REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

ANALYSIS OF INTRA-TUMORAL HETEROGENEITY IN CONTEXT OF TISSUE SPECIFIC GENE EXPRESSION WITH COMPUTATIONAL APPROACH

HATİCE BÜŞRA KONUK

MSc. THESIS
DEPARTMENT OF BIOENGINEERING
PROGRAM OF BIOENGINEERING

ADVISER

ASSIST. PROF. DR. ALPER YILMAZ

İSTANBUL, 2018

REPUBLIC OF TURKEY

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

ANALYSIS OF INTRA-TUMORAL HETEROGENEITY IN CONTEXT OF TISSUE SPECIFIC GENE EXPRESSION WITH COMPUTATIONAL APPROACH

A thesis submitted by Hatice Büşra KONUK in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** is approved by the committee on 15.01.2018 in Department of Bioengineering, Bioengineering Program.

Thesis Adviser

Assist. Prof. Dr. Alper YILMAZ

Yıldız Technical University

Approved By the Examining Committee

Assist. Prof. Dr. Alper YILMAZ

Yıldız Technical University _____

Prof. Dr. Musa TÜRKER

Yıldız Technical University _____

Assoc. Prof. Dr. Tunahan ÇAKIR

Gebze Technical University _____

ACKNOWLEDGEMENTS

I would like to thank all those people who made this thesis possible and a valuable experience for me.

Firstly, I would like to express my deepest gratitude to my supervisor Assist. Prof. Dr. Alper YILMAZ for his excellent guidance, valuable suggestions, encouragement, patience, and conversations full of motivation. I am glad to have the chance to study with him.

I would like to express my deepest gratitude to Res. Assist. Muhammet Raşit CESUR for the constant support while learning the programming language.

I would like to express my deepest gratitude to Dr. Ensar YALÇINKAYA for his meaningful support during categorization of tissue types.

I would like to express my deepest gratitude for the constant support in every step in my life, understanding, their belief in me and love that I received from my parents, especially my little brother Mehmet Emre KONUK.

February, 2018

Hatice Büşra KONUK

TABLE OF CONTENTS

Page

LIST OF SYMBOLS

μ	mean
σ	standard deviation
τ	tau score
Hg	entropy

LIST OF ABBREVIATIONS

A	Adenine
AI	Artificial intelligence
BCT	Breast conservation therapy
BGLAP	Bone gamma-carboxyglutamate protein
BLCA	Bladder Urothelial Carcinoma
Bp	Billion base pair
BRCA	Breast Invasive Carcinoma
BTO	Brenda Tissue Ontology
CAGE	Cap analysis of gene expression
Chıpchip	Chromatin immunoprecipitation coupled to DNA microarray
Chıp-seq	Chromatin immunoprecipitation coupled to DNA sequencing
COAD	Colon Adenocarcinoma
CST1	Cystatin SN
DAVID	The Database for Annotation, Visualization and Integrated Discovery
DEG	Differentially expressed genes
DKKL1	Dickkopf like acrosomal protein 1
DSA	Digital Sorting Algorithm
ECM	Extracellular matrix
EE	Expression enrichment
ELAVL3	ELAV like RNA binding protein 3
ENCODE	Encyclopedia of DNA Elements Consortium Project
ESCA	Esophageal Carcinoma
ESTs	Expressed sequence tags
FATE1	Fetal and adult testis expressed 1
FC	Fold change
FCM	Fuzzy c-means
FPKM	Fragments per kilobase of exon model per million mapped reads
G	Guanine
GEJ	Gastroesophageal junction
GTEx	Genotype-Tissue Expression
GWAS	Genome-wide association studies
HCC	Hepatocellular carcinoma
Hg	Shannon Entropy
HGP	The Human Genome Project
HIVE	High Performance Integrated Virtual Environment

Kb	kilobases
KIRC	Kidney Renal Clear Cell Carcinoma
KIRP	Kidney renal papillary cell carcinoma
KRT6C	Keratin 6C
K–S	The Kolmogorov–Smirnov
IDE	Integrated Developmental Environment
LEMD1	LEM Domain 1
lncRNAs	Long non-coding RNAs
LIHC	Liver Hepatocellular Carcinoma
LR	Likelihood ratio test
lRNAs	Long RNAs
LUAD	Lung Adenocarcinoma
MARCH10	Membrane associated ring-CH-type finger 10
MAT1A	Methionine adenosyltransferase 1A
miRNAs	Micro RNAs
MGFM	Marker Gene Finder in Microarray Data
MMP13	Matrix metalloproteinase
mRNAs	Messenger RNAs
MT	Metastatic Tumor
NCI	National Cancer Institute
ncRNAs	Non-protein-coding RNAs
NGS	Next-generation sequencing
NSCLC	Non-small cell lung cancer
NT	Solid Tissue Normal
ORA	Over-Representation Analysis
OS	Overall survival
PAAD	Pancreatic Adenocarcinoma
PDAC	Pancreatic ductal adenocarcinoma
PEM	Preferential Expression Measure
PGM	Personal Genome Machine
PLS	Partial least squares
RA	Regression analysis
RCC	Renal cell carcinoma
RPKM	Reads per kilobase of exon per million reads mapped
rRNA	Ribosomal RNA
RNA-Seq	RNA-Sequencing
SAGE	Serial analysis of gene expression
SCC	Squamous cell carcinoma
SFTPD	Surfactant protein D
SLFNL1	Schlafen like 1
SPM	Specificity Measure
sncRNAs	Short non-coding RNAs
solid	Sequencing by Oligo Ligation Detection
sRNAs	Short RNAs
T	Thymine
TCGA	The Cancer Genome Atlas
TFs	Transcription factors

