpretation

of the data has become possible due to computational and statistical analysis that also supports the true meaning of molecular events in the context of any disease (García-Campos et al., 2015). Several software tools are developed for better pathway analysis that ultimately leads to understanding in different fields like metabolomics, proteomics, and genomics. Some of the publicly available pathway analysis tools are; DAVID, Reactome, pathvisio, KEGG, and Cytoscape. These tools help researchers to identify critical pathways in diseases (Xia and Wishart, 2010).

Chapter 2

LITERATURE REVIEW

This chapter illustrates the review of studies that has been carried out on thyroid carcinoma. The aim is to summarize the findings and outcomes of significant conducted studies, the responsible etiological agents and the use of microarray and high throughput sequencing techniques in context with the disease.

2.1 Role of miRNA, mRNA and lncRNA in Thyroid Carcinoma

miRNA and lncRNA emerge as key figures of the many molecules and mechanisms identified in recent years in the field of oncology. This happened because of their action on the regulation of known cancer genes and/or their products (tumor suppressor genes, oncogenes, and apoptotic proteins). In past, it's also been proposed that benign and malignant tumors can be differentially dignosed with the help of some of the miRNAs and/or lncRNAs. However little content is known about their role in prognosis. Interestingly, there are some miRNAs which have been observed repetitively dysregulated, especially in papillary thyroid carcinoma such as; , miR-181b, miR-187, miR-221, miR-146b and miR-222. In some studies, the same set of molecules was associated with tumor aggressiveness. Sadly, the specific collection of miRNAs differs from report to report, making any concrete conclusions currently difficult or even impossible to draw. The ambiguity of the existing information on lncRNAs is enormous because these long (over 200nt) RNAs can play a role at both the transcriptional and the post-transcriptional gene regulation level. NAMA, AK023948, lncRNAs and PTCSC3AA belong to the (still) reduced number of PTC-associated lncRNAs. To far it has not been possible to establish any function for lncRNA in the diagnosis and treatment of patients with thyroid carcinoma (Tavares et al., 2016).

There had been an elaborated review of the literature evaluating those researches

which showed or cited the evidence of direct interaction (e.g., luciferase assays) between the recognised miRNA and its suggested target(s) for mRNA against thyroid carcinoma. Reference was made to details about 44 different miRNAs. Many of those miRNAs are identified in the tumor biology which are involved as potential major elements. Nonetheless, some studies usually include only 9 miRNAs. Six of these are well known for their presence in many other forms of cancer: miR-145-5p, miR-221-3p, miR-222-3p, miR-21-5p, and miR-101-3p. Only half of the studies published on human thyroid carcinoma cell lines, however, conducted luciferase assays. Therefore, further confirmation regarding the direct suppression of the targeted mRNA(s) is needed for the remaining studies (Saiselet et al., 2016).

2.2 Role of miR-592 expression in thyroid carcinoma

A further detailed thyroid cancer analysis showing its incidence that accounts for more than ninety eight percent of all thyroid malignant tumors. The undifferentiated thyroid carcinoma (UTC) is one of the most vigorous malignancies in humans while the Papillary thyroid carcinoma (PTC) may just have a mild course of action. The number of different of genes have been identified to participate in thyroid carcinoma pathogenicity. In case of UTC and medullary thyroid carcinoma (MTC), effective therapeutic agents are almost non-existent once thyroid cancer has spread to distant organs. Primary prevention is usually feasible to allow for the possibility of cure. Nevertheless, our growing understanding of the genes involved in thyroid oncogenesis will help to establish more successful therapeutic approaches. Previous researches have also been examined which revealed that different microRNAs (miRs) are abnormally expressed in thyroid carcinoma and play a significant role in malicious thyroid cancer. The abnormal expression of miR-592 has been documented extensively in multiple forms of human cancer; moreover, its pattern of expression and functional areas in thyroid cancer incompletely understood. In order to determine the expression pattern of miR-592 in thyroid cancer tissues and cell lines, a reverse

transcription-quantitative polymerase chain reaction was conducted. The findings suggested that miR-592 in thyroid carcinoma samples was massively reduced, and its down-regulation was correlated with tumor-node-metastasis and lymph node metastasis. Thyroid cancer individuals with poor miR-592 expression showed a considerably low survival rate compared with individuals with severe miR-592 expression. An elevated level of miR-592 leads to reduced cell proliferation , migration and incursion of thyroid carcinoma.. However, the neuro-oncological ventral antigen 1 (NOVA1) has been reported in thyroid infected cells as a novel target gene of miR-592. The findings demonstrate that *in vitro* and *in vivo*, the NEAT1 / miR-592 / NOVA1 pathway can play critical role in the regulation of thyroid cancer malignancy (Yoo et al., 2019).

2.3 Role of miR-146b Expression in Thyroid Carcinoma

The recent research has shown that downregulation of microRNA-146b (miR-146b) is related to PTC viciousness and pathogenesis. Next we highlighted the current understanding of the biological functions, controlled target genes and therapeutic potential of miR-146b in PTC and explored how well these findings offered better guidance into the main role of miR-146b as a pathogenic regulator that promotes cellular differentiation as well as a diagnostic predictor for tumor progression in PTC patients. Together with the existing views on miRNAs in a wide range of human tumors, the analysis will ideally transform these revised results on miR-146b into more detailed diagnostic and therapeutic information on treatment in PTC patients before surgery and follow-up approaches. Even though the underlying principles and therapeutic research of miR-146b have yet to be fully understood, miR-146b expression in PTC not only predicts a specific complementary method for diagnosis and prediction, but can also serve as a potential biomarker and therapeutic target for PTC in the coming years (Chou et al., 2017).

Table 2.1. Genes implicated in thyroid tumorigenesis.

Cancer types	***Oncogenes***	***Tumor suppressor genes***
Papillary thyroid carcinoma	RET,MET,RAS,BRAF	p53
Medullary thyroid carcinoma	RET	-
Anaplastic thyroid carcinoma	BRAF	p53

2.4 Significant Differentially Expressed Genes

There is another detailed analysis containing data on gene expression of seventy eight thyroid carcinoma (C) samples, seventeen thyroid adenoma (A) samples, and four healthy thyroid epithelial tissues (N), retrieved through the Genome Expression Omnibus (GEO) database via the GSE27155 entry/accession ID. Of the one hundred and ninty (96 highly expressed in A), 294 (102 overexpressed in C) and 425 (with 183 overexpressed in C), respectively, Differentially Expressed Genes (DEGs) found between A vs N, C vs N and C vs A. In the N versus C genome expression analysis, MLLT3, RUNX1, FOSB, EGR2, KIT, and CTGF were suggested as helpful diagnostic tools for the enhanced group/cluster. EGR2 was identified as one of the five genes over a decade ago to efficiently recognize disease with an accuracy of 98.5 percent in follicular carcinoma diagnosis through regression technique. Along the complete absence of genetic variations, the lower or undetectable c-kit expression-pattern tried to argue against the significant role of c-kit in the prevalence of undifferentiated thyroid cancer cells, and RUNX1 was classified to be one of the forty three most appropriate genetic markers in PTC. The C vs. A distinction describes a total of 14 core genes, BCL2, CTGF, MMP7, EGR1, KDR, TIMP1, APOE, VWF, CCND1, BCL2L1, LGALS3, MCL1, DDIT3, and PGF. In conclusion, for the identification of DEGs, gene expression patterns of carcinoma cells in individuals with thyroid adenoma/cancer were compared with healthy epithelial cells, producing three comparative tables in pairs. The screened DEGs from GEO2R were successively characterized using review of the GO and pathway enrichment (Wang et al., 2018).

In this review, three datasets were combined consisting 114 Papillary Thyroid Carci-

noma (PTC) tissues and 126 healthy tissues, which included the greatest population of PTC tissue samples in related bioinformatics analyses, and reported 831 Differentially Expressed Genes (DEGs) containing 410 up-regulated and 421 down-regulated. Furthermore, the PPI network was designed for DEGs and explored the list of top ten hub genes (LRRK2, CD44, CCND1, JUN, DCN, BCL2, ACACB, TGFB1, CXCL8, and CXCL12) with maximum connectivity. At last, the three most influential modules were filtered out from the PPI network. Hence, in these modules the resultant genes have been connected to chemokine signaling pathways, cancer pathways, and PI3K-Akt signaling pathways. In addition, experimental verification is necessary to validate our expected output through the bioinformatics review. Therefore, in next step, the low expression of DCN, BCL2, ACACB, JUN, and CXCL12 and the up-regulation of these reported genes CXCL8, LRRK2, CD44, TGFB1, and CCND1 were confirmed by RT-PCR tests in thirty two groups of PTC samples and their corresponding healthy tissue. Amongst all, the hub genes were classified as clinical important genes CXCL8, DCN, BCL2, and ACACB through expression profile of 504 PTC samples from the Cancer Genome Atlas (TCGA) group. In the meantime, addressing the particularly clinically relevant genes (DCN, BCL2, CXCL8, and ACACB), would provide therapeutic interventions for PTC diagnosis. Nevertheless, there seems to be a big challenge to have these genes transmitted clinically for stratification of patients, used as biological markers for diagnosis as well as for immunotherapeutic or for oncovaccine production (Li et al., 2019).

2.5 Signalling Pathways and Networks

Throughout this research, a network-based integrative analysis of Follicular Thyroid Carcinoma (FTC) and lesion transcriptomes of benign follicular thyroid adenoma (FTA) were used to classify essential genes and pathways that vary among them. A dataset of microarray gene expression (GSE82208, samples = 52) was used, obtained from the tissues of FTC and FTA to classify those genes which expressed differ-

ently. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) tools were then used to investigate potential significant pathways and protein-protein interactions (PPIs) were analyzed to determine hub genes. In this study, 598 DEGs, 133 over-expressed genes and 465 down-regulated genes were reported in FTCs. 4 important pathways such as progesterone-mediated oocyte maturation signaling, cell cycle pathways , one carbon pool by folate, and p53 signalling were discovered which connected to DEGs having over-expressed FTCs. While eight pathways connected to DEGs with lower relative FTC expression-profile were identified. Furthermore, top 10 GO categories were closely related to FTC-over-expressed DEGs and 80 with low-expressed DEG (Hossain et al., 2020).

Different strategies such as whole genome sequencing (WGS) presented tremendous insight into the genetic abnormalities that are responsible for formation, development and de-differentiation of numerous categories of thyroid carcinomas. Such activities have resulted in the emergence of the MAPK and PI3 K signaling cascades as the key activation pathways involved in thyroid tumor progression. However, the existence of these important pathways is massively complicated, with thousands of elements, numerous crosstalk points, various subcellular morphological features and the ability to control several cellular processes potentially. As novel therapies, small-molecule inhibitors that target main kinases of these pathways offer great potential and many have entered medical research. Although some notable statements have been published, the production of susceptibility remains a concern and constraints patient reward. Throughout this study, the latest findings on the key components of the MAPK and PI3 K pathways were addressed, which included their activation mechanisms in physiological and pathological contexts, their genetic alterations regarding the various forms of thyroid carcinomas and the more important therapeutic agents designed to inhibit their function (Zaballos and Santisteban, 2017).

2.6 Potential Therapeutic Targets and Biomarkers

The research found that the symmetric chain of the main class II histocompatibility complex was CD74 and also a receptor for the inhibitory component of macrophage migration (MIF). MIF and CD74 were associated with histopathologic and solid tumor development and metastatic spread. It is observed in this study that sixty and sixty five percent of papillary thyroid cancers were positive for immunohistochemical staining of MIF and CD74, accordingly. For MIF anaplastic thyroid cancer was negative, but often positive for expression of CD74. Regular thyroid tissue and follicular adenocarcinomas have been adverse to expression of CD74. There was no affiliated clinicopathological criterion to the expression of MIF. Diagnosis in thyroid cancer cells with anti-CD74 antibody inhibited cell development, colony formation, vascular endolethial growth and cell tissue regeneration. Conversely, recombinant MIF treatment caused an enhancement in cell invasion. Treatment with anti-CD74 decreased phosphorylation of AKT, and triggered activation of AMPK. Our results indicate that overexpression of CD74 is closely linked with intensive tumor level, and may serve as a potential therapeutic. Briefly , we study CD74 expression specification in thyroid carcinoma and illustrate that anti-CD74 antibody therapy efficiently mediates tumor cells morphology. While the findings seem to negate the initial hypothesis that CD74/MIF mediates the connection among both inflammation and thyroid carcinoma. Our findings indicate that CD74 may serve as a therapeutic target in highly developed thyroid carcinoma (Cheng et al., 2015).

In general, strong prognosis of follicular thyroid carcinoma (FTC) degrades if the tumor does not hold radioactive iodine. Moreover, increased competition for this group of patients is on new druggable targets. After this, the prognostic and biological role of survivin and XIAP in FTC was studied in detail. The expression of XIAP and survivin was analyzed through tissue microarray in 44 FTC and subsequent non - cancerous thyroid samples. shRNAs induced inhibition of both apoptosis protein in-

hibitors (IAPs), or particular small molecule antagonists, and biochemical changes were reported in vitro and in vivo. Production of survivin associated with progressive tumor stage and persistent disease. Moreover, survivin has proven an unbiased adverse prognostic marker. Knockdown of XIAP or survivin resulted in a significant decrease in viable cells and propagation, stimulated caspase3/7 and was affiliated with reduction in vivo cancer progression. A decrease in enzyme activity, differentiation and cell cycle phase caused by IAP-targeting compounds with an increase in apoptosis. YM155 a small Survivin expression molecule inhibitor strongly suppressed tumor growth in vivo. Both IAPs show important functional effects in FTC tumorigenesis and therefore demonstrate to be potential targets in patients with developed FTC (Werner et al., 2017).

In the risk assessment of recurrence using SVM algorithm researchers further conducted a study to evaluate significance of the 4 independent lncRNA genetic markers as a predictor. Results of this study have examined the usefulness and possible impact of the 4 independent biomarkers of lncRNA in forecasting the recurrence risk. It was concluded that the present study conducted genome-wide analysis for the expression of lncRNA in the PTC infected individuals from the large population of TCGA and showed altering patterns of expression among cancer and non – cancerous specimens as well as between recurring samples and samples free of recurrences. four lncRNAs including RP11-508M8.1, AC026150.8, RP11-536N17.1, and CTD-2139B15.2 have been reported by a lncRNA signature, that can be used as an alternative diagnostic indicator to rigorously determine the recovery and relapse of PTC individuals. Such recognized lncRNAs, further with experimental prediction methods in prospective cohort studies, may support as potential therapeutic triggers and biomarkers for PTC infected patients (Li et al., 2017).

Several drugs have also been reported, approved from U.S based Food and Drug Administration (FDA) for the thyroid cancer treatment;

Table 2.2. Drugs approved by FDA for thyroid cancer treatment.

Sorafenib	Vandetanib	Axitinib	Sunitinib
Lenvatinib	Motesanib	Cabozantinib	Pazopanib

The medications currently available have side effects to their respective targets, i.e. asthenia, hypertension, nausea and myositis and drug resistance, which restrict their maximum potential for use. Following are the causes of drug resistance found in literature;

- All tumors of about 2.0 cm in size but limited to the thyroid gland are now classified as Thyroid Cancer (T1), while previously only tumors of around 1.0 cm were classified as T1. This may lead to undertreatment in some patients , leading them to a higher risk of recurrence due to a less intensive initial diagnosis (Grande et al., 2012) & (Naoum et al., 2018).

- There is recent evidence that thyroid cancer is a disease of the stem cells with infinite growth capacity and resistance to traditional therapeutic regimens (Naoum et al., 2018).

- Elevated levels of Tg (thyroglobulin) after surgery that causes immunologic destabilisation (Grande et al., 2012) & (Naoum et al., 2018).

2.7 Study Rationale

- Despite of major advances in screening and treatment, which includes radiological tests, ultrasound, MRI, and surgical resection, mortality rates continue to rise.

- There were many reported genes as well miRNAs but their expressions were different according to different regions.

- For better understanding of molecular basis of the disease of different regions we have to study that from sequence level and have to check both mRNA and miRNA expressions.

2.8 Problem Statement & Proposed Solution

Reoccurrence is the major problem for thyroid cancer patients because drugs show resistance to the treatment. As,overall pathogenicity still needs to be explored at molecular level. Therefore, we need to identifying effective therapeutic targets through gene expression profiles based on differential expressed genes and pathways via microarray and high throughput sequencing data

2.9 Objectives

- Identification of differentially expressed genes (DEGs) using microarray and high throughput sequencing technology through datasets of thyroid carcinoma of different platform
- To find common pathways and metabolic processes in which differentially expressed genes are involved

Chapter 3

MATERIALS AND METHODS

The primary purpose of this research is the use of microarray, RNA Seq to find DEGs, and transcripts of miRNAs and mRNAs that can serve as therapeutic targets. These DEGs are further used to perform pathway analysis through the approach of systems biology. The general workflow used in this study has been shown in Figure 3.1

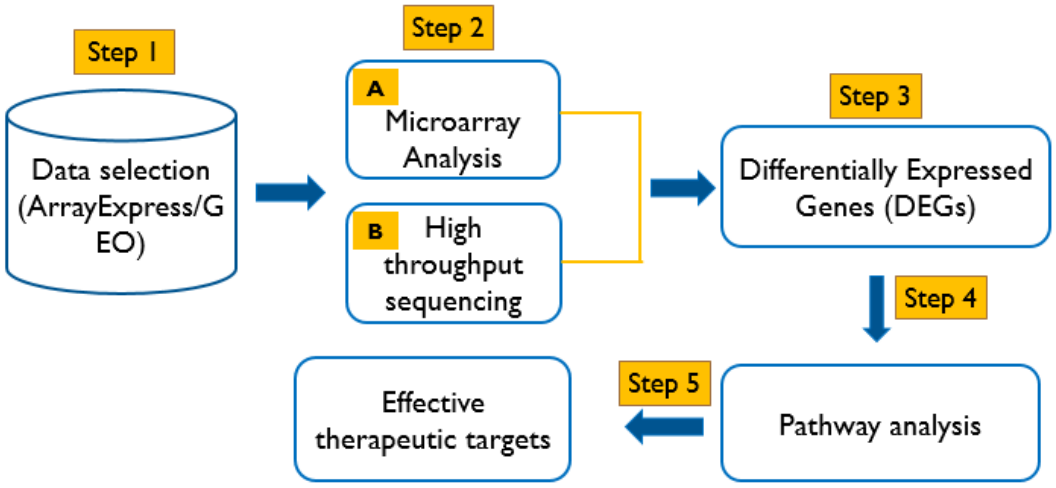


Figure 3.1. General workflow of the study.

3.1 Datasets Retrieval

Datasets of Thyroid Carcinoma are selected from publicly available repositories such as Array Express and NCBI-Gene Expression Omnibus (GEO). Table 3.1 shows the sample information regarding the microarray analysis of thyroid carcinoma datasets. While Table 3.2 shows datasets for RNA seq analysis.

Table 3.1. Datasets for microarray analysis of thyroid carcinoma.

Accession No	*Organism*	*Phenotype*	*Region*	*Platform*	*Samples*
E-GEOD-65144	Homo Sapiens	Diseased/ Normal	USA	Affymetrix (µ-array)	25
E-GEOD-3467	Homo Sapiens	Diseased/ Normal	USA	Affymetrix (µ-array)	17
E-GEOD-40807	Homo Sapiens	Diseased/ Normal	France	Agilent (µ-array)	80

Table 3.2. Sample information regarding RNA seq analysis.

Accession No	*Organism*	*Phenotype*	*Region*	*Platform*	*Samples*
GSE57780	Homo Sapiens	Diseased/ Normal	Belgium	RNA seq (miRNAs)	6
GSE57780	Homo Sapiens	Diseased/ Metastasis	Belgium	RNA seq (miRNAs)	6
GSE64912	Homo Sapiens	Diseased/ Normal	Italy	RNA seq (mRNAs)	22

3.2 Microarray Analysis

For Microarray data analysis, a recently published maEndToEnd (end to end differential gene expression) pipeline for the Affymetrix platform was used (Klaus and Reisenauer, 2016). Whereas in the datasets of Agilent single-channel platform, different codes are utilized for background correction and between array normalization. The complete analysis is performed on R studio (R 3.5.1) by using R script (Team, 2018). This pipeline was based on various packages.

General Bioconductor packages

• Biobase

Data collection as quantitative values i.e. gene expressions are calculated in several samples with several features i.e. molecules and genes. It is a standard micro-array expression data container, which can also be used for other data types i.e. medication panels (Huber et al., 2015).

• OligoClasses

It includes class descriptions, validity tests, and methods of initialization for classes used in the packages oligo (Carvalho and Scharpf, 2011).

Quality control and preprocessing packages

• Oligo

This evaluate Probe level oligonucleotide arrays and actually facilitates arrays of NimbleGen and Affymetrix (CEL files) (Carvalho and Irizarry, 2010).

• ArrayQualityMetrics

This manages almost all of the latest microarray techniques and is suitable for use in integrated applications for research or for automated data analysis, and even for personal access (Kauffmann et al., 2009).

Analysis and statistics packages

• Limma

It is used for analysis of linear concepts and microarray deferentially expressed data (Ritchie et al., 2015).

• ClusterProfiler

This executes methodologies for analysis and visualization of functional patterns (GO and KEGG) of clusters of genes (Yu et al., 2012).

Plotting and color options packages

• ggplot2 & pheatmap

For visualizing and plotting the genomics data (Gentleman, 2018) & (Fisher et al., 2019).

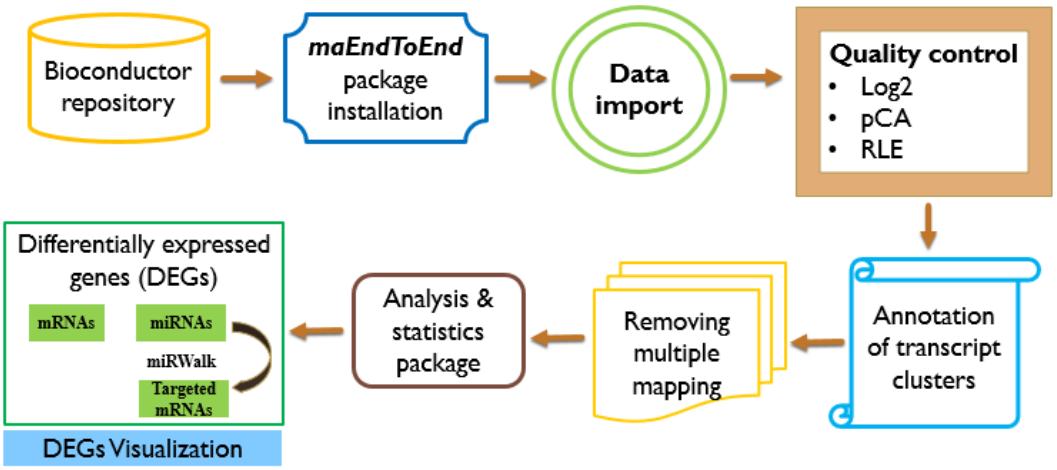


Figure 3.2. Workflow for microarray analysis.

3.2.1 maEndToEnd pipeline

As indicated by the name, Bioconductor packages were arranged end to end to analyze microarray data for the Affymetrix platform. An R version of 3.6.1 was used for analysis. maEndToEnd was installed using BiocManager, which is explicitly made for the installation of packages that are designed for handling and analysis of high throughput data of RNA seq and microarray. Once it is installed, the library of maEndToEnd was called. While the library of all the integrated packages would automatically be called by calling maEndToEnd (Klaus and Reisenauer, 2016).

3.2.2 Data Import

Expression data from all the datasets are imported using the Bioconductor package ArrayExpress. ADF (array design format), IDF (investigation description format), SDRF (sample and data relationship format), raw data files, and processed data files are fetched. In these files, the IDF, SDRF, and raw data files are utilized in our analysis. Raw data obtained from the Affymetrix platform is in the form of CEL files, while Agilent raw data is present in the form of text files. Both consist of measured probe intensities (Klaus and Reisenauer, 2016).

A) SDRF file

It contains essential information on the samples, such as the cel file. SDRF file imports with the read.delim function in order to acquire the sample annotation. To create an Expression Set for our data, the SDRF table is sorted by selecting columns of interest into an AnnotatedDataFrame from the Biobase package. The package Biobase contains genomic data (Klaus and Reisenauer, 2016).

B) The Expression Set

This class contained different sources of information. This information, later on, converted into a single, convenient structure. The expression set consists of assay data, Metadata, and experiment data (Klaus and Reisenauer, 2016).

3.2.3 Quality Assessment

Quality assessment is performed to check the quality of the gene expression data based on arrays and to detect quality errors or problems (Cohen Freue et al., 2007). After normalization, it is an essential step of the analysis for which Bioconductor package arrayQualityMatrics is used. Principal Component Analysis (PCA) plot, boxplot, Relative log expression (RLE) plot are generated using package oligo and ggplot. An expression data is usually analyzed by taking a logarithmic scale, so we took log2 of the raw expression data. To check the quality of the data, different plots were produced by using different packages (Wickham, 2016).

A) PCA plot

PCA deals with multi-dimensional data and reduces the dimensionality along with simplifying the complexity. It clusters the data according to different phenotypes or treatment groups by finding patterns in the absence of a reference (Ringnér, 2008). PCA graph is generated, which represents expression data in the form of points to show the samples with different phenotypes. It is performed based on a log intensity scale. PCA was performed on both raw and processed data that has been given as input. The data is composed of two phenotypes i.e. tumor vs. normal (Klaus and

Reisenauer, 2016).

B) Boxplot

Boxplots are generated to statistically analyze the biological expression data and present us with its distribution. It consists of an upper quartile (Q3), median and lower quartile(Q1), and Q3-Q1 gives us interquartile range (IQR). To find outliers in our data, Q3+1.5xIQR and Q1-1.5xIQR are calculated. Boxplots are generated in which each box represented a sample on the basis of probe intensities. It displays the intensity differences between samples and signifies whether the data should be normalized or not (Thirumalai et al., 2017).

C) RLE plot

To visualize and analyze undesirable variation in gene expression data, RLE plots are generated, also to check if normalization has been performed accurately (Gandolfo and Speed, 2018). RLE is performed to calculate the median of log2 intensities for the expression of raw data. For each probe-set, the values of RLE data are computed by taking the ratio of the expression values by subtracting median expression values across arrays. An RLE plot is generated on the data to check the quality as log2 intensities must be centered near zero along with a similar spread (Gandolfo and Speed, 2018).

D) Robust Multi-array Average (RMA)

Oligo, a Bioconductor package used for preprocessing oligonucleotide microarray datasets. It allows the user to perform background correction, normalization, and summarization in a single go by using the RMA algorithm. The package provides a combined framework for parallel execution of different steps with the support of Bioconductor (Carvalho and Irizarry, 2010). In this study, quantile normalization is used for the standardization of the raw data. Quantile normalization is such an adjustment method that considers the statistical distribution of samples in a dataset is the same. The observed distributions of all samples are normalized to the equal median by taking an average of each quantile across all samples. Summarization is another

critical step as in a microarray experiment. Multiple probes are located at numerous locations that are needed to summarize into one quantity (Hicks and Irizarry, 2015)

3.2.4 Preprocessing

Preprocessing was performed using background adjustment, calibration, and summarization. These are briefly described below.

A) Background Adjustment

After the quality assessment of microarray data, the background of experimental values was adjusted. It is performed to cope with non-specific hybridization of a probe with target and optimal noise generated by the scanner. The scanner catches the fluorescence signals due to the insignificant or non-specific binding of a target with a probe. Consequently, the background of the probe intensities must be adjusted prior to further analysis to get significant results (Irizarry et al., 2003).

B) Calibration

It refers to the normalization of experimental microarray data. Associating and relating to experimental values generated from different microarray chips give rise to issues like complications with microarray chips, varied laboratory conditions, and batch effects. To tackle these issues, calibration across biochips is essential. After calibration, experimental values from different biochips were comparable. Quantile normalization was implemented to that step (Irizarry et al., 2003).

C) Summarization

It is an essential step to perform after calibration while handling data from the Affymetrix platform. Probes are merely the oligonucleotide sequences of about 25 base pairs. In the Affymetrix platform, each transcript is depicted by many probes. Thus the background adjusted and calibrated experimental values of probe intensities must be transformed into single values for every specific gene. That transformation values from multiple probes would help access the relative quantity of the RNA transcripts. Robust multiple average (RMA) algorithm was used to summarize the

background adjusted and calibrated experimental data of intensities of probes (Giorgi et al., 2010).

3.2.5 Heatmap

Heatmap is a graphical visualization of the multivariate high-throughput data. It shows the similarity level of the gene expression patterns of expression data by clustering different samples. Heatmap is generated to analyze distance within and among the samples where the distance algorithm used is Manhattan (Klaus and Reisenauer, 2016).

3.2.6 Linear models

Linear models are a general class of models that is used on continuous data such as gene expression values. The gene expression data obtained from microarray experiments usually consists of log-ratios or log intensity values on which linear modeling is performed by using the package limma as it can model and fit a wide range of genomic data. To perform differential expression analysis, the limma package is utilized, which works on both single-channel and two-color channel microarray data. Linear modeling is performed using the limma package to find DEGs. Empirical Bayes (eBayes) method is used, which is already administered in the limma package. It borrowed information across genes and modeled the relationship between gene expression and the variance of the genes. In the process of linear models, two matrices are formed; design and contrast matrix (Robinson et al., 2010).

A) Design matrix

The design matrix is constructed to check the variation in gene expression data. A design matrix has entries of zeros and ones. Row names of the design matrix were patients, while the column names were the defined variables of the linear model i.e., phenotype abbreviations. One in column depicts turning on of phenotype of a respective sample while zero represents turning off of phenotype of respective row

(Klaus and Reisenauer, 2016).

B) Contrast matrix

The factors defined by the design matrix are then allowed to be joined into a contrast of interest by forming a contrast matrix. In the later matrix, each contrast relates to a comparison of interest between the disease and control samples (Klaus and Reisenauer, 2016).

3.2.7 Differential expression analysis

DEGs are obtained after fitting the linear model for gene expression data by applying the topTable function. Results are sorted by t-statistics (Ritchie et al., 2015). In this study, a significant cut-off for negative log of p.value equals to 0.05 and log2FC (log2 fold change) equals to 0.5 was set to extract the genes that were differentially expressed.

3.2.8 For Agilent platform

The dataset E-GEOD-40807 is from an Agilent platform. For microarray data analysis of miRNA from the Agilent platform, the AgiMicroRna package was used.

A) AgiMicroRna

As this study also includes datasets of miRNAs of different platforms, so different packages are used to find differentially expressed miRNAs. AgiMicroRna has such useful functionality for the preprocessing, quality control, and differential expression analysis of Agilent microRNA array data. The package uses a limma package for gene expression analysis of microRNAs by fitting the linear model http: //bioconductor.org/packages/AgiMicroRna/.

B) miRWalk

The database used to identify the targeted genes of differentially expressed microRNAs. It contains validated targets associated with genes, pathways, diseases, cell lines, etc. of humans, rats, and mice. It is the only database that finds the com-

plementary binding sites of microRNAs on complete genomes includes a promoter, 5' UTR, CDS, and 3' UTR. The results are validated through 8 microRNAs target predicted programs i.e., miRanda, miRDB, PITA, PicTar and TargetScan, DIANA-microT (Dweep et al., 2014).

3.2.9 Volcano-plot

Volcano-plot, a type of scatter-plot, is generated to obtain and visualize the highly expressed DEGs, whether they are up- or down-regulated. It compares the expression level of each gene and shows log2FC plotted against the negative log of P.value (W. Li, 2012). To visualize the DEGs volcano-plot is formed using the Bioconductor package EnhancedVolcano. By setting a cut-off value, enhanced volcano-plot highlights the DEGs of interest with high or low expression (Blighe et al., 2019).

3.3 RNA sequencing Analysis

RNA-seq analysis is performed using workflow adapted from galaxy shown in Figure 4.3 on three publicly available thyroid carcinoma datasets shown in Table 3.2 to obtain DEGs.

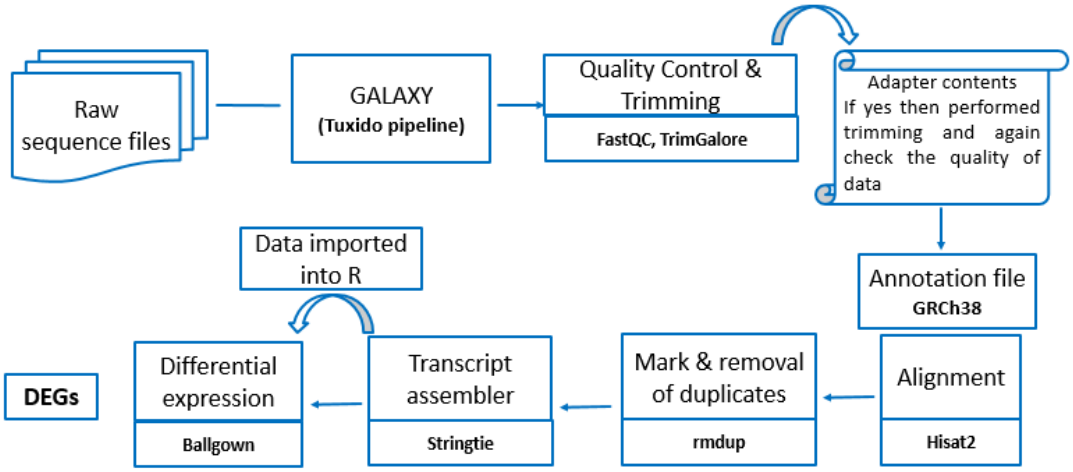


Figure 3.3. Workflow for RNA-seq analysis.

3.3.1 Galaxy Pipeline

Galaxy is a simple interface to powerful tools that is available at https:// usegalaxy.eu/ accessed on January 10, 2020. It automatically manages the tools and their updates. Galaxy is present in the form of publicly available web service and downloadable package. Galaxy contains different tools for the analysis of genomics, comparative genomics, and functional genomics data. It is used to execute complex level analysis of high throughput data, including RNA-Seq. A pipeline in the galaxy platform is made to perform quality control, preprocessing, mapping to the human reference genome, and normalization. For this purpose, some tools are used that are available in the Galaxy interface. These tools are FastQC, RseQC, HISAT2, and StringTie. Galaxy interface provides users with data storage of 250 GB for high-throughput analysis (Blankenberg et al., 2010).

A) Data import

The data from different datasets are imported from the ENA database (as it directly links to Galaxy software) in the form of FastQ files. One of the datasets is paired-end while two data sets are single end. The FastQ files are in the format of text files, which consists of raw sequence reads.

B) Quality control

On the raw data, which is in FastQ.gz format, to check the quality of raw sequence data obtained from high throughput sequencing technology, FastQC version 0.72 is utilized to check for any biases. In the process, a quality control report is obtained, which might consist of some technical or biological errors. The quality of data is visually represented in a graphical form. FastQC is performed to check the quality scores, sequence quality, per sequence GC content, adapter content, and sequence duplication level (Andrews et al., 2010).

C) Trim Galore

Trim Galore is a wrapped script that is used to trim any nature of adapter contents if present in the sequence. It increases the quality of data as well. Trim

Galore will try to auto-find whether the Illumina universal, Nextera transposase, or Illumina small RNA adapter sequence is used. If adapters are their in sequence, it identifies the nature of adapters and trims these adapters to enhance the quality of sequence (Leggett et al., 2013).

D) Alignment

Alignment is performed once the FastQC files with better quality scores are achieved. The files obtained after analysis on the FastQ quality trimmer tool are selected for the sequence alignment. HISAT2 version 2.1.0 is an alignment tool that aligns reads to the reference genome (Hg38), which exists in it by default. In the end, a BAM file is generated, which contains the aligned sequence reads (Kim et al., 2015) & (Rosenbloom et al., 2015).

E) Read duplicates

Read duplication (RSeQC) version 2.6.4 tool is used to identify sequence and mapping based duplicate reads. RSEQC package consists of modules like basic and RNA-seq specific modules, which helps in evaluating sequence data. The BAM file is then analyzed for identifying the principal component analysis (PCR) duplicates by using a read duplication tool. It is a quality control tool for RNAseq data through which we know about the level of duplication in our data due to PCR replicates. The data is then visualized graphically through RSEQC plots (Wang et al., 2012).

F) MarkDuplicates and RmDup

The aligned duplicate reads in BAM files are marked or located with the help of the MarkDuplicate tool. Through a default method, it ranks the reads by sums of their base quality score and so it differentiates the duplicate reads from the original sequence. MarkDuplicate tool is used to identify and mark the replicates in our sequence data present in the form of the BAM file. While RmDup tool version 2.0.1 is used for duplicates deletion, which helps in removing the duplicates as reads with the highest mapping quality is removed (Li et al., 2009).

G) StringTie

StringTie tool is used for transcriptome analysis. StringTie can also perform an abundance estimation. StringTie provides estimated abundance in the form of FPKM and RPKM. FPKM estimation is used for paired-end, and single-end reads. It has high coverage to estimate the abundance of FPKM and RPKM than other tools (Pertea et al., 2015). For the assembly of transcripts a reference annotation file of GRCH38 is used in GTF format and creates essential multiple isoforms. For differential expression analysis, ballgown output files are retrieved to calculate the expression values of genes and transcripts (Guo et al., 2017).

H) Differential expression analysis through Ballgown

The Ballgown package of R can be used to read the data for downstream analysis. Ballgown uses a flexible linear model framework for differential expression analysis. It uses the StringTie files to plot the read level coverage for transcripts of interest. Ballgown can work with any assembly tool that produces assembled transcripts and expression estimates as an output. The Ballgown object loads the data for the expression level of introns, exons, and transcripts for genomics measurements (Frazee et al., 2014)

3.3.2 Pathway Analysis

Pathway analysis is performed on the DEGs obtained from microarray and RNA-seq data analysis. The DAVID database was used for analysis, which is an online database and provides us with functional annotation tools (Dennis et al., 2003). Validation is performed by using Enrichr on the pathways obtained from KEGG. Enrichr is a publicly available and published web-server for performing Gene Set Enrichment Analysis (GSEA). It consists of a wide range of gene set libraries for analysis and helps in biological discoveries with the help of the gathered biological knowledge (Kuleshov et al., 2016). KEGG is a database in itself, while DAVID utilizes KEGG as a reference. KEGG consists of computationally generated manually drawn organism-

specific pathways. The database contains knowledge about the molecular interactions, networks, and reactions in the metabolism, human diseases, drugs, genome and cellular along with organism systems. Major pathways involved in different diseases, along with sub-pathways, can also be studied in the KEGG database (Kanehisa et al., 2017).

Chapter 4

RESULTS

The principal goal of this study is to identify differential expression of genes through microarray and RNA-seq analysis. Pathway analysis is performed to analyse common pathways in all the datasets. Datasets of mRNAs and miRNAs of different regions and platforms have been used in order to identify differentially expressed genes. In the end, comparative analysis on the basis of identified pathways give us a common pathway among all datasets.

4.1 Results of Microarray Data Analysis

The results of microarray data analysis of Thyroid Carcinoma are discussed with the help of figures and tables.

4.1.1 Microarray Dataset-1 (E-GEOD-65144)

On the dataset E-GEOD-65144, analysis was performed on 25 Homo sapien samples where the organism part was thyroid tissue. Thirteen normal samples and 12 anaplastic thyroid carcinoma samples were selected. First, we used PCA-plots, boxplots, and RLE-plots to analyze the quality of the raw and calibrated data. Then heatmap and enhanced volcano-plot were also generated to enable the quick visual identification of genes that were also statistically significant.

Figure 4.1 represents the PCA-plot for the log2 transformation of the raw expression data containing two groups of samples (Normal and ATC). Principal component analysis (PCA) was performed for the dimension reduction of high dimensional data. Samples were color-coded w.r.t phenotypes. In the PCA plots, the blue dots represent the normal condition, while orange dots represent the infected samples.

The first principal component (PC1) is plotted on the x-axis, and the second principal component (PC2) plotted on the y-axis.

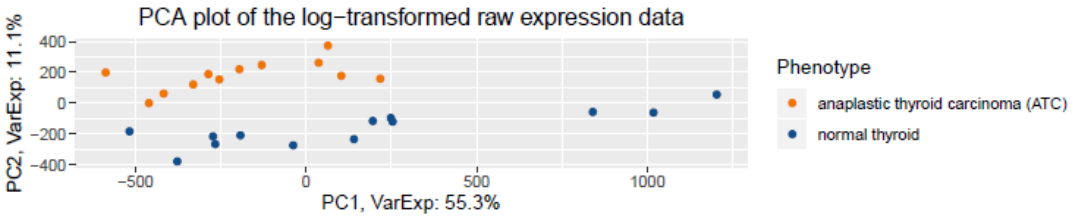


Figure 4.1. PCA-plot of the raw expression data (E-GEOD-65144).

The PCA plot of the raw data shows very little difference between the two phenotypes (orange and blue dots) along the PC1. Therefore, it was hard to differentiate two phenotypes based on the raw data. The raw data was then calibrated and summarized before the PCA. In calibration, the background was adjusted, followed by quantile normalization and summarization. Figure 4.2 shows the PCA-plot of the summarized data. In this plot, the different phenotypes are separately clustered, and there is a significant difference between these two phenotypes along the PC1.

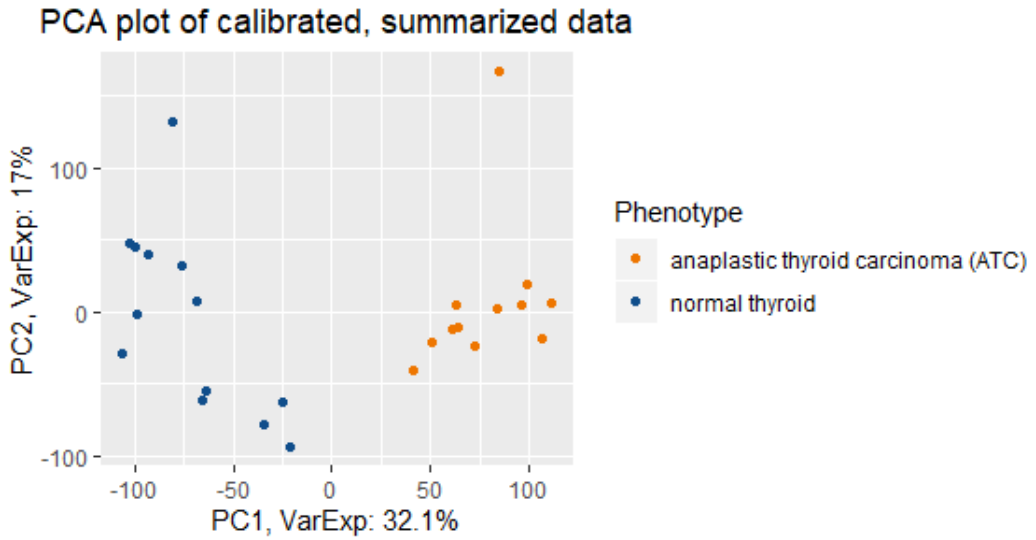
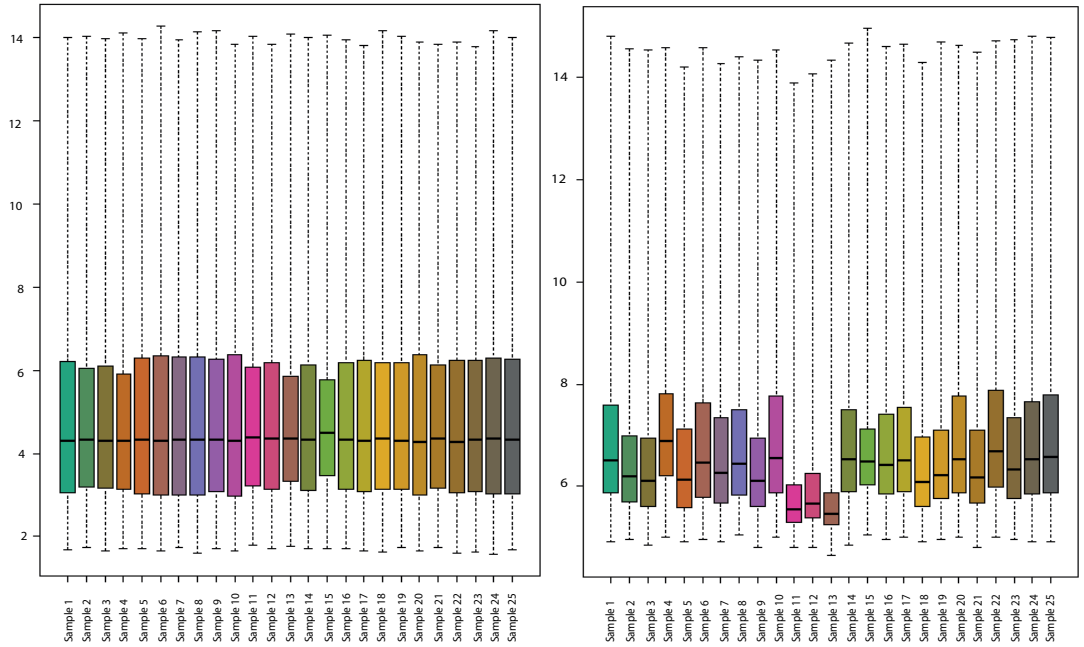


Figure 4.2. PCA-plot of the normalized data (E-GEOD-65144).

Figure 4.1 and 4.2 highlights the significance of the calibration chosen for this specific dataset which was also supported by the boxplots of the raw and calibrated data, shown in Figures 4.3**A** & 4.3**B** respectively. In these boxplots, each box rep-

resents the probe intensity of an individual sample, and the center-line of each box represents median intensity. The samples of both normal and ATC phenotypes can be seen on the x-axis, while the y-axis shows the expression values of genes. The boxplot of the raw data highlights the variation among samples alongwith outliers, as shown in Figure 4.3**A**. This variation among samples was reduced by calibrating the raw data. The boxplot of this calibrated data is shown in Figure 4.3**B**

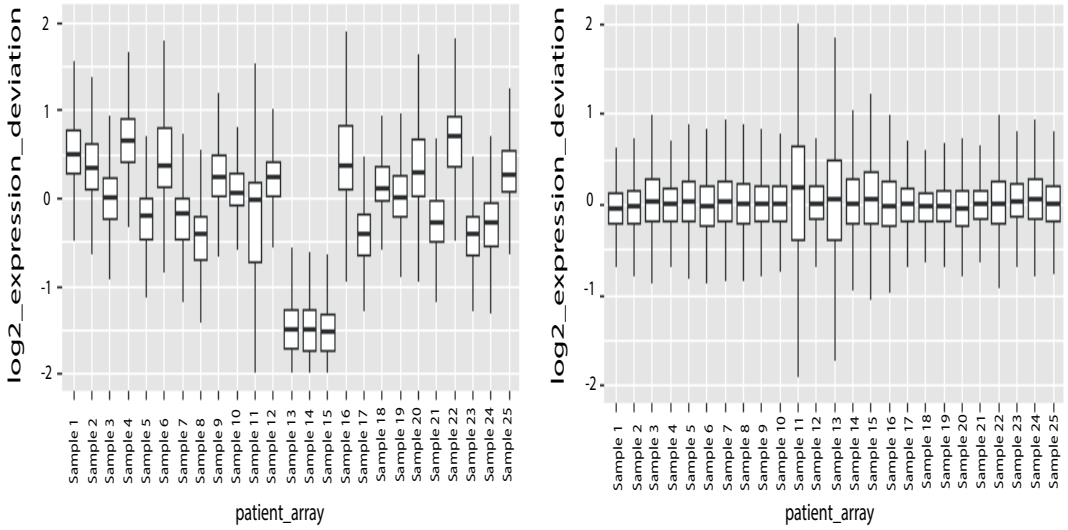


(a) Boxplot of raw data. **(b)** Boxplot of calibrated data.

Figure 4.3. (A) Boxplot of raw data & (B) Boxplot of calibrated data (E-GEOD-65144).

To further support the notion of calibration of the raw data, RLE plots were also extracted for both raw and calibrated data, shown in Figures 4.4**A** and 4.4**B** respectively. As shown in figure, the calibration of the raw data made sample medians symmetric along the zero line.

After calibration we were able to get two distinct phenotypes with their sample intensity symmetric around zero. The next step was to look at the distance between the individual samples because highly correlated samples (small distances) represents a phenotype. We used the heat map to highlight this correlation among the samples. In Figure 4.5, the phenotypes are color coded (brown for normal samples and green



(a) RLE plot of raw data. **(b)** RLE plot of calibrated data.

Figure 4.4. (A) RLE plot of raw data & (B) RLE plot of calibrated data (E-GEOD-65144).

for ATC samples). Each row represents a sample and are labelled with their IDs. The dendrogram indicates distance between the samples. Large distance between the samples is indicated by low intensity shade (lighter tone). The small distance between the samples is indicated by high color intensity (color tone towards red shade).

As mentioned earlier that before Principle Component Analysis (PCA), data standardisation is necessary. Prior implementing PCA, we must standardise the data, otherwise PCA would not be able to find the optimal key components. Therefore it was hard to differentiate two phenotypes based on the raw data. The raw data was then calibrated and summarized before the PCA. Therefore different phenotypes are separately clustered. PCA can be used primarily for highly correlated variables. PCA does not work well to reduce data if the relationship is weak between variables. However, the manhattan distance is based on absolute value which could produce more consistent performance. So the next step was to look at the distance between the individual samples because highly correlated samples (small distances) represents a phenotype. We used the heat map to highlight this correlation among the samples. So if the correlation among two samples was low, it showed larges distance between them indicated

by lighter color tone and vise versa.

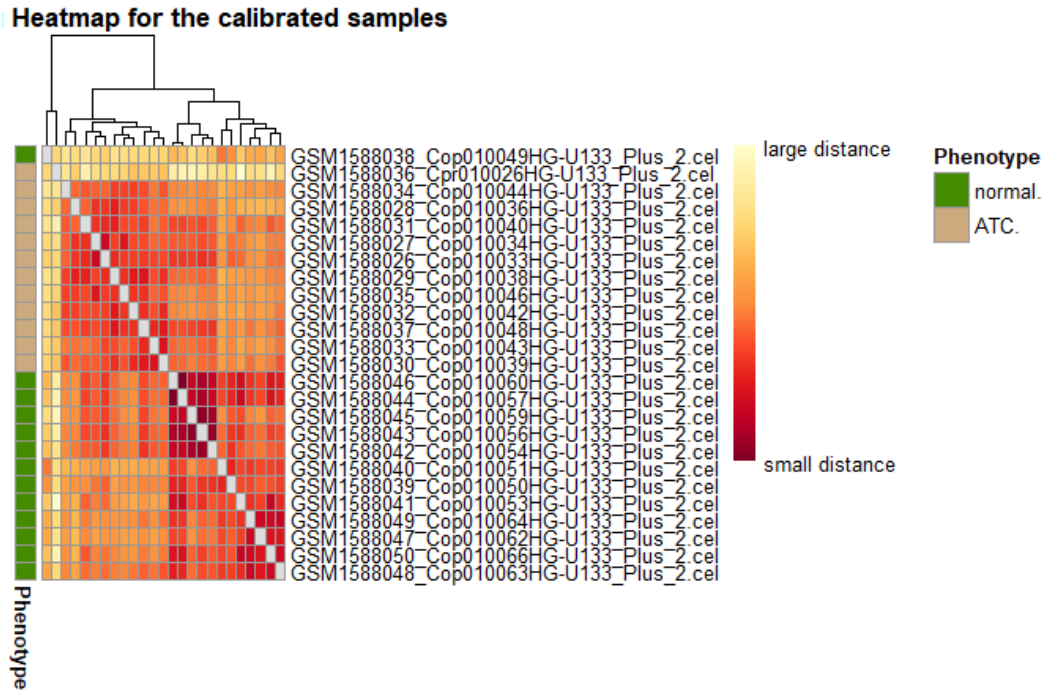


Figure 4.5. Heatmap of the summarised data (E-GEOD-65144).

Enhanced volcano-plot was generated to visualize the deferentially expressed genes (DEGs). In Figure 4.6, Log 2 transformed fold changes (log2FC) are plotted on the x-axis while the negative log of p-values are plotted on the y-axis. Each cell in a plot is a gene. The cells of the plot are color coded. The threshold for p-values is 0.05 whereas log2FC is 0.5. In negative log scale, smaller p-values appears at top whereas the higher p-values are at the base of the y-axis.

In this plot, grey color of cells indicate that cells have not crossed both the set criteria of p-value and fold change. The green cells have only passed the fold change criterion. Red cells are the significant DEGs as they have crossed both the thresholds of significance and fold change. The negative value of fold change indicates the under-expression of genes w.r.t normal while positive fold change indicates the over-expression of genes w.r.t normal reference. The plot shows relatively same number of over-expressed and under-expressed genes in ATC w.r.t normal condition.

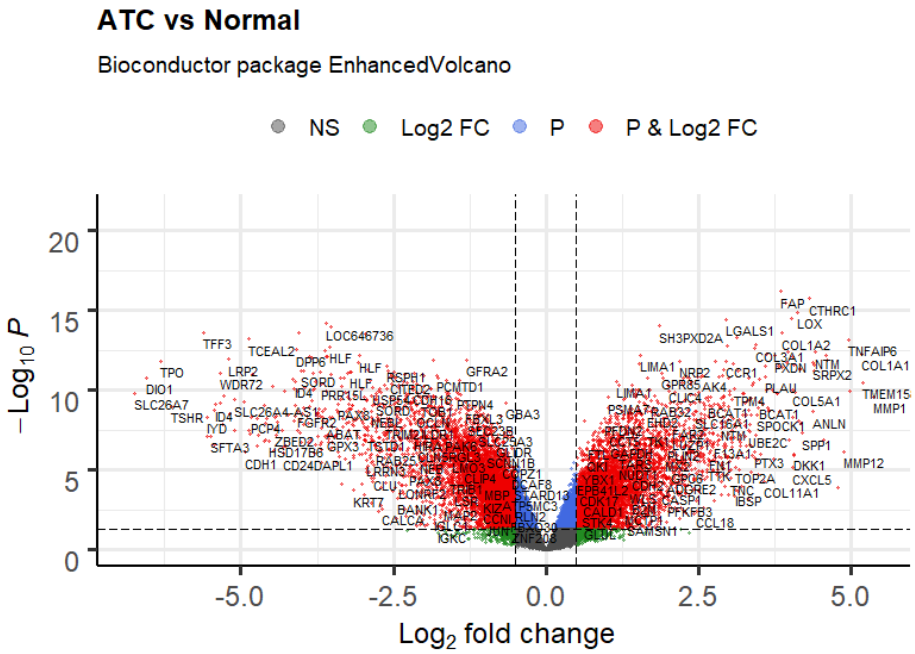


Figure 4.6. Enhanced Volcano plot of the summarised data (E-GEOD-65144).

Through differential expression analysis 49809 DEGs are obtained out of which top10 DEGs are shown in Table 4.1. For these ten DEGs, probe IDs along with gene symbols are given along with the log2 fold change and p.value. The negative and positive log2FC values are showing the down- and up-regulation of the DEGs.

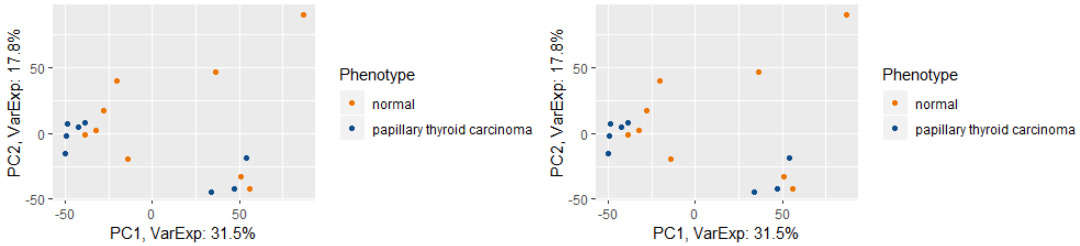
Table 4.1. Top 10 DEGs from dataset 1 (E-GEOD-65144).

S.NO	*PROBEID*	*SYMBOL*	*logFC*	*P.Value*
1	240008_at	NA	0.871986	1.72E-06
2	227498_at	SOX6	0.637477	8.91E-06
3	242146_at	SNRPA1	1.130373	3.45E-05
4	238620_at	NA	0.526659	5.95E-05
5	226041_at	NAPEPLD	0.696292	6.56E-05
6	218340_s_at	UBA6	0.988885	0.000208
7	202057_at	KPNA1	0.540031	0.000248
8	238459_x_at	SPATA6	0.708539	0.000309
9	219584_at	PLA1A	-0.56462	0.000345
10	220770_s_at	ZBED8	0.664875	0.000347

4.1.2 Microarray Dataset 2 (E-GEOD-3467)

The dataset 2 belongs to papillary thyroid carcinoma with the accession number E-GEOD-3467 and it has seventeen samples. Among these, eight samples represent the phenotype of tumor whereas other nine samples are of non-tumor nature. We followed the same pipeline as we did for the first dataset. First we compared raw with calibrated data using PCA plots, boxplots, and RLE plots. After this comparison heatmap and volcano plots were used to highlight the characteristics of the calibrated data.

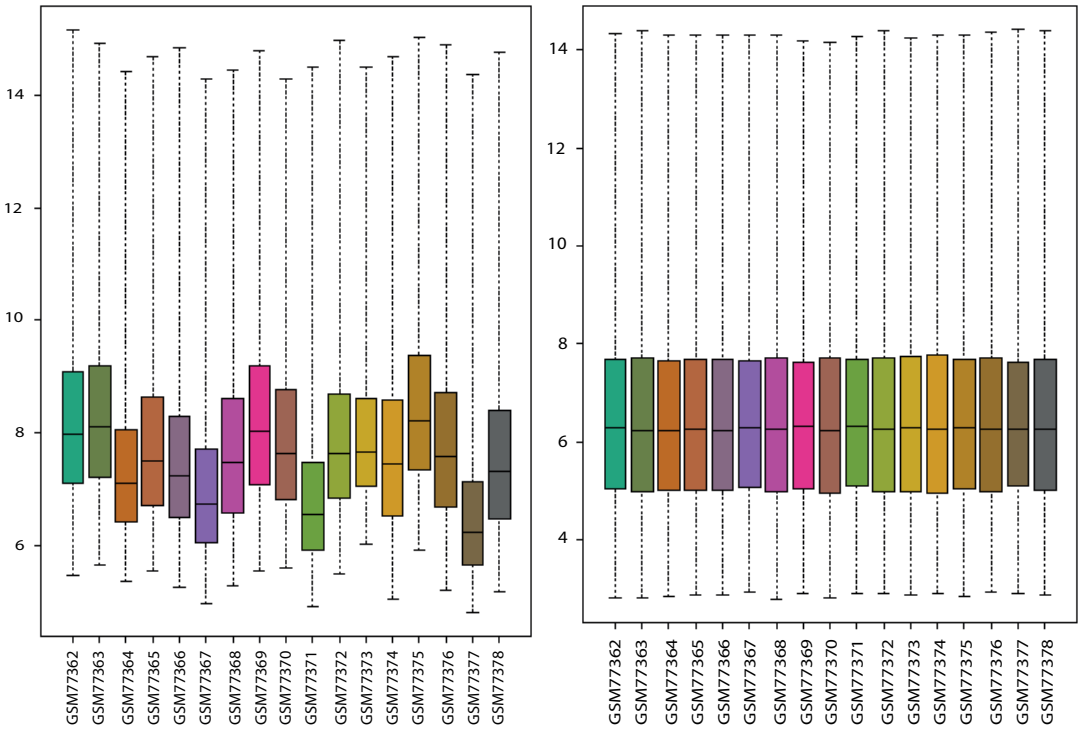
Figures 4.7**A** and 4.7**B** shows the PCA plots for the raw and calibrated data for this dataset. There was no difference between the two plots. In this dataset the calibration did not affect the distribution of the samples as the sample were fairly random and there was no clear categorization of two phenotypes.



(a) PCA plot of raw data. **(b)** PCA plot of calibrated data.

Figure 4.7. (A) PCA plot of raw data (B) PCA plot of calibrated data (E-GEOD-3467).

Although the calibration did not help in categorizing the phenotypes it reduced the variation in the sample medians. A significant improvement regarding sample medians was achieved via calibration. Figures 4.8**A** and 4.8**B** represents the box plot of raw and calibrated data respectively.

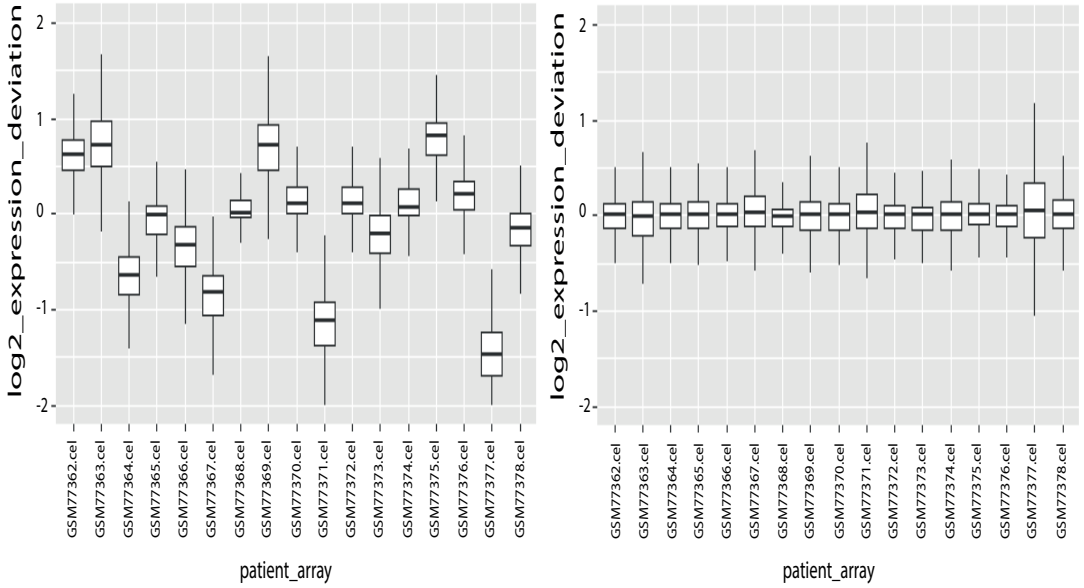


(a) Boxplot of raw data. **(b)** Boxplot of calibrated data.

Figure 4.8. (A) Boxplot of raw data & (B) Boxplot of calibrated data (E-GEOD-3467).

Another advantage of calibrating raw data from this dataset can be highlighted using RLE plots. As shown in figures 4.9**A** and 4.9**B**, the sample medians were more

symmetric around zero for the calibrated data as compared to the raw data.



(a) RLE plot of raw data. **(b)** RLE plot of calibrated data.

Figure 4.9. (A) RLE plot of raw data & (B) RLE plot of calibrated data (E-GEOD-3467).

After looking at the distribution of two phenotypes in this dataset it was obvious that similar phenotypes are not clustered together (see figures 4.7**A** and 4.7**B**). To further investigate the data, heatmap for this dataset was generated to understand sample to sample distance relationship. This heatmap is shown in Figure 4.10 which depicts that some tumor samples are not clustered with other tumor samples and share large distances (lighter tone). Similarly, there were few samples of normal state that were not clustered with other normal samples.

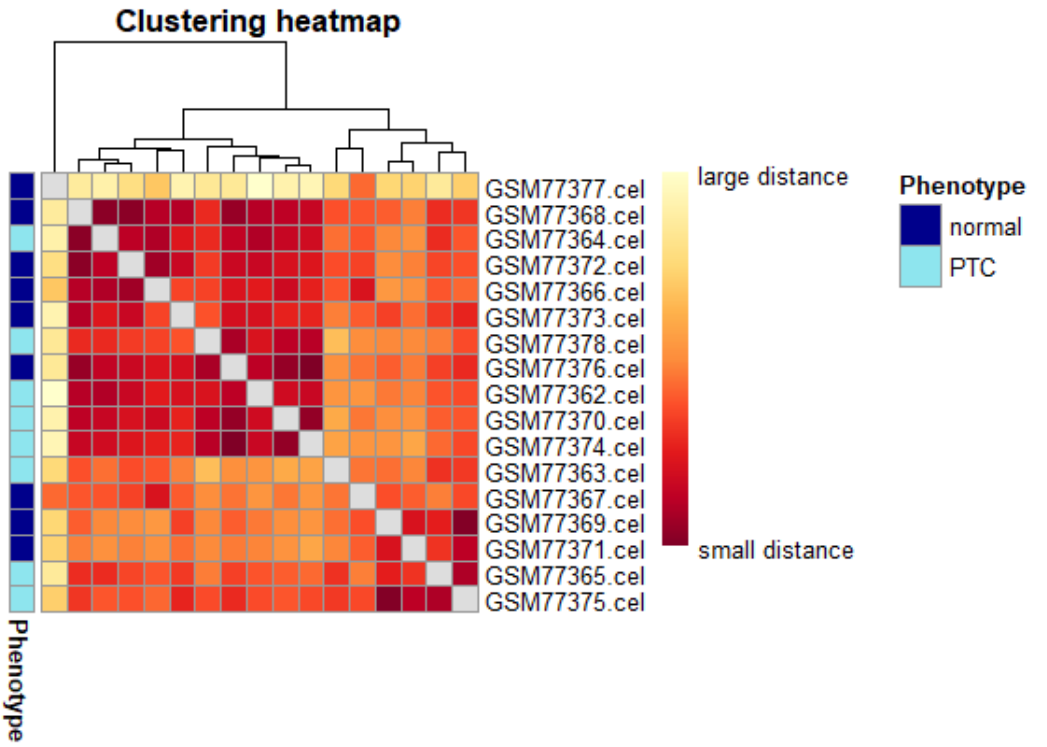


Figure 4.10. Heatmap of the summarised data (E-GEOD-3467).

Figure 4.11 represents the volcano plot of E-GEOD-3467 where threshold for p-values is 0.05 and log2FC is 0.5. As the over-expression increases, so does the value of positive fold changes. The negative sign of fold change is the indicator of under-expression of the genes in disease state w.r.t reference. The number of DEGs would increase if the tumor samples were analysed with reference to normal individuals.

Figure 4.11. Enhanced Volcano Plot of the summarised data (E-GEOD-3467).

Top 10 deferentially expressed genes have been shown in Table 4.2.

Table 4.2. DEGs of phenotype Normal vs PTC of E-GEOD-3746.

S.No	*PROBEID*	*SYMBOL*	*logFC*	*P.Value*
1	240008_at	NA	0.871986	1.72E-06
2	227498_at	SOX6	0.637477	8.91E-06
3	242146_at	SNRPA1	1.130373	3.45E-05
4	238620_at	NA	0.526659	5.95E-05
5	226041_at	NAPEPLD	0.696292	6.56E-05
6	218340_s_at	UBA6	0.988885	0.000208
7	202057_at	KPNA1	0.540031	0.000248
8	238459_x_at	SPATA6	0.708539	0.000309
9	219584_at	PLA1A	-0.56462	0.000345
10	220770_s_at	ZBED8	0.664875	0.000347

4.1.3 Microarray Dataset 3 (E-GEOD-40807)

This study also includes the analysis of microRNA (miRNAs). The messenger RNA (mRNA) targets of the deferentially expressed miRNA are identified using miRwalk database. On dataset E-GEOD-40807, analysis was performed on eighty samples of medullary thyroid carcinoma where forty tumor and 40 normal samples were selected. As the platform is Agilent, the data was normalized before the analysis. The PCA plot, RLE plot and the box plots of the normalized data were then generated to look at the distribution of the phenotypes.

Figure 4.12 shows the PCA-plot, which was generated after the data was normalized. In this plot, the different phenotypes were separately clustered along the first principal component (PC1).

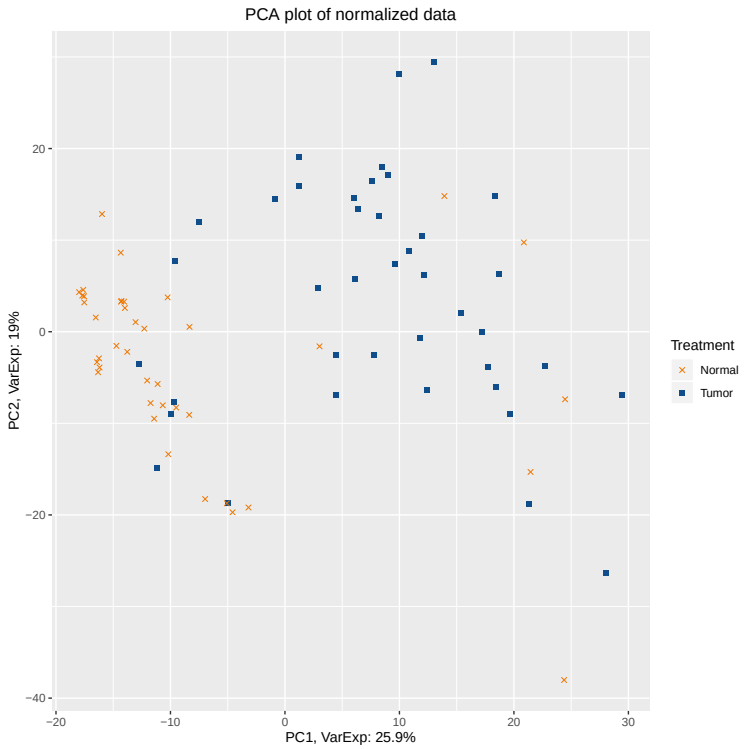


Figure 4.12. PCA plot of the summarised data (E-GEOD-40807).

The RLE-plot was then generated to look at the distribution of sample medians in the normalized data. Figure 4.13 shows that the medians of the normal and tumor samples were aligned around zero with a normal distribution.

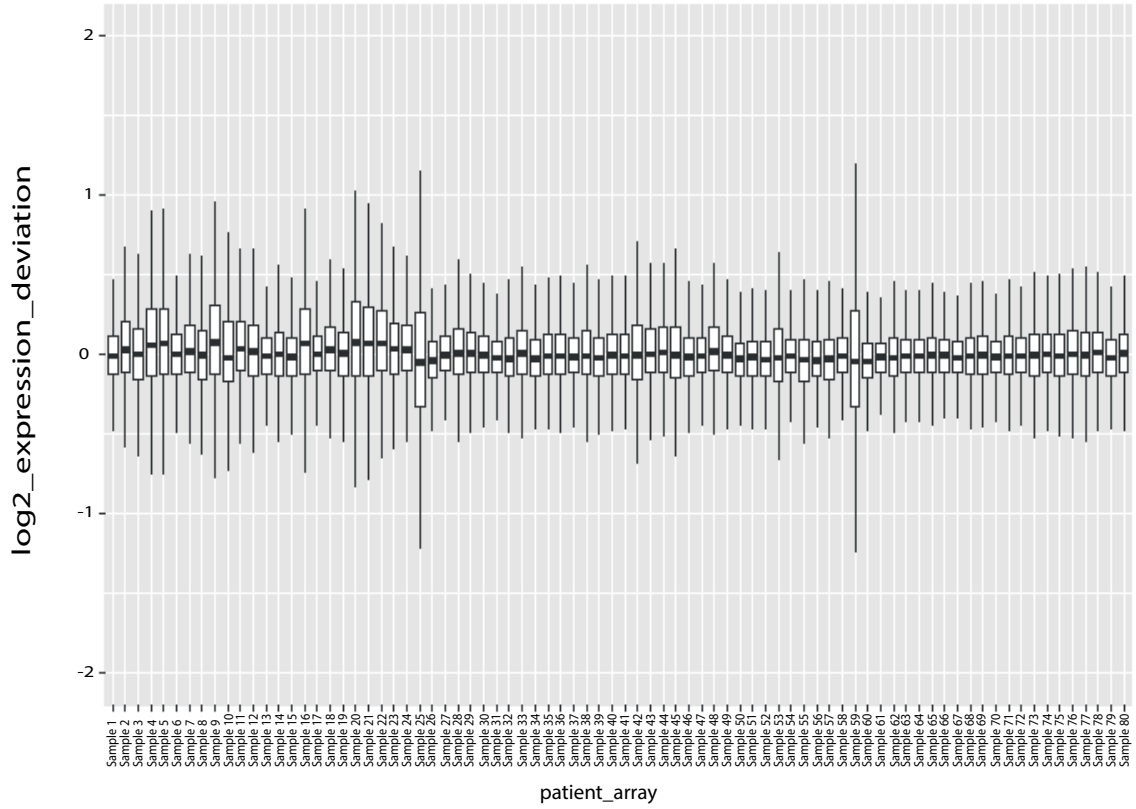


Figure 4.13. RLE plot of the summarised data (E-GEOD-40807).

The box plot of the normalized data also showed that the sample medians are normally distributed As shown in Figure 4.14 the Quartile 1 and 2 along with maximum and minimum are normally distributed and the medians of all samples are approximately at five.

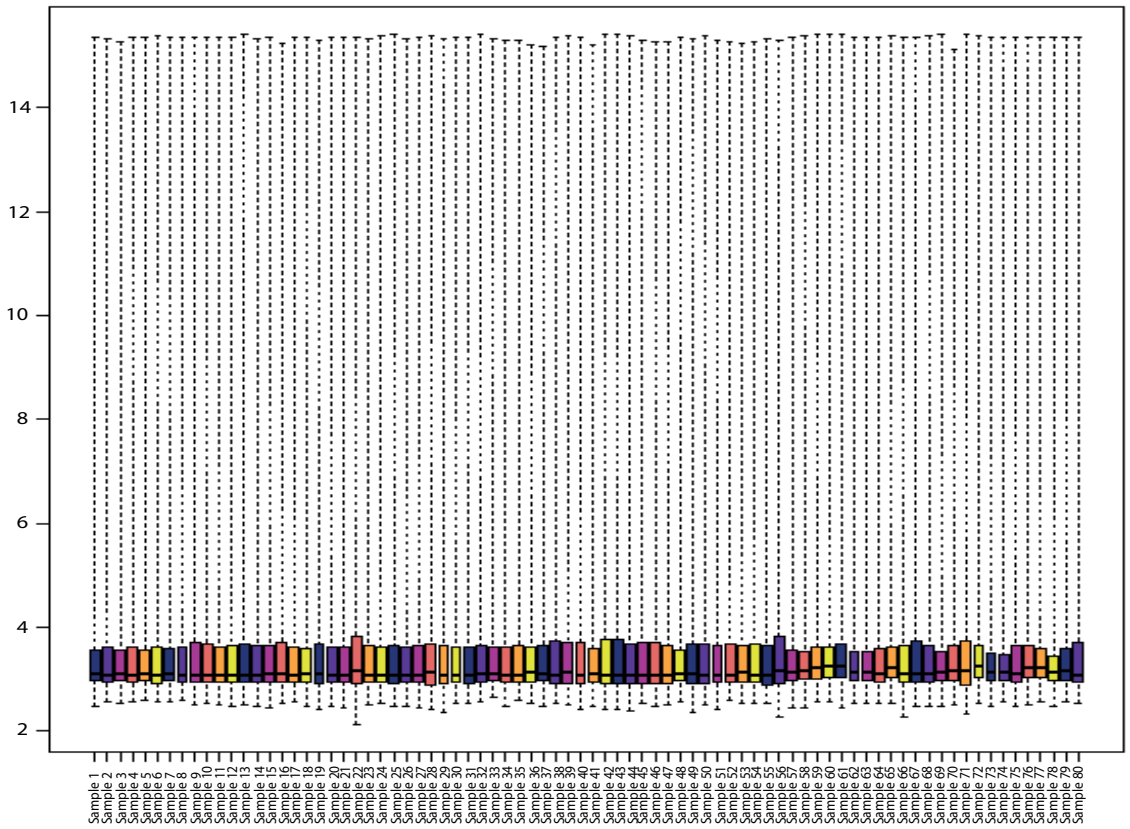


Figure 4.14. Boxplot of the summarised data (E-GEOD-40807).

After highlighting the effects of normalization of the data we started our analysis on the normalized data. An enhanced volcano-plot was generated where the threshold for p-value was 0.05 and log2FC was 0.5.

Figure 4.15 highlights the deferentially expressed genes (DEGs) by their probe IDs. Right side of the plot represents up-regulated DEGs while left side are down-regulated DEGs with negative value of log2FC.

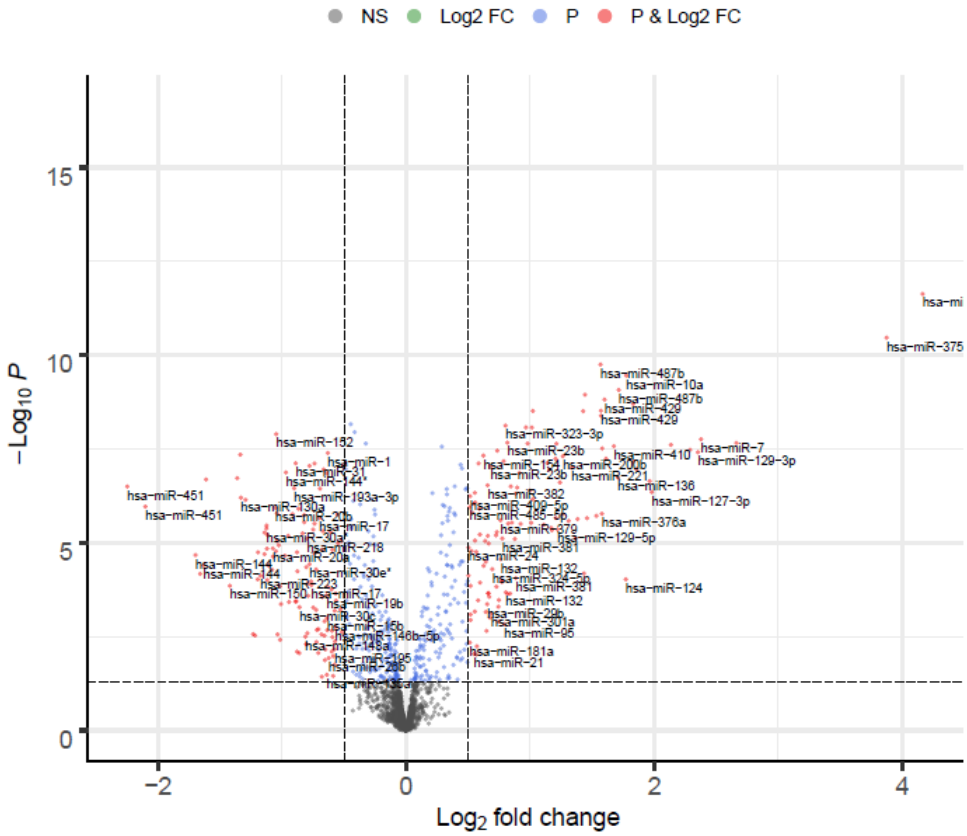


Figure 4.15. Enhanced Volcano plot of the summarised data (E-GEOD-40807).

Total of 2466 DEGs were retrieved from differential expression analysis out of which top10 DEGs are mentioned in Table 4.3.

Table 4.3. miRNAs of phenotype Normal vs MTC of E-GEOD-40807.

S.NO	***SystematicName***	***logFC***	***P.Value***
1	hsa-miR-375	4.162761	2.35E-12
2	hsa-miR-375	3.872235	3.43E-11
3	hsa-miR-487b	1.566938	1.77E-10
4	hsa-miR-10a	1.773394	3.51E-10
5	hsa-miR-487b	1.71407	8.46E-10
6	hsa-miR-200a	1.442414	1.14E-09
7	hsa-miR-429	1.599356	1.52E-09
8	hsa-miR-153	1.830175	2.00E-09
9	hsa-miR-429	1.569651	3.03E-09
10	hsa-miR-136*	1.021476	3.08E-09

Targets of miRNAs has been predicted through miRWalk with the cutoff of 0.05 for pvalue. Few targets of miRNAs have been shown Table 4.4

Table 4.4. miRNAs targets of E-GEOD-40807 using miRWalk.

miRNAs	***Targeted mRNAs***
hsa-miR-375	-
hsa-miR-133b	RAB3B
hsa-miR-129-5p	VAMP1
hsa-miR-127-3p	RGS14
hsa-miR-429	BTBD11
hsa-miR-557	TCN2
hsa-miR-142-3p	SYT2

4.2 Results of RNA-seq Data Analysis

The results of RNA-seq data analysis of Thyroid Carcinoma are discussed below in pictorial and tabular forms.

4.2.1 RNA-seq Dataset 1 (E-GEOD-64912)

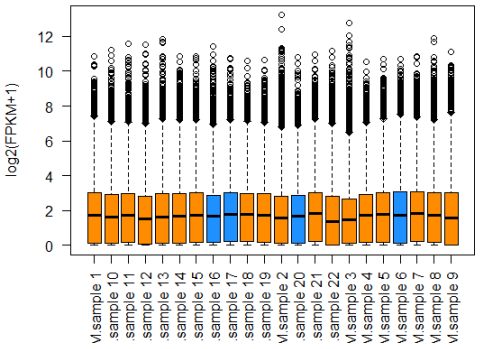
Analysis is performed on 22 Homo sapiens samples with an accession number E-GEOD-64912. It includes four normal and eighteen diseased samples.

In Figure 4.16**A** Boxplot represents distribution of samples on basis of FPKM expression values of the data. In this plot, each box represents normal and diseased samples on x-axis while on y-axis are the FPKM expression values displaying variation in the data. Mid layer in each box of boxplot represents medians of expression values while upper and lower layers represents first and third quartile. It shows us outliers in the data that are represented by dotted lines. Medians of all the samples are lying around 2.

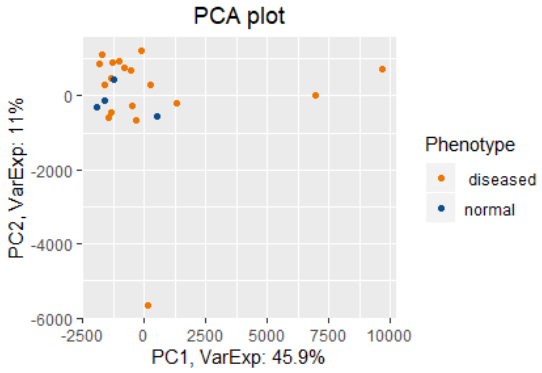
Figure 4.16**B** represents principle component analysis (PCA) which is used for dimension reduction where the expression data is clustered according to their groups (i-e; normal and diseased samples). But large number of diseased samples are close enough to normal samples because they are highly correlated to each other. These samples can be excluded but this is not a good approach because exclusion of the such samples may lead to lose some information. Therefore, both PC1 and PC2 fails to differentiate the diseased samples from the normal ones. So PCA is not the suitable technique to differentiate the phenotypes in that scenario.

Figure 4.16**C** depicts a heatmap that represents expression patterns of genes across all samples along with distance. Manhattan distance is used to cluster genes and samples based on expression patterns and displayed with the help of dendrogram. The map is also annotated with the phenotype. Distance among the diseased samples are coded with light color intensity (top right) indicates large distance which represents that

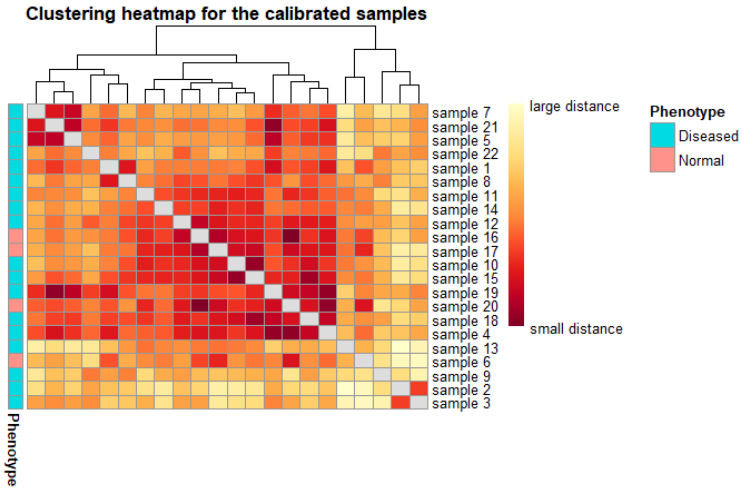
these are less closely related to each other than normal. While on top left of the plot, distance from diseased to normal samples are coded with high color intensity (red) indicates small distance (closely related).



(a) Boxplot



(b) PCA plot



(c) Heatmap

Figure 4.16. Boxplot (A), PCA plot (B) and heatmap (C) of E-GEOD-64912.

In Figure 4.17, the enhanced volcano-plot is showing deferentially expressed genes (DEGs) where log2FC on the x-axis is plotted against negative log of p.value on the y-axis. DEGs are visualized by giving a threshold of p.value equals to 0.05 and log2FC equals to 0.5. The DEGs (red color) on left side of the plot are down-regulated while the ones on right side of the plot are up-regulated. The plot shows relatively higher number of over expressed genes w.r.t normal condition.

Figure 4.17. Enhanced volcano plot of E-GEOD-64912.

26035 DEGs are obtained in the differential expression analysis process, out of which top10 DEGs are listed in Table 4.5.

Table 4.5. Top 10 DEGs of phenotype Normal vs Diseased of E-GEOD-64912.

S.no	*geneNames*	*Transcript-IDs*	*P-value*	*Log2FC*
1	ISG15	ENSG00000187608.10	0.027034	-0.91855
2	TNFRSF4	ENSG00000186827.11	0.006495	-0.99232
3	UBE2J2	ENSG00000160087.20	0.046834	0.71336
4	INTS11	ENSG00000127054.20	0.019537	-1.16895
5	INTS11	ENSG00000127054.20	0.02912	1.642657
6	CPTP	ENSG00000224051.7	0.000343	-0.91356
7	DVL1	ENSG00000107404.20	0.013896	-1.575
8	MXRA8	ENSG00000162576.16	0.029839	-2.17209
9	VWA1	ENSG00000179403.12	0.015297	-0.98636
10	AL691432.2	ENSG00000272106.1	0.005452	0.894066

4.2.2 RNA-seq Dataset 2 (GSE57780)

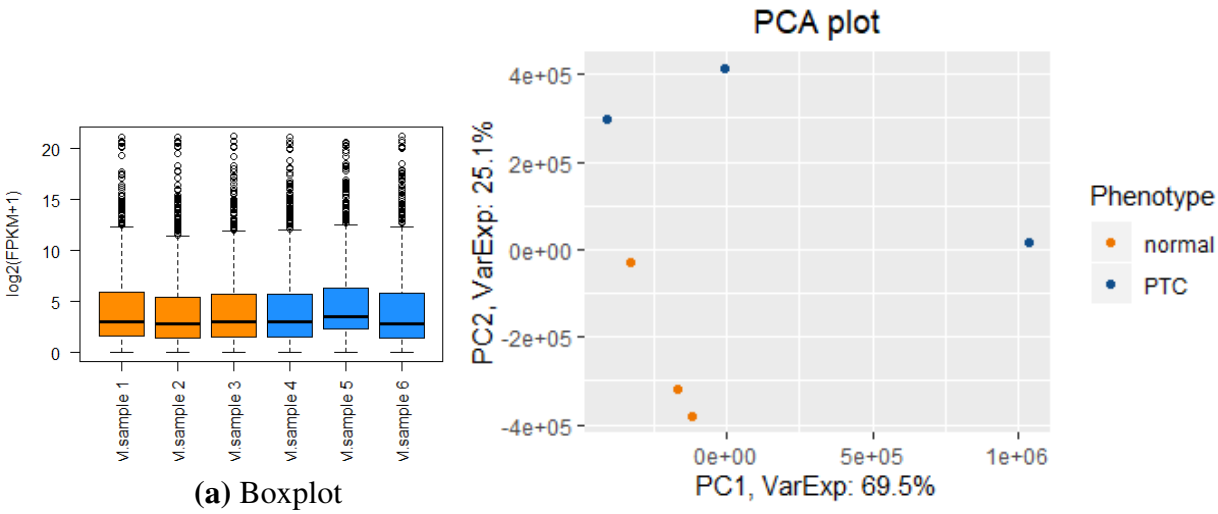
This dataset is comprised of nine samples. Among these, three samples represent the phenotype of tumor, other three depicts normal and remaining three samples are of metastatic nature. To ease our analysis, dataset is divided into two parts. In first part analysis is performed on normal vs tumor samples while in other part analysis is performed on normal vs metastatic samples.

(A) Normal vs Tumor (GSE57780)

In Figure 4.18**A**, distribution of samples is represented with boxplot where median of the expression data almost lie at the same point 2.5. Each box in the plot symbolizes one sample. Outliers are also detected in the plot.

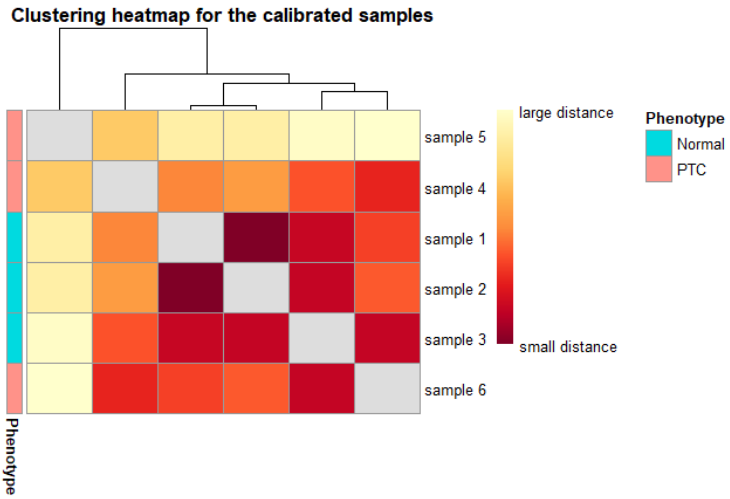
Figure 4.18**B** represents PCA plot of GSE57780 where each cell in PCA plot accounts for one sample. As shown in plot, the sample size is small therefore we can easily discriminate between two phenotypes i.e. normal (orange) and tumor (blue).

In Figure 4.18**C**, heatmap shows distance between the samples where Manhattan distance is applied. The diagonally arranged boxes express the distance of specific sample from its self that must be zero. The dendogram at the top of the heatmap depicts closeness and the order of clustering. It is cleared from the top left of the plot that distance among the tumor samples is indicated by low color intensity depicts large distance (less closely related) as compared to normal samples represented by high color intensity. Likewise, at bottom right of the plot, distance from tumor to normal is indicated by high color tone (red) shows more closely relation.



(a) Boxplot

(b) PCA plot



(c) Heatmap

Figure 4.18. Boxplot (A), PCA plot (B) and heatmap (C) of GSE57780 (Normal vs Tumor).

In enhanced volcano-plot, log2FC is plotted against negative log of p.value displaying deferentially expressed genes (DEGs) as shown in Figure 4.19. A threshold of negative log of p.value equal to 0.05 and log2FC equals to 0.5 is set for obtaining DEGs. The red cells with positive fold change values are up-regulated and those with negative values of fold change are down regulated genes.

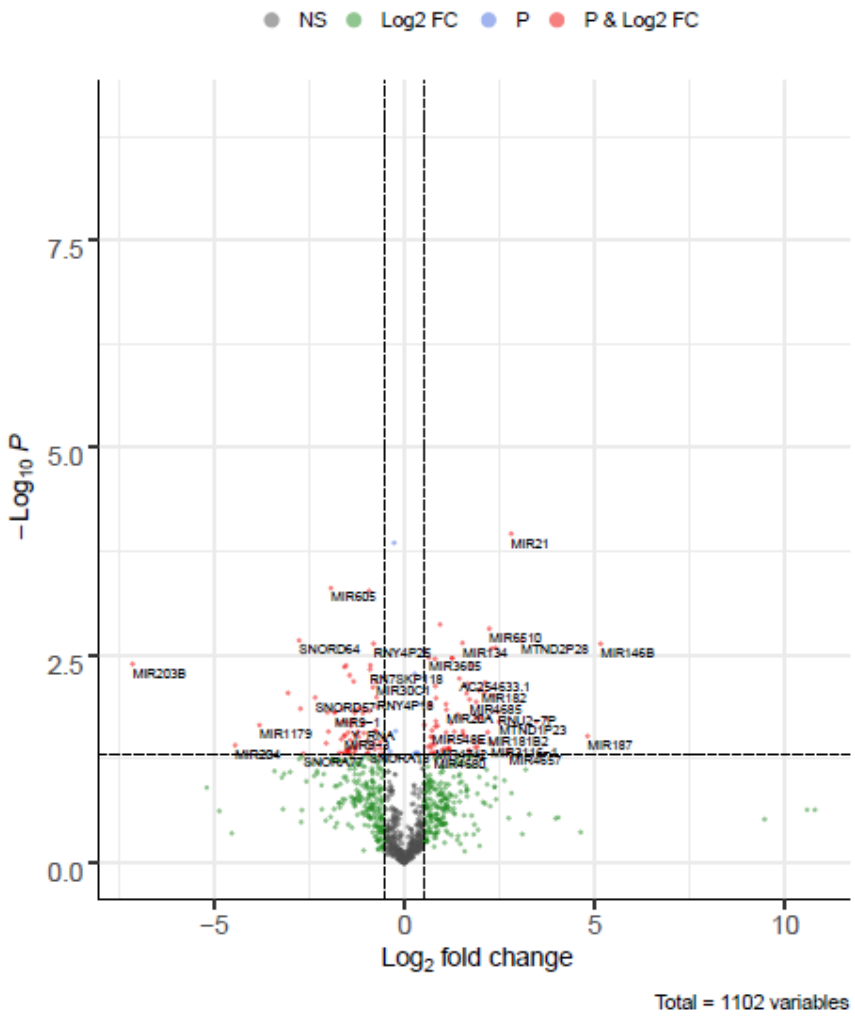


Figure 4.19. Enhanced volcano plot of GSE57780 (Normal vs Tumor).

The plot shows nearly equal number of over-expressed and under-expressed genes. Top10 out of 1102 DEGs are listed in Table 4.6.

Table 4.6. Top 10 DEGs of phenotype Normal vs Tumor of GSE57780 (Normal vs Tumor).

S.No	*geneNames*	*transcriptNames*	*P-value*	*Log2FC*
1	MTND1P23	ENST00000416931.1	0.019897	2.494101
2	MTND2P28	ENST00000457540.1	0.002122	3.071711
3	MIR6730	ENST00000622213.1	0.002575	2.390667
4	AC254633.1	ENST00000606790.1	0.006038	1.444086
5	MIR3605	ENST00000583214.3	0.003397	0.646543
6	MIR30C1	ENST00000385227.1	0.006684	-0.74109
7	MIR3116-1	ENST00000584654.1	0.035682	2.276628
8	MIR3116-2	ENST00000636415.1	0.036797	2.338783
9	RNU1-120P	ENST00000363009.1	0.020832	1.228026
10	RNY4P25	ENST00000459254.1	0.002299	-0.8075

Targets of miRNAs has been predicted through miRWalk with the cutoff of 0.05 for P-value. Few targets of miRNAs have been shown Table 4.7

Table 4.7. miRNAs for Normal vs Tumor (GSE57780).

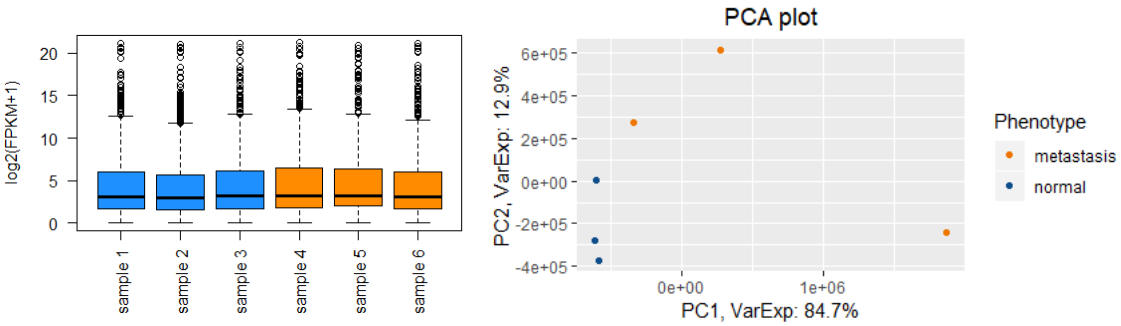
Micro RNAs	*Targeted mRNAs*
MIR1179	METTL8
MIR1181	DPP8
MIR3154	IYD
MIR3605	-
MIR30C1	-
MIR3929	PDS3

(B) Normal vs Metastasis (GSE57780)

Figure 4.20**A** depicts the boxplot of log 2 transformed data of all samples. It is cleared from the plot that medians of all samples are lying around a single point 2.5.

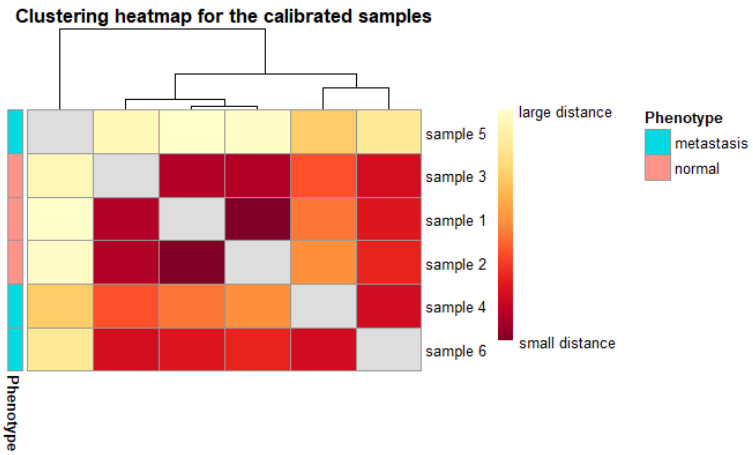
Figure 4.20**B** represents Principal Component Analysis of the data which is performed for the dimension reduction of high dimensional data. PCA has transformed the correlations between all samples in a 2-D graph. Plot shows that PC1 is discriminating the data with 84.7% variance.

Figure 4.20**C** depicts the heatmap clustering analysis of GSE57780. It is cleared at the bottom right from the heatmap, that distance from the metastatic samples to normal samples is small indicated by high color intensity (red). Similarly, at top left of the plot distance among the diseased samples is large, indicated by low color tone (towards yellow) and distance among normal samples is represented by high color intensity. The heatmap nicely clusters the samples in that scenario.



(a) Boxplot

(b) PCA plot



(c) Heatmap

Figure 4.20. Boxplot (A), PCA plot (B) and heatmap (C) of GSE57780 (normal Vs metastasis).

Figure 4.21 shows enhanced volcano plot between normal and metastasis samples, where log2FC is plotted against negative log of p.value. To visualize the deferentially expressed genes (DEGs), a threshold of negative log of p.value equals to 0.05 and log2FC equals to 0.5 is applied. The positive value of fold change indicates over-expression and negative value depicts under expression of genes w.r.t normal reference. The plot shows relatively equal number of up regulated and down regulated genes. Top10 out of 1046 DEGs are listed in Table 4.8.

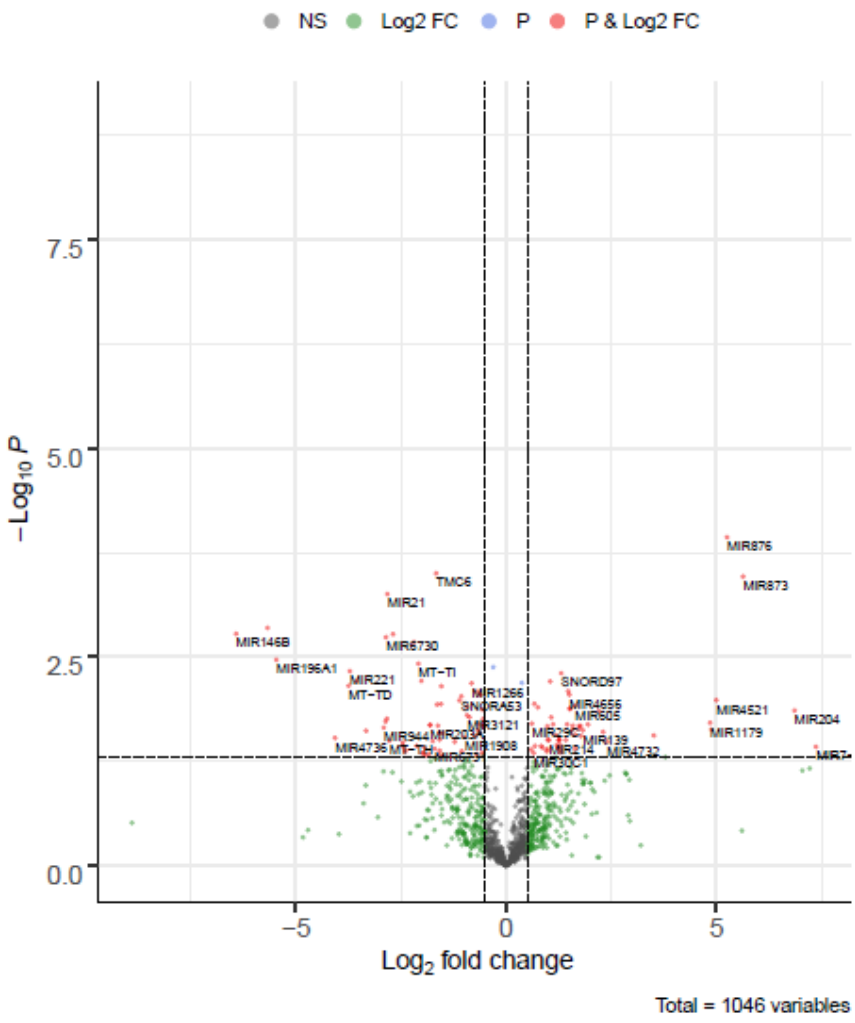


Figure 4.21. Enhanced volcano plot of GSE57780 (normal Vs metastasis).

Table 4.8. Top 10 DEGs of phenotype Normal vs Metastasis of GSE57780.

S.no	*geneNames*	*transcriptNames*	*P-value*	*Log2FC*
1	MIR6730	ENST00000622213.1	0.00184	-2.85628
2	MIR6731	ENST00000614863.1	0.039961	-1.69756
3	MIR30C1	ENST00000385227.1	0.044831	0.657086
4	MIR7156	ENST00000620979.1	0.03105	-1.57877
5	MIR214	ENST00000385214.1	0.031473	1.016545
6	MIR199A2	ENST00000385289.1	0.031377	1.25806
7	MIR3121	ENST00000579680.1	0.015857	-0.9284
8	MIR181B1	ENST00000385240.1	0.048351	-1.91924
9	MIR181A1	ENST00000385026.1	0.046181	-1.8392
10	SNORA77	ENST00000408716.1	0.036127	1.289324

Targets of miRNAs has been predicted through miRWalk with the cutoff of 0.05 for p-value. Few targets of miRNAs have been shown Table 4.9

Table 4.9. miRNAs for Normal vs Metastasis (GSE57780)

Micro RNAs	*Targeted mRNAs*
MIR944	SYN2
MIR599	REL
MIR6730	-
MIR3131	BAHD1
MIR1179	METTL8
MIR1273c	DLC1

4.3 Comparative Analysis of Microarray Datasets

The microarray data analysis was performed on three datasets. Two datasets were of mRNA type while remaining one was of miRNA type. Target genes were found for differentially expressed miRNAs. Deferentially expressed genes (DEGs) for every dataset were obtained at p-value = 0.05 and log2FC = 0.5.

- 280 common DEGs were obtained among all the microarray datasets of thyroid carcinoma shown in Figure 4.22

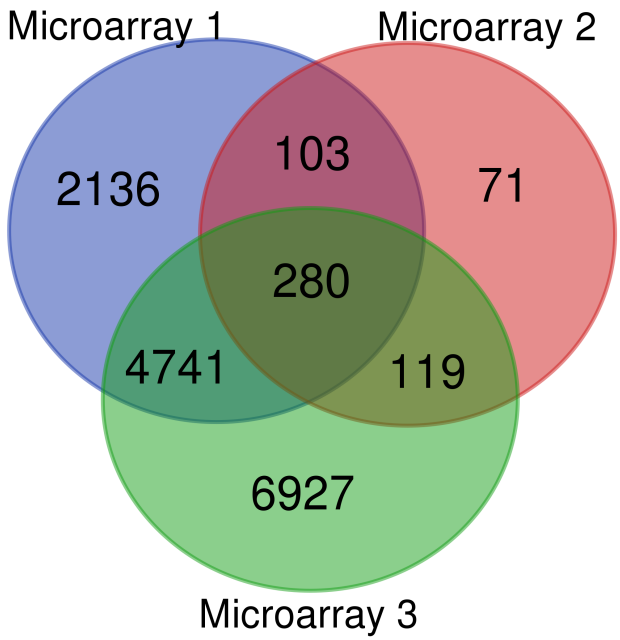


Figure 4.22. Common DEGs among microarray datasets.

4.4 Comparative Analysis of RNA-seq Datasets

Comparative analysis on the DEGs of RNA-Seq analysis was performed. Two datasets were of miRNA type while remaining one was of mRNA type.

- 1247 common DEGs were obtained among all the RNA-seq datasets of thyroid carcinoma shown in Figure 4.23

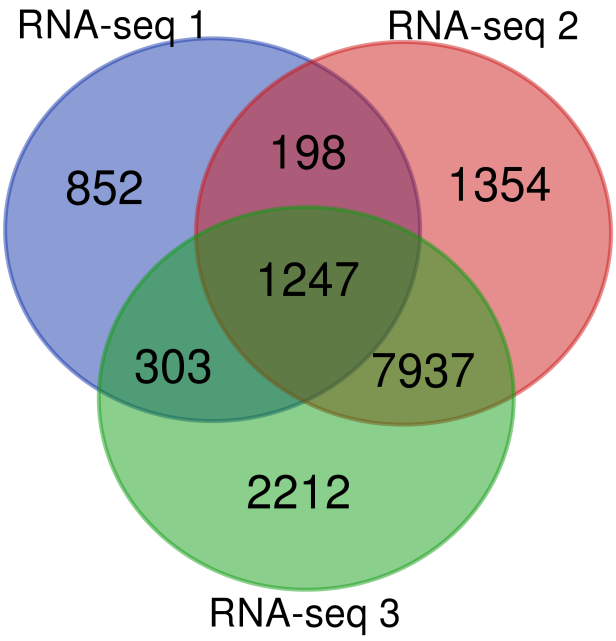


Figure 4.23. Common DEGs among RNA-seq datasets.

However, as depicted in Figure 4.24, 36 common genes were obtained after comparative analysis on DEGs obtained from thyroid carcinoma datasets of microarray and RNA-Seq analysis.

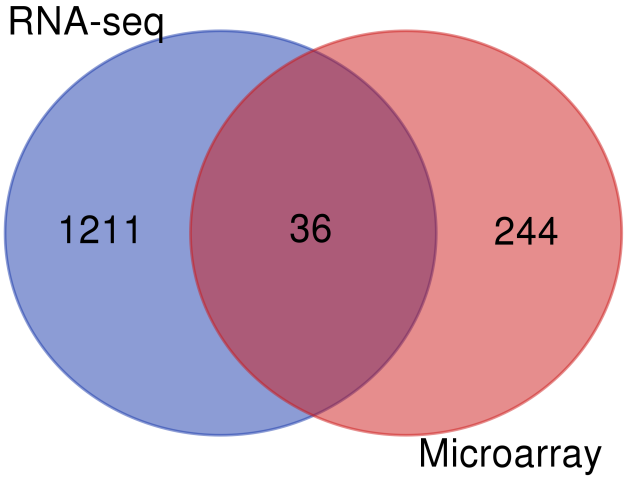


Figure 4.24. Common DEGs among RNA-seq and microarray datasets.

4.5 Pathway Analysis

Pathway analysis for both microarray and RNA-seq datasets was performed via Enrichr. KEGG pathway database is selected to know about common pathways in which the DEGs are involved. Top pathways in order of significance are given below.Tables i.e. 4.10 for E-GEOD-65144, 4.11 for E-GEOD-3467, 4.12 for E-GEOD-40807 and 4.13 for common DEGs among all datasets, shows the pathways involved during microarray analysis of Deferentially Expressed Genes (DEGs) for Thyroid Carcinoma separately.

While Tables such as 4.14 for E-GEOD-64912, 4.15 for GSE57780 (Normal vs Tumor), 4.16 GSE57780 (Normal vs Metastasis) and 4.17 for common DEGs among all datasets, shows the pathways involved during RNA seq analysis of Deferentially Expressed Genes (DEGs) for Thyroid Carcinoma separately.

4.6 Microarray Datasets Pathway Analysis

The list of top 10 pathways involved in all three microarray datasets along with number of genes and P-Value are shown below. Pathways involved in dataset 1 (E-GEOD-65144) of microarray shown in Table 4.10.

Table 4.10. Pathways involved in DEGs obtained for dataset E-GEOD-65144.

S.no	*Pathways*	*No of genes*	*P-value*
1	Cell cycle	74/124	9.51E-08
2	AGE-RAGE signaling pathway	60/100	1.17E-06
3	Proteoglycans in cancer	105/201	2.60E-06
4	FoxO signaling pathway	74/132	2.78E-06
5	Autophagy	72/128	3.19E-06
6	Prostate cancer	57/97	5.39E-06
7	Focal adhesion	101/199	1.98E-05
8	Rap1 signaling pathway	104/206	2.01E-05
9	Endocytosis	120/244	2.33E-05
10	PI3K-Akt signaling pathway	166/354	2.40E-05

Pathways involved in dataset 2 (E-GEOD-3467) of microarray shown in Table 4.11.

Table 4.11. Pathways involved in DEGs obtained for dataset E-GEOD-3467.

S.no	*Pathways*	*No of genes*	*P-value*
1	NF-kappa B signaling pathway	13/95	3.11E-06
2	Chagas disease (American trypanosomiasis)	12/103	3.89E-05
3	Wnt signaling pathway	15/158	5.08E-05
4	NOD-like receptor signaling pathway	16/178	5.61E-05
5	IL-17 signaling pathway	11/93	7.06E-05
6	TNF signaling pathway	12/110	7.46E-05
7	Epithelial signaling in H-pylori infection	9/68	1.34E-04
8	Human cytomegalovirus infection	17/225	2.79E-04
9	Osteoclast differentiation	12/127	2.94E-04
10	Proteoglycans in cancer	13/201	0.00540

Pathways involved in dataset 3 (E-GEOD-40807) of microarray shown in Table 4.12

Table 4.12. Pathways involved in DEGs obtained for dataset E-GEOD-40807.

S.no	*Pathways*	*No of genes*	*P-value*
1	Axon guidance	155/181	7.09E-14
2	MAPK signaling pathway	226/295	1.92E-09
3	Signaling pathways regulating pluripotency of stem cells	116/139	2.83E-09
4	Pathways in cancer	383/530	3.98E-09
5	Proteoglycans in cancer	159/201	9.49E-09
6	ErbB signaling pathway	75/85	1.23E-08
7	Colorectal cancer	75/86	3.93E-08
8	Neurotrophin signaling pathway	99/119	5.67E-08
9	Cellular senescence	128/160	7.65E-08
10	Chronic myeloid leukemia	67/76	8.16E-08

Pathways involved in 280 common DEGs obtained for microarray analysis shown in Table 4.13

Table 4.13. Pathways in common DEGs for microarray.

S.no	*Pathways*	*No of genes*	*P-value*
1	Wnt signaling pathway	7/158	0.006
2	Hippo signaling pathway	7/160	0.007
3	Bladder cancer	3/41	0.019
4	Proteoglycans in cancer	7/201	0.023
5	Valine, leucine and isoleucine degradation	3/48	0.029
6	Platelet activation	5/124	0.030
7	Hepatocellular carcinoma	6/168	0.031
8	Vascular smooth muscle contraction	5/132	0.038
9	mRNA surveillance pathway	4/91	0.038
10	Signaling pathways regulating pluripotency of stem cells	5/139	0.046

4.7 RNA-seq Datasets Pathway Analysis

The list of top 10 pathways involved in all three RNA-seq datasets along with number of genes and P-Value are shown below. Pathways involved in dataset 1 (E-GEOD-64912) of RNA-seq shown in Table 4.14.

Table 4.14. Pathways involved in DEGs obtained for dataset E-GEOD-64912.

S.no	*Pathways*	*No of genes*	*P-value*
1	Adherens junction	23/72	2.42E-05
2	Transcriptional misregulation in cancer	44/186	4.84E-05
3	Pathways in cancer	99/530	1.10E-04
4	Fluid shear stress and atherosclerosis	34/139	1.69E-04
5	Spliceosome	33/134	1.82E-04
6	Small cell lung cancer	25/93	2.54E-04
7	Longevity regulating pathway	26/102	4.79E-04
8	Autophagy	30/128	8.47E-04
9	Colorectal cancer	22/86	0.00119
10	Proteoglycans in cancer	40/201	0.00373

Pathways involved in dataset 2 (GSE57780-Normal vs Tumor) of RNA-seq shown in Table 4.15.

Table 4.15. Pathways involved in DEGs for dataset GSE57780 (Normal vs Tumor).

S.no	*Pathways*	*No of genes*	*P-value*
1	Ras signaling pathway	165/232	3.49E-08
2	Oxytocin signaling pathway	113/153	2.09E-07
3	Pathways in cancer	340/530	4.89E-07
4	Rap1 signaling pathway	145/206	6.28E-07
5	Axon guidance	129/181	8.72E-07
6	MAPK signaling pathway	198/295	1.58E-06
7	Glioma	60/75	1.81E-06
8	Proteoglycans in cancer	140/201	2.50E-06
9	Leukocyte transendothelial migration	83/112	6.69E-06
10	GABAergic synapse	68/89	7.71E-06

Pathways involved in dataset 3 (GSE57780-Normal vs Metastasis) of RNA-seq shown in Table 4.16.

Table 4.16. Pathways involved in DEGs for dataset GSE57780 (Normal vs Metastasis).

S.no	*Pathways*	*No of genes*	*P-value*
1	Axon guidance	145/181	4.83E-10
2	Proteoglycans in cancer	153/201	1.08E-07
3	Cellular senescence	125/160	1.19E-07
4	MAPK signaling pathway	215/295	1.61E-07
5	Adrenergic signaling in cardiomyocytes	112/145	1.51E-06
6	Adherens junction	61/72	1.51E-06
7	Leukocyte transendothelial migration	89/112	2.09E-06
8	Sphingolipid signaling pathway	93/119	4.72E-06
9	ErbB signaling pathway	69/85	7.31E-06
10	Cell adhesion molecules (CAMs)	110/145	8.44E-06

Pathways involved in 1247 common DEGs obtained for RNA-seq analysis shown in Table 4.17

Table 4.17. Pathways involved in common DEGs for RNA-seq analysis.

S.no	*Pathways*	*No of genes*	*P-value*
1	Transcriptional misregulation in cancer	26/186	9.35E-05
2	Mitophagy	13/65	1.59E-04
3	Proteoglycans in cancer	26/201	3.30E-04
4	Autophagy	19/128	3.65E-04
5	Adherens junction	13/72	4.56E-04
6	Pathways in cancer	53/530	4.67E-04
7	Endocytosis	28/244	0.00135
8	Leukocyte transendothelial migration	16/112	0.00155
9	Cellular senescence	20/160	0.00231
10	Fluid shear stress and atherosclerosis	18/139	0.00253

Chapter 5

DISCUSSION

As stated in chapter 1, thyroid carcinoma is increasingly prevailing across all over the world over time. The right and timely diagnosis of carcinoma is a challenge. To tailor treatment appropriately, it is crucial to undertake a proper diagnostic workup before treatment is started (Cabanillas et al., 2016). Therefore, the discovery of therapeutic targets would help in quick diagnosis of disease at molecular level. By this, the survival rate will increase by many folds.

To our understanding from literature, the complexity in thyroid carcinoma was not properly understood, although many metabolic and signaling pathways were reported which elaborates the mechanism and pathogenicity of the disease to some extent. But there were some loop-holes in understanding the molecular basis of the disease associated with their etiological agents. Expression profiling of mRNAs and miRNAs gain importance after advancement in high throughput sequencing techniques as it enlightens the path for researchers to find potential therapeutic targets of the disease. High throughput sequencing techniques make it very easy to understand the pathogenicity of the disease. There is a need to completely understand the pathogenicity of the disease and to identify biomarkers and potential therapeutic targets for the disease. Different studies reported many genes and miRNAs as biomarkers but they didn't use multiple data of different regions and also did not use integrated sequencing analysis techniques to predict the biomarkers and therapeutic targets.

Biomarkers can be categorized into four types: diagnostic , prognostic, therapeutic and predictive. A diagnostic biomarker enables the cancer to be identified early in a non-invasive manner, and hence the secondary cancer prevention. A predictive biomarker enables the prediction of the patient's reaction to a targeted therapy and thus identifies sub-populations of patients likely to benefit from a specific therapy.

A prognostic biomarker is a clinical or biological function that provides information about the disease's probable course; it provides information about the patient's outcome. In general, a therapeutic biomarker is a protein that could be used as a therapy target (Carlomagno et al., 2017). A prognostic biomarker provides information on the overall cancer outcome of patients irrespective of treatment (Oldenhuis et al., 2008).

The study was designed to interpret the meaning behind the microarray and high throughput data of thyroid carcinoma. The aim was to predict the potential therapeutic targets for the disease. In order to acheive goals and objective of the study, microarray data analysis using R based published pipeline and RNA-seq data analysis using Galaxy were performed.

Pathway analysis was then performed to interpret and detect the effect of differential expression of group of genes in disease state at biological network level using DAVID and Enrichr software. The purpose of using datasets of different regions is to check any kind of genetic variations that may be different from region to region. Although microarray and RNA-seq have already performed on selected datasets at experimental level but they did not compare their results of differential expressions with other datasets of the same disease. We performed the whole analysis again by using some computational and statistical techniques i.e. the use of different quality checks, algorithms for normalization and clustering to statistically visualize and analyze the differentially expressed genes and miRNAs. We have generated different plots to check the distribution and variation of samples. Box plot shows the distribution of the data between upper quarties, medians, lower quartiles, minimum and maximum values and also give information about outliers if exist.

Comparative analysis was then performed on all microarray datasets with threshold of p-value=0.05 and log2FC=0.5 shown on Figure 4.22, 280 common DEGs were obtained among all the microarray datasets of thyroid carcinoma. While in Figure 4.23, 1247 common DEGs were obtained among all the RNA-seq datasets of thyroid carcinoma when comparative analysis was performed. DisGeNET is a flexible tool that can

be used for various research purposes, including the analysis of the genetic implications and co-morbidities of human diseases, the study of the characteristics of disease genes, the production of hypotheses on the therapeutic action of medicinal products and the adverse effects, and the validation of disease genes which are predicted computationally (Pinero et al., 2020). Therefore to check the association between gene and its corresponding disease, DAVID and DisGeNet plateforms were used. As illustrated in Figure 4.24, comparative analysis between microarray and RNA-seq datasets revealed 36 common significant genes i.e. (discs large MAGUK scaffold protein 1) DLG1, (ADP ribosylation factor like GTPase 1) ARL1, (solute carrier family 39 member 14) SLC39A14, (atlastin GTPase 2) ATL2, (zinc finger protein 148) ZNF148, (UBX domain protein 4) UBXN4, (DLG associated protein 4) DLGAP4, (dead-box helicase 40) DHX40, (N-acylsphingosine amydohydrolase 1) ASAH1, (mucin 15, cell surface associated) MUC15, (jade family PHD finger 1) JADE1, (Ras related glycolysis inhibitor and calcium channel regulator) RRAD, (zinc finger protein 644) ZNF644, (thioredoxin like 1) TXNL1, (centosomal protein 68) CEP68, (interleukin enhancer binding factor 3) ILF3, (syntrophin beta 2) SNTB2, (protein inhibitor of activated stat 1) PIAS1, (transgelin) TAGLN, (eukaryotic translation initiation factor 5) EIF5, (serine and arginine rich splicing factor 10) SRSF10, (SMAD family member 1) SMAD1, (scavenger receptor class B member 2) SCARB2, (deiodinase, iodothyronine type 2) DIO2, (Rho GTPase activating protein 5) ARHGAP5, (solute carrier family 25 member 36) SLC25A36, (myeloid/lymphoid or mixed lineage leukemia; translocated to 10) MLLT10, (dead-box helicase 17) DDX17, (activity regulated cytoskeletal) ARC, (heterogeneous nuclear ribonucleoprotein A3) HNRNPA3, (prolyl endopeptidase like) PREPL, (zinc finger and BTB domain containing 44) ZBTB44, (elongator acytyltransferase complex subunit 2) ELP2, (caldesmon 1) CALD1, (lamin A/C) LMNA and (solute carrier family 4 member 4) SLC4A4. Out of these 36 differentially expressed genes, only ZBTB44 was not considered a prognostic therapeutic target for thyroid cancer but for other carcinomas patients which needs further in-

vestigation to overcome the disease. While remaining were also validated through literature.

The genes with highest significance (lowest p-value) from every subset of dataset are stated in previous chapter. Those genes could be regarded as potential therapeutic targets for thyroid carcinoma. FAP, CTHRC1, LOX, NA, SOX6, SNRPA1, EPB41L4A, RGS14, BLOC1S5 were the highest significant genes from the differential expression results of microarray data analysis, which could be proposed as potential therapeutic targets to diagnose the disease. The previous studies also validated the role of FAP, LOX, NA, SOX6, SNRPA1 in tumor formation in thyroid carcinoma (Sada et al., 2019), (Pan et al., 2019). ISG15, TNFRSF4, UBE2J2, ARHGAP19, METTL8, TBX22 were also proposed as potential therapeutic targets in RNA-seq data analysis. A study proposed ISG15 as significant potential biomarker for thyroid carcinoma (Lin et al., 2019). The role of UBE2J2 in thyroid cancer prognosis is also reported by the study (Hosseini et al., 2019). A research also reported that METTL8 leads to thyroid cancer progression (Gao et al., 2019). In a recent study TBX22 is proposed as potential therapeutic target in prognosis of papillary thyroid cancer (Chang et al., 2016). One of the findings seem to negate the initial hypothesis that CD74/MIF mediates the connection among both inflammation and thyroid carcinoma. But their findings indicate that CD74 may serve as a therapeutic target in highly developed thyroid carcinoma (Cheng et al., 2015).

Pathway enrichment analysis is performed by utilizing the DEGs acquired from microarray and RNA-seq analysis to find important metabolic and signaling pathways in which these are enriched. Pathways of all datasets have been identified. Comparative analysis on the basis of pathways has been performed where Proteoglycans in cancer, Transcriptional misregulation in cancer, PI3K-Akt signaling pathway, Wnt signalling pathway and MAPK signalling pathway and many other are involved in thyroid carcinoma. While Proteoglycans in cancer is found to be common. Many of the important significant genes in these pathways have been associated with the disease.

A previous study declared that transcriptional misregulation in cancer plays critical role in migration and proliferation of tumor cells which contibutes to understanding of genetic basis of papillary thyroid carcinoma (Ao et al., 2018). Another study reported that an essential constituents of the extracellular matrix, Proteoglycans (PGs), were associated with cancer pathogenicity. Moreover, expression profiles of PGs and their integrating proteins were characterized as being unique to the development of diseases in various forms of cancer. Notably, PGs largely regulate the bio-activity of hormones, growth factors, and cytokines and also the triggering of their specific receptor that control recurrence levels, phenotypic diversity and gene expression in different types of tumours (Baciu et al., 2017) & (Nikitovic et al., 2018). A research also analyzed that pathway (PI3K-PKB/AKT) one of the most important molecular signal transduction involved in key cellular activities. The persistent activation in its downstream effectors by multiple abnormal receptor tyrosine kinases (RTKs) and hereditary abnormalities results in high cell growth in a wide range of cancers including thyroid carcinomas. Normally such pathway is deactivated by tumor suppressor phosphatase. It demonstrated the importance of dual role of mTOR and AKT and analyze that over expression of mTOR/AKT signaling pathway enhance the chances of tumor growth. They also identified some genetic alterations in isoform of PI3K (Nozhat and Hedayati, 2016). There were some other pathways that were reported like PTEN/PDK1/BRAF, Ras/Raf/MAPK signaling pathways (Zaballos and Santisteban, 2017).

Genes associated with reported pathways are deferentially expressed during our Microarray and RNA-seq analysis. In this study we focused on some entities or pathways. Rest of the pathways will be our major concerns of future. We will design a pathway and further analyze it by using quantitative modeling approach of systems biology.

Chapter 6

CONCLUSION AND FUTURE PERSPECTIVES

This study has proposed the therapeutic targets by using high throughput sequencing techniques. Microarray and RNA-sequencing based analysis are performed by using different datasets of mRNAs and miRNAs. Differentially expressed mRNAs and miRNAs has been identified. Targets of miRNAs are also analyzed. Motive of the study is to perform differential expression analysis to obtain differentially expressed genes (DEGs), comparative study of DEGs through pathway analysis. Comparative analysis of microarray datasets revealed 280 common genes and 1247 common genes in RNA-seq datasets. Differential expression analysis shows the upregulated and downregulated genes which is further used to perform pathway analysis. The genes are declared as significant because 'Apoptosis' and 'Proliferation' are found to be sensitive towards those genes. Protoglycans in cancer is appeared as common pathway in all the datasets of microarray and RNA-seq used in this study. To analyse complex mechanisms, expression and interactions, further investigation through wet lab techniques is necessary. To minimize the susceptibility of cancer, the differential level and alterations in those significant genes should be observed in thyroid carcinoma patients. Further study on remaining pathways by using different approaches of systems biology and wet lab techniques can provides us more important targets associated with the disease.

REFERENCES

Abdullah, M. I., Junit, S. M., Ng, K. L., Jayapalan, J. J., Karikalan, B. and Hashim, O. H. (2019). Papillary thyroid cancer: genetic alterations and molecular biomarker investigations. International journal of medical sciences *16*, 450.

Andrews, S. et al. (2010). FastQC: a quality control tool for high throughput sequence data.

Ao, Z.-X., Chen, Y.-C., Lu, J.-M., Shen, J., Peng, L.-P., Lin, X., Peng, C., Zeng, C.-P., Wang, X.-F., Zhou, R. et al. (2018). Identification of potential functional genes in papillary thyroid cancer by co-expression network analysis. Oncology Letters *16*, 4871–4878.

Baciu, A., Uyy, E., Suica, V., Boteanu, R., Popescu, A., Giulea, C., Manda, D., Badiu, C. and Antohe, F. (2017). Proteomic analysis of plasma molecular markers as predictors of differentiated thyroid cancer. Romanian Reports in Physics *69*, 601.

Blankenberg, D., Kuster, G. V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A. and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. Current protocols in molecular biology *89*, 19–10.

Blighe, K., Rana, S. and Lewis, M. (2019). EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. R package version *1*.

Cabanillas, M. E., McFadden, D. G. and Durante, C. (2016). Thyroid cancer. The Lancet *388*, 2783–2795.

Carlomagno, N., Incollingo, P., Tammaro, V., Peluso, G., Rupealta, N., Chiacchio, G., Sandoval Sotelo, M. L., Minieri, G., Pisani, A., Riccio, E. et al. (2017). Diagnostic, predictive, prognostic, and therapeutic molecular biomarkers in third millennium: a breakthrough in gastric cancer. BioMed research international *2017*.

Carvalho, B. and Scharpf, R. (2011). oligoClasses: Classes for High-Throughput Arrays Supported by oligo and crlmm. R package version *1*.

Carvalho, B. S. and Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. Bioinformatics *26*, 2363–2367.

Chang, F., Xing, P., Song, F., Du, X., Wang, G., Chen, K. and Yang, J. (2016). The role of T-box genes in the tumorigenesis and progression of cancer. Oncology letters *12*, 4305–4311.

Cheng, S.-P., Liu, C.-L., Chen, M.-J., Chien, M.-N., Leung, C.-H., Lin, C.-H., Hsu, Y.-C. and Lee, J.-J. (2015). CD74 expression and its therapeutic potential in thyroid carcinoma. Endocr Relat Cancer *22*, 179–190.

Chmielik, E., Rusinek, D., Oczko-Wojciechowska, M., Jarzab, M., Krajewska, J., Czarniecka, A. and Jarzab, B. (2018). Heterogeneity of thyroid cancer. Pathobiology *85*, 117–129.

Chou, C.-K., Liu, R.-T. and Kang, H.-Y. (2017). MicroRNA-146b: a novel biomarker and therapeutic target for human papillary thyroid cancer. International journal of molecular sciences *18*, 636.

Clancy, S. and Brown, W. (2008). Translation: DNA to mRNA to protein. Nature Education *1*, 101.

Cohen Freue, G. V., Hollander, Z., Shen, E., Zamar, R. H., Balshaw, R., Scherer, A., McManus, B., Keown, P., McMaster, W. R. and Ng, R. T. (2007). MDQC: a new quality assessment method for microarrays based on quality control reports. Bioinformatics *23*, 3162–3169.

Costa-Silva, J., Domingues, D. and Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. PloS one *12*, e0190152.

Dal Maso, L., Bosetti, C., La Vecchia, C. and Franceschi, S. (2009). Risk factors for thyroid cancer: an epidemiological review focused on nutritional factors. Cancer Causes & Control *20*, 75–86.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. and Lempicki, R. A. (2003). DAVID: database for annotation, visualization, and integrated discovery. Genome biology *4*, 1–11.

Dweep, H., Gretz, N. and Sticht, C. (2014). miRWalk database for miRNA–target interactions. In RNA Mapping pp. 289–305. Springer.

Fisher, B. F., Snodgrass, H. M., Jones, K. A., Andorfer, M. C. and Lewis, J. C. (2019). Site-Selective C–H Halogenation Using Flavin-Dependent Halogenases Identified via Family-Wide Activity Profiling. ACS central science *5*, 1844–1856.

Gandolfo, L. C. and Speed, T. P. (2018). RLE plots: Visualizing unwanted variation in high dimensional data. PLoS One *13*, e0191629.

Gao, X., Wang, J. and Zhang, S. (2019). Integrated bioinformatics analysis of hub genes and pathways in anaplastic thyroid carcinomas. International journal of endocrinology *2019*.

García-Campos, M. A., Espinal-Enríquez, J. and Hernández-Lemus, E. (2015). Pathway analysis: state of the art. Frontiers in physiology *6*, 383.

Gentleman, R. (2018). Biocore. geneplotter: Graphics related functions for Bioconductor. R package version *1*.

Giorgi, F. M., Bolger, A. M., Lohse, M. and Usadel, B. (2010). Algorithm-driven artifacts in median polish summarization of microarray data. BMC bioinformatics *11*, 553.

Grande, E., Díez, J. J., Zafon, C. and Capdevila, J. (2012). Thyroid cancer: molecular aspects and new therapeutic strategies. Journal of Thyroid Research *2012*.

Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C. and Shyr, Y. (2017). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. Genomics *109*, 83–90.

Han, Y., Gao, S., Muegge, K., Zhang, W. and Zhou, B. (2015). Advanced applications of RNA sequencing and challenges. Bioinformatics and biology insights *9*, BBI–S28991.

Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Hughes, J., Snesrud, E., Lee, N. and Quackenbush, J. (2000). A concise guide to cDNA microarray analysis. Biotechniques *29*, 548–562.

Hicks, S. C. and Irizarry, R. A. (2015). Quantro: a data-driven approach to guide the choice of an appropriate normalization method. Genome biology *16*, 117.

Hossain, M., Asa, T. A., Rahman, M., Uddin, S., Moustafa, A. A., Quinn, J. M., Moni, M. A. et al. (2020). Network-Based Genetic Profiling Reveals Cellular Pathway Differences Between Follicular Thyroid Carcinoma and Follicular Thyroid Adenoma. International Journal of Environmental Research and Public Health *17*, 1373.

Hosseini, S. M., Okoye, I., Chaleshtari, M. G., Hazhirkarzar, B., Mohamadnejad, J., Azizi, G., Hojjat-Farsangi, M., Mohammadi, H., Shotorbani, S. S. and Jadidi-Niaragh, F. (2019). E2 ubiquitin-conjugating enzymes in cancer: Implications for immunotherapeutic interventions. Clinica Chimica Acta *498*, 126–134.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T. et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nature methods *12*, 115–121.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249–264.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic acids research *45*, D353–D361.

Kauffmann, A., Gentleman, R. and Huber, W. (2009). arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. Bioinformatics *25*, 415–416.

Kim, D., Langmead, B. and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nature methods *12*, 357–360.

Klaus, B. and Reisenauer, S. (2016). An end to end workflow for differential gene expression using Affymetrix microarrays. F1000Research *5*.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A. et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research *44*, W90–W97.

Leboulleux, S., Baudin, E., Travagli, J.-P. and Schlumberger, M. (2004). Medullary thyroid carcinoma. Clinical endocrinology *61*, 299–310.

Leggett, R. M., Ramirez-Gonzalez, R. H., Clavijo, B., Waite, D. and Davey, R. P. (2013). Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Frontiers in genetics *4*, 288.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, Q., Li, H., Zhang, L., Zhang, C., Yan, W. and Wang, C. (2017). Identification of novel long non-coding RNA biomarkers for prognosis prediction of papillary thyroid cancer. Oncotarget *8*, 46136.

Li, X., He, J., Zhou, M., Cao, Y., Jin, Y. and Zou, Q. (2019). Identification and Validation of Core Genes Involved in the Development of Papillary Thyroid Carcinoma via Bioinformatics Analysis. International journal of genomics *2019*.

Lin, P., Yao, Z., Sun, Y., Li, W., Liu, Y., Liang, K., Liu, Y., Qin, J., Hou, X. and Chen, L. (2019). Deciphering novel biomarkers of lymph node metastasis of thyroid papillary microcarcinoma using proteomic analysis of ultrasound-guided fine-needle aspiration biopsy samples. Journal of proteomics *204*, 103414.

Liu, X., He, M., Hou, Y., Liang, B., Zhao, L., Ma, S., Yu, Y. and Liu, X. (2013). Expression profiles of microRNAs and their target genes in papillary thyroid carcinoma. Oncology reports *29*, 1415–1420.

Naoum, G. E., Morkos, M., Kim, B. and Arafat, W. (2018). Novel targeted therapies and immunotherapy for advanced thyroid cancers. Molecular cancer *17*, 51.

Nguyen, Q. T., Lee, E. J., Huang, M. G., Park, Y. I., Khullar, A. and Plodkowski, R. A. (2015). Diagnosis and treatment of patients with thyroid cancer. American health & drug benefits *8*, 30.

Nikiforov, Y. E. (2008). Thyroid carcinoma: molecular pathways and therapeutic targets. Modern Pathology *21*, S37–S43.

Nikitovic, D., Berdiaki, A., Spyridaki, I., Krasanakis, T., Tsatsakis, A. and Tzanakakis, G. N. (2018). Proteoglycans—biomarkers and targets in cancer therapy. Frontiers in endocrinology *9*, 69.

Nozhat, Z. and Hedayati, M. (2016). PI3K/AKT pathway and its mediators in thyroid carcinomas. Molecular diagnosis & therapy *20*, 13–26.

Oldenhuis, C., Oosting, S., Gietema, J. and De Vries, E. (2008). Prognostic versus predictive value of biomarkers in oncology. European Journal of Cancer *44*, 946–953.

O'Neill, J. P. and Shaha, A. R. (2013). Anaplastic thyroid cancer. Oral oncology *49*, 702–706.

Pan, Z., Li, L., Fang, Q., Qian, Y., Zhang, Y., Zhu, J., Ge, M. and Huang, P. (2019). Integrated bioinformatics analysis of master regulators in anaplastic thyroid carcinoma. BioMed research international *2019*.

Peng, Y. and Croce, C. M. (2016). The role of MicroRNAs in human cancer. Signal transduction and targeted therapy *1*, 1–9.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T. and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology *33*, 290–295.

Pinero, J., Ramrez-Anguita, J. M., Sauch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F. and Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic acids research *48*, D845–D855.

Reuter, J. A., Spacek, D. V. and Snyder, M. P. (2015). High-throughput sequencing technologies. Molecular cell *58*, 586–597.

Ringnér, M. (2008). What is principal component analysis? Nature biotechnology *26*, 303–304.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research *43*, e47–e47.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M. et al. (2015). The

UCSC genome browser database: 2015 update. Nucleic acids research *43*, D670–D681.

Sada, H., Hinoi, T., Ueno, H., Yamaguchi, T., Inoue, Y., Konishi, T., Kobayashi, H., Kanemitsu, Y., Ishida, F., Ishida, H. et al. (2019). Prevalence of and risk factors for thyroid carcinoma in patients with familial adenomatous polyposis: results of a multicenter study in Japan and a systematic review. Surgery Today *49*, 72–81.

Saiselet, M., Pita, J. M., Augenlicht, A., Dom, G., Tarabichi, M., Fimereli, D., Dumont, J. E., Detours, V. and Maenhaut, C. (2016). miRNA expression and function in thyroid carcinomas: a comparative and critical analysis and a model for other cancers. Oncotarget *7*, 52475.

Smallridge, R. and Copland, J. (2010). Anaplastic thyroid carcinoma: pathogenesis and emerging therapies. Clinical Oncology *22*, 486–497.

Soon, W. W., Hariharan, M. and Snyder, M. P. (2013). High-throughput sequencing for biology and medicine. Molecular systems biology *9*, 640.

Tavares, C., Melo, M., Cameselle-Teijeiro, J. M., Soares, P. and Sobrinho-Simoes, M. (2016). Genetic predictors of thyroid cancer outcome. Eur J Endocrinol *174*, 117–126.

Team, R. C. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Austria: Vienna.

Thirumalai, C., Vignesh, M. and Balaji, R. (2017). Data analysis using Box and Whisker plot for Lung Cancer. In 2017 Innovations in Power and Advanced Computing Technologies (i-PACT) pp. 1–6, IEEE.

Tomari, Y. and Zamore, P. D. (2005). MicroRNA biogenesis: drosha can't cut it without a partner. Current Biology *15*, R61–R64.

Wang, L., Wang, S. and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics *28*, 2184–2185.

Wang, Q., Shen, Y., Ye, B., Hu, H., Fan, C., Wang, T., Zheng, Y., Lv, J., Ma, Y. and Xiang, M. (2018). Gene expression differences between thyroid carcinoma, thyroid adenoma and normal thyroid tissue. Oncology reports *40*, 3359–3369.

Werner, T. A., Dizdar, L., Nolten, I., Riemer, J. C., Mersch, S., Schütte, S. C., Driemel, C., Verde, P. E., Raba, K., Topp, S. A. et al. (2017). Survivin and XIAP–two potential biological targets in follicular thyroid carcinoma. Scientific reports *7*, 1–11.

Wickham, H. (2016). ggplot2: elegant graphics for data analysis. springer.

Xia, J. and Wishart, D. S. (2010). MetPA: a web-based metabolomics tool for pathway analysis and visualization. Bioinformatics *26*, 2342–2344.

Xing, M. (2013). Molecular pathogenesis and mechanisms of thyroid cancer. Nature Reviews Cancer *13*, 184–199.

Yoo, S.-K., Song, Y. S., Lee, E. K., Hwang, J., Kim, H. H., Jung, G., Kim, Y. A., Kim, S.-j., Cho, S. W., Won, J.-K. et al. (2019). Integrative analysis of genomic and transcriptomic characteristics associated with progression of aggressive thyroid cancer. Nature communications *10*, 1–12.

Yu, G., Wang, L.-G., Han, Y. and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a journal of integrative biology *16*, 284–287.

Zaballos, M. A. and Santisteban, P. (2017). Key signaling pathways in thyroid cancer. Journal of Endocrinology *235*, R43–R61.

Appendix A

Source code of microarray analysis for Affymetrix

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

# The following initializes usage of Bioc devel
BiocManager::install(version='3.9')
BiocManager::install("maEndToEnd")

browseVignettes("maEndToEnd")
suppressPackageStartupMessages({library("maEndToEnd")})

install.packages("devtools")
library(devtools)

devtools::install_github("r-lib/remotes")
library(devtools)
library(remotes)
packageVersion("remotes") # has to be 1.1.1.9000 or later

BiocManager::install("pd.hugene.1.0.st.v1")
BiocManager::install("GenomeInfoDbData")
BiocManager::install("Cairo")
BiocManager::install("arrayQualityMetrics")

remotes::install_github("b-klaus/maEndToEnd", ref="master")



#General Bioconductor packages
library(Biobase)
library(oligoClasses)

#Annotation and data import packages
library(ArrayExpress)
library(pd.hugene.1.0.st.v1)
library(hugene10sttranscriptcluster.db)
library(hgu133plus2.db)
library(pd.hg.u133.plus.2)

#Quality control and pre-processing packages
library(oligo)
library(arrayQualityMetrics)

#Analysis and statistics packages
```

```
library(limma)
library(topGO)
library(ReactomePA)
library(clusterProfiler)

#Plotting and color options packages
library(gplots)
library(ggplot2)
library(geneplotter)
library(RColorBrewer)
library(pheatmap)

#Formatting/documentation packages
#library(rmarkdown)
#library(BiocStyle)
library(dplyr)
library(tidyr)

#Helpers:
library(stringr)
library(matrixStats)
library(genefilter)
library(openxlsx)
#library(devtools)
library(maEndToEnd)


getwd()
setwd('D:/aqsadataset1')
raw_data_dir <- getwd()

#raw_data_dir <- tempdir()

if (!dir.exists(raw_data_dir)) {
  dir.create(raw_data_dir)
}
#anno_AE <- getAE("E-GEOD-65144", path = raw_data_dir, type = "raw")

raw_data_dir <- ("D:/aqsadataset1/data1")

if (!dir.exists(raw_data_dir)) {
  dir.create(raw_data_dir)
}
sdrf_location <- file.path(raw_data_dir, "E-GEOD-65144.sdrf.txt")
SDRF <- read.delim(sdrf_location)
write.csv(SDRF,"sdrf.csv")
rownames(SDRF) <- SDRF$Array.Data.File
SDRF <- AnnotatedDataFrame(SDRF)
```

```
getwd()
raw_data <- getwd()

raw_data <- oligo::read.celfiles(filenames = file.path(raw_data_dir,
                                                       SDRF$Array.Data.File),

                                 verbose = FALSE, phenoData = SDRF)
stopifnot(validObject(raw_data))


#head(Biobase::pData(raw_data))
Biobase::pData(raw_data)


Biobase::pData(raw_data) <- Biobase::pData(raw_data)[, c("Source.Name",
                                                         "Characteristics..
                                                            organism.",
                                                         "FactorValue..tissue
                                                            .type.")]



Biobase::exprs(raw_data)[1:5, 1:5]
exp_raw <- log2(Biobase::exprs(raw_data))
PCA_raw <- prcomp(t(exp_raw), scale. = FALSE)


percentVar <- round(100*PCA_raw$sdev^2/sum(PCA_raw$sdev^2),1)
sd_ratio <- sqrt(percentVar[2] / percentVar[1])

dataGG <- data.frame(PC1 = PCA_raw$x[,1], PC2 = PCA_raw$x[,2],
                     Phenotype = pData(raw_data)$FactorValue..tissue.type.

                     )

write.csv(dataGG,"dataGG.csv")

ggplot(dataGG, aes(PC1, PC2)) +
  geom_point(aes(colour = Phenotype)) +
  ggtitle("PCA plot of log-transformed raw expression data") +
  xlab(paste0("PC1, VarExp: ", percentVar[1], "%")) +
  ylab(paste0("PC2, VarExp: ", percentVar[2], "%")) +
  theme(plot.title = element_text(hjust = 0.5))+
  coord_fixed(ratio = sd_ratio) +
  scale_shape_manual(values = c(4,15)) +
  scale_color_manual(values = c("darkorange2", "dodgerblue4"))
```

```
oligo::boxplot(raw_data ,las=2, cex.axis= 0.35,

             main = "Boxplot of log2-intensitites for raw data")

arrayQualityMetrics(expressionset = raw_data,
                  outdir = tempdir(),
                  force = TRUE, do.logtransform = TRUE,
                  intgroup = c("FactorValue..tissue.type."))

head(ls("package:hgu133plus2.db"))

#ls("pd.hg.u133.plus.2")
#head(ls("package:pd.hg.u133.plus.2"))

palmieri_eset <- oligo::rma(raw_data, normalize = FALSE)

warnings()

?rma

row_medians_assayData <-
  Biobase::rowMedians(as.matrix(Biobase::exprs(palmieri_eset)))
RLE_data <- sweep(Biobase::exprs(palmieri_eset), 1,
    row_medians_assayData)
RLE_data <- as.data.frame(RLE_data)
write.csv(RLE_data,"RLE_data.csv")
RLE_data_gathered <-
  tidyr::gather(RLE_data, patient_array, log2_expression_deviation)
ggplot2::ggplot(RLE_data_gathered, aes(patient_array,
                                       log2_expression_deviation)) +
  geom_boxplot(outlier.shape = NA) +
  ylim(c(-2, 2)) +
  theme(axis.text.x = element_text(colour = "aquamarine4",
                                   angle = 90, size = 6.5, hjust = 1 ,
                                   face = "bold"))
dev.off()

RLE_data_gathered
#RMA calibration of the data
palmieri_eset_norm <- oligo::rma(raw_data)

palmieri_eset_norm
#RLE after normalization
row_medians_assayData <-
  Biobase::rowMedians(as.matrix(Biobase::exprs(palmieri_eset_norm)))

RLE_data <- sweep(Biobase::exprs(palmieri_eset_norm), 1,
    row_medians_assayData)
```

```
RLE_data <- as.data.frame(RLE_data)
RLE_data_gathered <-
  tidyr::gather(RLE_data, patient_array, log2_expression_deviation)
ggplot2::ggplot(RLE_data_gathered, aes(patient_array,
                                       log2_expression_deviation)) +
  geom_boxplot(outlier.shape = NA) +
  ylim(c(-2, 2)) +
  theme(axis.text.x = element_text(colour = "aquamarine4",
                                   angle = 90, size = 6.5, hjust = 1 ,
                                   face = "bold"))
dev.off()

write.csv(RLE_data_gathered,"RLE_data_gathered.csv")
palmieri_eset_norm <- oligo::rma(raw_data)

exp_palmieri <- Biobase::exprs(palmieri_eset_norm)
write.csv(exp_palmieri,"exp_palmieri.csv")
PCA <- prcomp(t(exp_palmieri), scale = FALSE)
percentVar <- round(100*PCA$sdev^2/sum(PCA$sdev^2),1)
sd_ratio <- sqrt(percentVar[2] / percentVar[1])
dataGG <- data.frame(PC1 = PCA$x[,1], PC2 = PCA$x[,2],
                     Phenotype = pData(raw_data)$FactorValue..tissue.type
                         .)

ggplot(dataGG, aes(PC1, PC2)) +
  geom_point(aes(colour = Phenotype)) +
  ggtitle("PCA plot of calibrated, summarized data") +
  xlab(paste0("PC1, VarExp: ", percentVar[1], "%")) +
  ylab(paste0("PC2, VarExp: ", percentVar[2], "%")) +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_fixed(ratio = sd_ratio) +
  scale_shape_manual(values = c(4,15)) +
  scale_color_manual(values = c("darkorange2", "dodgerblue4"))


#boxplot

oligo::boxplot(palmieri_eset_norm,las=2, cex.axis= 0.4,

               main = "Boxplot of log2-intensitites for the normalized
                   data")



#heatmaps
phenotype_names <- ifelse(str_detect(pData
                                     (palmieri_eset_norm)$FactorValue..
                                        tissue.type.,
                                     "nor"), "normal.", "ATC.")
```

```
annotation_for_heatmap <-
  data.frame(Phenotype = phenotype_names)
row.names(annotation_for_heatmap) <- row.names(pData(palmieri_eset_norm
    ))
write.csv(annotation_for_heatmap,"annotation_for_heatmap.csv")
dists <- as.matrix(dist(t(exp_palmieri), method = "manhattan"))
write.csv(dists,"dists.csv")
rownames(dists) <- row.names(pData(palmieri_eset_norm))
hmcol <- rev(colorRampPalette(RColorBrewer::brewer.pal(9, "YlOrRd"))
    (255))
colnames(dists) <- NULL
diag(dists) <- NA
ann_colors <- list(
  Phenotype = c(normal. = "chartreuse4", ATC. = "burlywood3"))
pheatmap(dists, col = (hmcol),
        annotation_row = annotation_for_heatmap,
        annotation_colors = ann_colors,
        legend = TRUE,
        treeheight_row = 0,
        legend_breaks = c(min(dists, na.rm = TRUE),
                          max(dists, na.rm = TRUE)),
        legend_labels = (c("small distance", "large distance")),
        main = "Clustering Heatmap for the calibrated samples")


#Median intensities
palmieri_medians <- rowMedians(Biobase::exprs(palmieri_eset_norm))
hist_res <- hist(palmieri_medians, 100, col = "cornsilk1", freq = FALSE
    ,
                main = "Histogram of the median intensities",
                border = "antiquewhite4",
                xlab = "Median intensities")


#with threshold

man_threshold <- 2.3
hist_res <- hist(palmieri_medians, 100, col = "cornsilk", freq = FALSE,
                main = "Histogram of the median intensities",
                border = "antiquewhite4",
                xlab = "Median intensities")
abline(v = man_threshold, col = "coral4", lwd = 2)

#Transcripts that do not have intensities larger than the threshold in
    at least as many arrays as the smallest
#experimental group are excluded.

  no_of_samples <-
  table(paste0(pData(palmieri_eset_norm)$FactorValue..tissue.type.))
```

```
  no_of_samples

  samples_cutoff <- min(no_of_samples)
  idx_man_threshold <- apply(Biobase::exprs(palmieri_eset_norm), 1,
                             function(x){
                               sum(x > man_threshold) >= samples_cutoff})
  table(idx_man_threshold)

  palmieri_manfiltered <- subset(palmieri_eset_norm, idx_man_threshold)


#hugene10sttranscriptcluster.db
anno_palmieri <- AnnotationDbi::select(hgu133plus2.db,
                                       keys = (featureNames(
                                           palmieri_manfiltered)),
                                       columns = c("SYMBOL", "GENENAME"),
                                       keytype = "PROBEID")
anno_palmieri <- subset(anno_palmieri, !is.na(SYMBOL))


#Removing multiple mappings

  anno_grouped <- group_by(anno_palmieri, PROBEID)
  anno_summarized <-
    dplyr::summarize(anno_grouped, no_of_matches = n_distinct(SYMBOL))
  head(anno_summarized)

  anno_summarized

  anno_filtered <- filter(anno_summarized, no_of_matches > 1)
  head(anno_filtered)
  probe_stats <- anno_filtered
  nrow(probe_stats)
  write.csv(probe_stats,"probe_stats.csv")
  ids_to_exlude <- (featureNames(palmieri_manfiltered) %in%
      probe_stats$PROBEID)
  table(ids_to_exlude)

  palmieri_final <- subset(palmieri_manfiltered, !ids_to_exlude)
  validObject(palmieri_final)

  head(anno_palmieri)
  fData(palmieri_final)$PROBEID <- rownames(fData(palmieri_final))
  fData(palmieri_final) <- left_join(fData(palmieri_final),
      anno_palmieri)

  # restore rownames after left_join
```

```
rownames(fData(palmieri_final)) <- fData(palmieri_final)$PROBEID
validObject(palmieri_final)


tissue <- str_replace_all(Biobase::pData(palmieri_final)$FactorValue..
    tissue.type.,
                          " ", "_")
tissue <- ifelse(tissue == "anaplastic_thyroid_carcinoma_(ATC)",
                 "ATC","normal")

##########
# Matrix
design_matrix <- model.matrix(~ 0 + tissue)
write.csv(design_matrix,"design matrix.csv")
contrast_matrix_C <- makeContrasts( tissueATC-tissuenormal, levels =
    design_matrix)
contrast_matrix_C
write.csv(contrast_matrix_C,"contrast_matrix_C.csv")
fit <- lmFit(palmieri_final, design_matrix)
                                          # contrast_matrix_C)
fit2 <- contrasts.fit(fit, contrast_matrix_C)
fit2 <- eBayes(fit2)
table_T <- topTable(fit2, number = Inf)
write.csv(table_T,"DEGs.csv" )
BiocManager::install("EnhancedVolcano")
library(EnhancedVolcano)
EnhancedVolcano(table_T,
                lab = table_T $SYMBOL,
                x = "logFC",
                y = "P.Value",
                pCutoff = 0.05,
                FCcutoff =0.5,
                title = "ATC vs Normal")
t1 <- subset(table_T, P.Value < 0.05)
t2 <- subset(t1, logFC < -0.5 | logFC> 0.5)
write.csv(t2,"final DEGs with cutoff.csv")
```

Appendix B

Source Code for Ballgown-RNA-seq

```
BiocManager::install("ballgown")
library(ballgown)

library(genefilter)
library(plyr)
library(devtools)


########################
library(ballgown)
getwd()

setwd("D:/rnaseq")
dwe<- "D:/rnaseq"
pheno = read.csv("PHENO.csv")

bg = ballgown(dataDir= dwe, samplePattern='sample', meas='all', pData=
    pheno)


save(bg, file='bg.rda')
structure(bg)$exon
bg

bg_filt = subset(bg,"rowVars(texpr(bg)) >1",genomesubset=TRUE)
bg_table = texpr(bg_filt, 'all')
bg_gene_names = unique(bg_table[, 9:10])
transcript_expression = as.data.frame(texpr(bg_filt))
head(transcript_expression)
row.names(transcript_expression)


############################### files ##############
write.csv(transcript_expression, "transcripts.csv")
write.csv(bg_gene_names, "bg_gene_names.csv")
write.csv(bg_table, "bg_table.csv")
write.csv(results_genes, "results_genes.csv")
write.csv(results_transcripts, "results_transcripts.csv")

##################transcripts analysis##############
results_transcripts = stattest(bg_filt,
                               feature="transcript",covariate="phenotype",
                                   adjustvars = NULL,
                               getFC=TRUE, meas="FPKM")
```

```
results_genes = stattest(bg_filt, feature="gene",
                      covariate="phenotype", adjustvars = NULL, getFC=
                          TRUE,
                      meas="FPKM")

results_transcriptsmer =
  data.frame(geneNames=ballgown::geneNames(bg_filt),
             geneIDs=ballgown::geneIDs(bg_filt), transcriptNames=
                 ballgown::transcriptNames(bg_filt),
             results_transcripts)

results_transcripts = arrange(results_transcripts,pval)
results_genes = arrange(results_genes,pval)

write.csv(results_transcripts, "chrX_transcript_results.csv",
          row.names=FALSE)
write.csv(results_genes, "chrX_gene_results.csv",
          row.names=FALSE)
tra <- subset(results_transcripts,results_transcripts$pval<0.05)
gen <- subset(results_genes,results_genes$pval<0.05)

write.csv(tra, "filtered transcripts.csv")
write.csv(gen, "filtered genes.csv")


##############################PCA-plot###############

pca_data=prcomp(t(transcript_expression))
percentVar <- round(100*pca_data$sdev^2/sum(pca_data$sdev^2),1)
percentVar
sd_ratio <- sqrt(percentVar[2] / percentVar[1])
sd_ratio
sd_ratio = 1.5
dataGG <- data.frame(PC1 = pca_data$x[,1], PC2 =pca_data$x[,2],
                     Phenotype = pheno$phenotype)
dataGG
ggplot(dataGG, aes(PC1, PC2)) +
  geom_point(aes(colour = Phenotype)) +
  ggtitle("PCA plot") +
  xlab(paste0("PC1, VarExp: ", percentVar[1], "%")) +
  ylab(paste0("PC2, VarExp: ", percentVar[2], "%")) +
  theme(plot.title = element_text(hjust = 0.5))+
  coord_fixed(ratio = sd_ratio) +
  scale_shape_manual(values = c(4,15)) +
  scale_color_manual(values = c("darkorange2", "dodgerblue4"))


#################### Box plot ######################
```

```
tropical= c('darkorange', 'dodgerblue',
          'hotpink', 'limegreen', 'yellow')
palette(tropical)

fpkm = texpr(bg_filt,meas="FPKM")
fpkm = log2(fpkm+1)
boxplot (fpkm,col=as.numeric(pheno$phenotype),las=2,ylab='log2(FPKM+1)
    ')


################################ABLINE plots##############

transcript_gene_table = indexes(bg_filt)$t2g
head(transcript_gene_table)
#Each row of data represents a transcript. Many of these transcripts
    represent the same gene.
#Determine the numbers of transcripts and unique genes#
length(row.names(transcript_gene_table))
length(unique(transcript_gene_table[,"g_id"]))

counts=table(transcript_gene_table[,"g_id"])
write.csv(counts, "COUNT table.csv")

c_one = length(which(counts == 1))
c_more_than_one = length(which(counts > 1))
c_max = max(counts)
hist(counts, breaks=50, col="bisque4", xlab="Transcripts per gene",
    main="Distribution of transcript count per gene")
legend_text = c(paste("Genes with one transcript =", c_one), paste("
    Genes with more than one transcript =", c_more_than_one), paste("
    Max transcripts for single gene = ", c_max))
legend("topright", legend_text, lty=NULL)

#Plot #2 - the distribution of transcript sizes as a histogram
full_table <- texpr(bg_filt , 'all')
hist(full_table$length, breaks=50, xlab="Transcript length (bp)", main
    ="Distribution of transcript lengths", col="steelblue")
################################################################

data_colors=(c("white", "blue", "#007FFF", "cyan","#7FFF7F", "yellow",
    "#FF7F00", "red", "#7F0000",
              "white", "blue","#007FFF", "cyan","#7FFF7F", "yellow", "#
                 FF7F00", "red", "#7F0000",
              "yellow", "#FF7F00", "red", "#7F0000","green","cyan"))

min_nonzero=1
#Set the columns for finding FPKM and create shorter names for figures
data_columns=c(1:22)
```

```
short_names=c("S1","S2","S3","S4","S5","S6","S7","S8", "S9","S10","S11
    ","S12","S13","S14","S15","S16","S17","S18","S19","S20","S21","S22")

#Plot #3 - View the range of values and general distribution of FPKM
    values for all libraries Create boxplots for this purpose
#Display on a log2 scale and add the minimum non-zero value to avoid
    log2(0)
boxplot(log2(transcript_expression [,data_columns]+min_nonzero), col=
    data_colors, names=short_names, las=2, ylab="log2(FPKM)", main="
    Distribution of FPKMs ")

################################################
colors = colorRampPalette(c("blue", "blue", "#007FFF", "cyan","#7FFF7F
    ", "yellow", "#FF7F00", "red", "#7F0000"))
#smoothScatter(x=log2(x+min_nonzero), xlab="FPKM (SRR218_N, Replicate
    1)", ylab="FPKM (SRR219_N, Replicate 2)", main="Comparison of
    expression values for a pair of replicates", colramp=colors, nbin
    =200)
#Compare the correlation 'distance' between all replicates
transcript_expression[,"sum"]=apply(transcript_expression[,data_columns
    ], 1, sum)
#Identify the genes with a grand sum FPKM of at least 5 - we will
    filter out the genes with very low expression across the board
i = which(transcript_expression[,"sum"] > 5)
#Calculate the correlation between all pairs of data
r=cor(transcript_expression[i,data_columns], use="pairwise.complete.obs
    ", method="pearson")
r
#Plot #8 - Convert correlation to 'distance', and use 'multi-
    dimensional scaling' to display the relative differences between
    libraries
d=1-r
data_columns=c(1:22)
write.csv(data_columns, "DAta columns.csv")
mds=cmdscale(d, k=2, eig=TRUE)
par(mfrow=c(1,1))
plot(mds$points, type="n", xlab="", ylab="", main="MDS distance plot ",
     xlim=c(-0.25,0.25), ylim=c(-0.25,0.25))
points(mds$points[,1], mds$points[,2], col="grey", cex=2, pch=16)
text(mds$points[,1], mds$points[,2], short_names, col=data_colors)

########################################
sig=which(results_transcriptsmer$pval<0.05)
results_transcriptsmer
sig
results_transcriptsmer[,"de"] = log2(results_transcriptsmer[,"fc"])

write.csv(results_transcriptsmer,"file de.csv")
```

```
hist(results_transcriptsmer[sig,"de"], breaks=50, col="seagreen", xlab
    ="log2(Fold change Normal-Diseased)",
     main="Distribution of differential expression values")
abline(v=-2, col="black", lwd=2, lty=2)
abline(v=2, col="black", lwd=2, lty=2)
legend("topleft", "Fold-change > 4", lwd=2, lty=2)

################################## enhanced volcano
    ##########################

library(EnhancedVolcano)

library(ggrepel)
EnhancedVolcano(results_transcriptsmer,
                lab = results_transcriptsmer$de,
                x = "de",
                y = "pval",
                pCutoff = 0.05,
                FCcutoff = 0.5,
                title = "RNAseq-dataset")
trans <- subset(results_transcriptsmer, pval < 0.05)
trans <- subset(trans, de < -0.5 | de > 0.5)
write.csv(trans,"DEGs with cutoffvolcano (pval vs de).csv")

############################## HEAT_MAP###############
library(stringr)
library(pheatmap)
pheno$phenotype

disease_names <- ifelse(str_detect(pheno$phenotype,
                                   "diseased"), "Diseased", "Normal")

annotation_for_heatmap <- data.frame(Phenotype = disease_names)
annotation_for_heatmap
row.names(annotation_for_heatmap) <- pheno$sample

##
dists <- as.matrix(dist(t(transcript_expression), method = "manhattan")
    )

dists
rownames(dists) <- pheno$sample #sample names
rownames(dists)
hmcol <- rev(colorRampPalette(RColorBrewer::brewer.pal(9, "YlOrRd"))
    (255))
colnames(dists) <- NULL
diag(dists) <- NA
```

```
ann_colors <- list(Disease= c(Diseased= "blue4", Normal = "cadetblue2")
    )
pdf("Heatmap.pdf",width=10,height=7,paper='special')
pheatmap(dists, col = (hmcol),
         annotation_row = annotation_for_heatmap,
         annotation_colors = ann_colors,
         legend = TRUE,
         treeheight_row = 0,
         legend_breaks = c(min(dists, na.rm = TRUE),
                           max(dists, na.rm = TRUE)),
         legend_labels = (c("small distance", "large distance")),
         main = "Clustering heatmap for the calibrated samples")
dev.off()
```

Appendix C

Source code of microarray analysis for Agilent platform

```
#agilent single channel

#libraries

library(limma)
library(convert)
library(Biobase)
library(openxlsx)
library(ggplot2)
library(oligo)
library(oligoClasses)
library(arrayQualityMetrics)
library(simpleaffy)
library(RColorBrewer)
library(pheatmap)
library(geneplotter)
library(stringr)
library(ArrayExpress)
library(gplots)
library(dplyr)
library(tidyr)
library(matrixStats)
library(genefilter)
#BiocManager::install("simpleaffy")

# Directory Setup
Tutorial_agilent <- "D:/dataset3"

# If Directory exist then good otherwise make one.

Tutorial_agilent
if(!dir.exists(Tutorial_agilent)){
  dir.create(Tutorial_agilent)
}

#set working directory

setwd("D:/dataset3")
raw_data_dir <- file.path(getwd(), "rawDataMAWorkdown")

if(!dir.exists(raw_data_dir)){
  dir.create(raw_data_dir)
}
```

```
# Fetching of data

#anno_AE <- getAE("E-GEOD-70394", path=raw_data_dir, type="raw")


# Targets

#targets <- readTargets(path = raw_data_dir,"targets1.txt",verbose=TRUE
    )
targets=readTargets(infile="targets1.txt",verbose=TRUE)
getwd()

# Now reading Files (normalization)
x <- read.maimages(targets, path=raw_data_dir, source="agilent",green.
    only=TRUE)

y <- backgroundCorrect(x, method="normexp", offset=16)

y <- normalizeBetweenArrays(y, method="quantile")

y.ave <- avereps(y, ID=y$genes$ProbeName)

# Just Analysis

y.ave$E

y.ave$genes

y.ave$targets

############################################################

## creating ExpressionSet object for merging this analysis to the end
    to end workflow tutorial:

SDRF <- AnnotatedDataFrame(targets)

all(colnames(y.ave)==rownames(SDRF))

eset<-new("ExpressionSet", exprs=as.matrix(y.ave),phenoData=SDRF)

pData(eset)

fData(eset)

exprs(eset)

pData(eset)$Treatment
```

```
## principal component analysis

PCA_raw <- prcomp(t(exprs(eset)), scale. = FALSE)

percentVar <- round(100*PCA_raw$sdev^2/sum(PCA_raw$sdev^2),1)
sd_ratio <- sqrt(percentVar[2] / percentVar[1])

## plotPCA

dataGG <- data.frame(PC1 = PCA_raw$x[,1], PC2 = PCA_raw$x[,2],
                     Treatment = pData(eset)$Treatment)

write.csv(dataGG,"dataGG.csv")

pdf("PCAplot.pdf",width=8,height=8,paper='special')
ggplot(dataGG, aes(PC1, PC2)) +
  geom_point(aes(shape = Treatment, colour = Treatment)) +
  ggtitle("PCA plot of normalized data") +
  xlab(paste0("PC1, VarExp: ", percentVar[1], "%")) +
  ylab(paste0("PC2, VarExp: ", percentVar[2], "%")) +
  theme(plot.title = element_text(hjust = 0.5))+
  scale_shape_manual(values = c(4,15)) +
  scale_color_manual(values = c("darkorange2", "dodgerblue4"))
dev.off()

## histogram of log 2 raw intensities

pdf("boxplot_normalized.pdf", width=10, height=10)
oligo::boxplot(eset,col=viridis_pal(option ="C")(n=6),las=2, cex.axis=
    0.4)
dev.off()

arrayQualityMetrics(expressionset = eset,
                    outdir = tempdir(),
                    force = TRUE, do.logtransform = TRUE,
                    intgroup = "Treatment")

#RLE

row_medians_assayData <-
  Biobase::rowMedians(as.matrix(Biobase::exprs(eset)))
RLE_data <- sweep(Biobase::exprs(eset), 1, row_medians_assayData)

RLE_data <- as.data.frame(RLE_data)
RLE_data_gathered <-
  tidyr::gather(RLE_data, patient_array, log2_expression_deviation)
write.csv(RLE_data,"RLEdata.csv")
```

```
#RLE plot

pdf("log2_expressionplot.pdf",width=8,height=8,paper='special') #RLE
    graph
ggplot2::ggplot(RLE_data_gathered, aes(patient_array,
                                       log2_expression_deviation)) +
  geom_boxplot(outlier.shape = NA) +
  ylim(c(-2, 2)) +
  theme(axis.text.x = element_text(colour = "aquamarine4",
                                   angle = 60, size = 6.5, hjust = 1 ,
                                   face = "bold"))
dev.off()

#heatmap

phenotype_names <- ifelse(str_detect(pData
                                     (eset)$Treatment,
                                     "Normal" ), "Normal", "Tumor")

annotation_for_heatmap <-
  data.frame(Phenotype=phenotype_names)

row.names(annotation_for_heatmap) <- row.names(pData(eset))
dists <- as.matrix(dist(t(exprs(eset)), method = "manhattan"))

rownames(dists) <- row.names(pData(eset))
hmcol <- rev(colorRampPalette(RColorBrewer::brewer.pal(9, "YlOrRd"))
    (255))
colnames(dists) <- NULL
diag(dists) <- NA

ann_colors <- list(Phenotype = c( Tumor = "chartreuse4" ,Normal = "
    burlywood3" ))

#Heatmap plot

pdf("heat_plot.pdf",width=8,height=8,paper='special')
pheatmap(dists, col = (hmcol),
         annotation_row = annotation_for_heatmap,
         annotation_colors = ann_colors,
         legend = TRUE,
         treeheight_row = 0,
         legend_breaks = c(min(dists, na.rm = TRUE),
                           max(dists, na.rm = TRUE)),
         legend_labels = (c("small distance", "large distance")))
dev.off()

## histogram
```

```
palmieri_medians <- rowMedians(Biobase::exprs(eset))
pdf("HISTOGRAMplot.pdf",width=10,height=10,paper='special')
hist_res <- hist(palmieri_medians, 100, col = "cornsilk1", freq = FALSE
    ,
                 main = "Histogram of the median intensities",
                 border = "antiquewhite4",
                 xlab = "Median intensities")
dev.off()
man_threshold <- 4.2 #YA pochna ha

pdf("HISTOGRAMplotline.pdf",width=10,height=10,paper='special')
hist_res <- hist(palmieri_medians, 100, col = "cornsilk", freq = FALSE,
                 main = "Histogram of the median intensities",
                 border = "antiquewhite4",
                 xlab = "Median intensities")

abline(v = man_threshold, col = "coral4", lwd = 2)

dev.off()

#linear model

f <- factor(targets$Treatment, levels = unique(targets$Treatment))

#design matrix

design <- model.matrix(~0 + f)

colnames(design) <- levels(f)

write.csv(design, "DESIGN_MATRIX.csv")

#fitting

fit <- lmFit(y.ave, design)

#contrast matrix

contrast.matrix <- makeContrasts("Tumor-Normal", levels=design)

write.csv(contrast.matrix, "contrastmatrix.csv")


fit2 <- contrasts.fit(fit, contrast.matrix)

fit2 <- eBayes(fit2)

output <- topTable(fit2, adjust="BH", coef="Tumor-Normal", genelist=y.
    ave$genes, number=Inf)
```

```
output

write.csv(output, "DEGs.csv")

fit2$genes

# Valcano plot
volcano_names <- ifelse(abs(fit2$coefficients)>=5,
                        fit2$genes, NA)
pdf("volcanoplot.pdf", width=10, height=10)
volcanoplot(fit2, coef = 1L, style = "p-value", highlight = 100,
           names = volcano_names,
           xlab = "Log2 Fold Change", ylab = NULL, pch=16, cex=0.35)

dev.off()

BiocManager::install("EnhancedVolcano")
library(EnhancedVolcano)
pdf("ENHANCED.pdf", width=10, height=10)
EnhancedVolcano(output,
                lab = output $SystematicName,
                x = "logFC",
                y = "P.Value",
                pCutoff = 0.05,
                FCcutoff =0.5,
                title = "MTC vs Normal")
dev.off()

t1<- subset(output, P.Value < 0.05)
t2 <- subset(t1, logFC < -0.5|logFC> 0.5)

write.csv(t2,"final DEGs with cutoff.csv")
#agilent single channel

#libraries

library(limma)
library(convert)
library(Biobase)
library(openxlsx)
library(ggplot2)
library(oligo)
library(oligoClasses)
library(arrayQualityMetrics)
library(simpleaffy)
library(RColorBrewer)
library(pheatmap)
library(geneplotter)
library(stringr)
```

```
library(ArrayExpress)
library(gplots)
library(dplyr)
library(tidyr)
library(matrixStats)
library(genefilter)
#BiocManager::install("simpleaffy")

# Directory Setup
Tutorial_agilent <- "D:/dataset3"

# If Directory exist then good otherwise make one.

Tutorial_agilent
if(!dir.exists(Tutorial_agilent)){
  dir.create(Tutorial_agilent)
}

#set working directory

setwd("D:/dataset3")
raw_data_dir <- file.path(getwd(), "rawDataMAWorkdown")

if(!dir.exists(raw_data_dir)){
  dir.create(raw_data_dir)
}

# Fetching of data

#anno_AE <- getAE("E-GEOD-70394", path=raw_data_dir, type="raw")


# Targets

#targets <- readTargets(path = raw_data_dir,"targets1.txt",verbose=TRUE
    )
targets=readTargets(infile="targets1.txt",verbose=TRUE)
getwd()

# Now reading Files (normalization)
x <- read.maimages(targets, path=raw_data_dir, source="agilent",green.
    only=TRUE)

y <- backgroundCorrect(x, method="normexp", offset=16)

y <- normalizeBetweenArrays(y, method="quantile")

y.ave <- avereps(y, ID=y$genes$ProbeName)
```

```
# Just Analysis

y.ave$E

y.ave$genes

y.ave$targets

############################################################

## creating ExpressionSet object for merging this analysis to the end
   to end workflow tutorial:

SDRF <- AnnotatedDataFrame(targets)

all(colnames(y.ave)==rownames(SDRF))

eset<-new("ExpressionSet", exprs=as.matrix(y.ave),phenoData=SDRF)

pData(eset)

fData(eset)

exprs(eset)

pData(eset)$Treatment

## principal component analysis

PCA_raw <- prcomp(t(exprs(eset)), scale. = FALSE)

percentVar <- round(100*PCA_raw$sdev^2/sum(PCA_raw$sdev^2),1)
sd_ratio <- sqrt(percentVar[2] / percentVar[1])

## plotPCA

dataGG <- data.frame(PC1 = PCA_raw$x[,1], PC2 = PCA_raw$x[,2],
                     Treatment = pData(eset)$Treatment)

write.csv(dataGG,"dataGG.csv")

pdf("PCAplot.pdf",width=8,height=8,paper='special')
ggplot(dataGG, aes(PC1, PC2)) +
  geom_point(aes(shape = Treatment, colour = Treatment)) +
  ggtitle("PCA plot of normalized data") +
  xlab(paste0("PC1, VarExp: ", percentVar[1], "%")) +
  ylab(paste0("PC2, VarExp: ", percentVar[2], "%")) +
  theme(plot.title = element_text(hjust = 0.5))+
  scale_shape_manual(values = c(4,15)) +
```

```
  scale_color_manual(values = c("darkorange2", "dodgerblue4"))
dev.off()

## histogram of log 2 raw intensities

pdf("boxplot_normalized.pdf", width=10, height=10)
oligo::boxplot(eset,col=viridis_pal(option ="C")(n=6),las=2, cex.axis=
    0.4)
dev.off()

arrayQualityMetrics(expressionset = eset,
                    outdir = tempdir(),
                    force = TRUE, do.logtransform = TRUE,
                    intgroup = "Treatment")

#RLE

row_medians_assayData <-
  Biobase::rowMedians(as.matrix(Biobase::exprs(eset)))
RLE_data <- sweep(Biobase::exprs(eset), 1, row_medians_assayData)

RLE_data <- as.data.frame(RLE_data)
RLE_data_gathered <-
  tidyr::gather(RLE_data, patient_array, log2_expression_deviation)
write.csv(RLE_data,"RLEdata.csv")

#RLE plot

pdf("log2_expressionplot.pdf",width=8,height=8,paper='special') #RLE
    graph
ggplot2::ggplot(RLE_data_gathered, aes(patient_array,
                                       log2_expression_deviation)) +
  geom_boxplot(outlier.shape = NA) +
  ylim(c(-2, 2)) +
  theme(axis.text.x = element_text(colour = "aquamarine4",
                                   angle = 60, size = 6.5, hjust = 1 ,
                                   face = "bold"))
dev.off()

#heatmap

phenotype_names <- ifelse(str_detect(pData
                                     (eset)$Treatment,
                                     "Normal" ), "Normal", "Tumor")

annotation_for_heatmap <-
  data.frame(Phenotype=phenotype_names)

row.names(annotation_for_heatmap) <- row.names(pData(eset))
```

```
dists <- as.matrix(dist(t(exprs(eset)), method = "manhattan"))

rownames(dists) <- row.names(pData(eset))
hmcol <- rev(colorRampPalette(RColorBrewer::brewer.pal(9, "YlOrRd"))
    (255))
colnames(dists) <- NULL
diag(dists) <- NA

ann_colors <- list(Phenotype = c( Tumor = "chartreuse4" ,Normal = "
    burlywood3" ))

#Heatmap plot

pdf("heat_plot.pdf",width=8,height=8,paper='special')
pheatmap(dists, col = (hmcol),
        annotation_row = annotation_for_heatmap,
        annotation_colors = ann_colors,
        legend = TRUE,
        treeheight_row = 0,
        legend_breaks = c(min(dists, na.rm = TRUE),
                          max(dists, na.rm = TRUE)),
        legend_labels = (c("small distance", "large distance")))
dev.off()

## histogram

palmieri_medians <- rowMedians(Biobase::exprs(eset))
pdf("HISTOGRAMplot.pdf",width=10,height=10,paper='special')
hist_res <- hist(palmieri_medians, 100, col = "cornsilk1", freq = FALSE
    ,
              main = "Histogram of the median intensities",
              border = "antiquewhite4",
              xlab = "Median intensities")
dev.off()
man_threshold <- 4.2 #YA pochna ha

pdf("HISTOGRAMplotline.pdf",width=10,height=10,paper='special')
hist_res <- hist(palmieri_medians, 100, col = "cornsilk", freq = FALSE,
              main = "Histogram of the median intensities",
              border = "antiquewhite4",
              xlab = "Median intensities")

abline(v = man_threshold, col = "coral4", lwd = 2)

dev.off()

#linear model

f <- factor(targets$Treatment, levels = unique(targets$Treatment))
```

```
#design matrix

design <- model.matrix(~0 + f)

colnames(design) <- levels(f)

write.csv(design, "DESIGN_MATRIX.csv")

#fitting

fit <- lmFit(y.ave, design)

#contrast matrix

contrast.matrix <- makeContrasts("Tumor-Normal", levels=design)

write.csv(contrast.matrix, "contrastmatrix.csv")


fit2 <- contrasts.fit(fit, contrast.matrix)

fit2 <- eBayes(fit2)

output <- topTable(fit2, adjust="BH", coef="Tumor-Normal", genelist=y.
    ave$genes, number=Inf)
output

write.csv(output, "DEGs.csv")

fit2$genes

# Valcano plot
volcano_names <- ifelse(abs(fit2$coefficients)>=5,
                        fit2$genes, NA)
pdf("volcanoplot.pdf", width=10, height=10)
volcanoplot(fit2, coef = 1L, style = "p-value", highlight = 100,
           names = volcano_names,
           xlab = "Log2 Fold Change", ylab = NULL, pch=16, cex=0.35)

dev.off()

BiocManager::install("EnhancedVolcano")
library(EnhancedVolcano)
pdf("ENHANCED.pdf", width=10, height=10)
EnhancedVolcano(output,
               lab = output $SystematicName,
               x = "logFC",
               y = "P.Value",
```

```
               pCutoff = 0.05,
               FCcutoff =0.5,
               title = "MTC vs Normal")
dev.off()

t1<- subset(output, P.Value < 0.05)
t2 <- subset(t1, logFC < -0.5|logFC> 0.5)

write.csv(t2,"final DEGs with cutoff.csv")
```