THCA	Thyroid carcinoma
TNNT1	T1 slow skeletal type
TNNI3	Troponin I3, cardiac type
TPM	Transcripts per million
TP	Primary Solid Tumor
TR	Recurrent Solid Tumor
tRNA	Transfer RNA
TSI	Tissue specificity index
NCBI	National Center for Biotechnology Information Search database
PaGenBase	A pattern gene database for the global and dynamic understanding of gene function
C	Cytosine
HK	Housekeeping
Pap	Papanicolaou
TCGA	The Cancer Genome Atlas
EMBL	The European Bioinformatics Institute
KEGG	Kyoto Encyclopedia of Genes and Genomes
OMIM	Online Mendelian Inheritance in Man
PDBE	Protein Data Bank in Europa
TiGER	Tissue-specific Gene Expression and Regulation
TiSGeD	A database for tissue-specific genes
TSGene	Tumor Suppressor Gene Database
UBC	Urinary bladder cancer
UCEC	Uterine Corpus Endometrial Carcinoma
WebGestalt	WEB-based Gene SeT AnaLysis Toolket

LIST OF FIGURES

Page

LIST OF TABLES

Page

ABSTRACT

ANALYSIS OF INTRA-TUMORAL HETEROGENEITY IN CONTEXT OF TISSUE SPECIFIC GENE EXPRESSION WITH COMPUTATIONAL APPROACH

Hatice Büşra KONUK

Department of Bioengineering

MSc. Thesis

Adviser: Assist. Prof. Dr. Alper YILMAZ

Identification of tissue-specific genes is essential for understanding molecular mechanisms. Intra-tumoral heterogeneity as diversity among cells and layers of tissues in only one single tumor limits therapeutic efficacy. Tumor heterogeneity is an important challenge for successful personalized medicine.

We aimed to identify tissue-specific genes rigorously for examining intersection between differentially expressed genes (DEG) and tissue specific genes, analyzing cancer gene expressions in terms of tissue specificity, demonstration and interpretation of intra-tumoral heterogeneity, revealing specific molecular targets for tumors and elucidating roles of tissue-specific genes in biological processes.

Gene expression, derived from five RNA-Sequencing projects, spanning 96 human tissues were used. Detection of tissue-specific genes not only included tau, but also required integrating of tau and statistical distance. This new method is defined as "extended tau". We investigated intersection of 16 different cancer DEG with specifically expressed genes in corresponding tissue. After that, gene expression data for 11 primary solid tumors, retrieved from The Cancer Genome Atlas (TCGA), were analyzed by integrating our tissue specificity findings. Significant genes obtained after all calculations were functionally annotated for identifying their roles.

Consequently, we successfully assigned genes to multiple tissues by extended tau. Then discovered that DEG in a cancer tissue has low overlap with genes specific to that tissue. Most importantly, we identified candidate biomarkers for heterogeneity exhibiting gene expression in cancer tissue although its expression is restricted to unrelated tissue. The

genes identified in our study were already shown in literature to be associated with tumor heterogeneity or genetic instability of cancer cells for various cancers. Our results are likely to contribute to the road paved by many researchers trying to understand cancer for many years by bringing the tissue-specific genes into the scene.

Keywords: Tissue specific genes, RNA-Sequencing, tau score, statistical distance, tumor heterogeneity, genetic instability of cancer cell.

ÖZET

TUMOR İÇİ HETEROJENİTENİN DOKUYA SPESİFİK GENLER İLE İLİŞKİSİNİN HESAPSAL YAKLAŞIMLA İNCELENMESİ

Hatice Büşra KONUK

Biyomühendislik Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Yrd. Doç. Dr. Alper YILMAZ

Dokulardaki moleküler mekanizmaların anlaşılması için dokuya özgü genlerin tanımlanmalarına ihtiyaç vardır. Tümör içi heterojenite tek bir tümör dokusu içinde çeşitli hücre ve doku tabakalarının bulunması ile ortaya çıkan karmaşık bir süreçtir ve tedavilerde önemli bir zorluktur.

Dokuya spesifik genlerin titiz bir şekilde tanımlanması, çeşitli kanserlerde ifadesi değişen genler ve doku spesifik genlerin kesişiminin incelenmesi, doku özgünlüğü kapsamında kanser gen profillerinin tümör içi heterojeniteyi göstermek için analiz edilmesi ve yorumlanması, farklı kanserler için moleküler hedeflerin ortaya konulması ve dokuya özgü genlerin biyolojik süreçlerdeki rollerinin anlaşılması amaçlanmıştır.

96 farklı insan dokusunu kapsayan beş RNA-dizileme projesinden elde edilen gen ifadesi verileri kullanılmıştır. Doku özgüllüğünün titizlikle belirlenmesi için tau ve istatistiksel mesafe entegre edilerek kullanılmıştır. Bu yeni yöntem "genişletilmiş tau" olarak tanımlanmıştır. 16 farklı kanser türüne ait ifadesi değişen genlerin (DEG) dokuya özgü genlerle kesişimleri incelenmiştir. Bundan sonra, 11 primer solid tümör için genlerin ifadesi, dokuya özgün gen sonuçlarının birleştirilmesiyle analiz edilmiştir. Tüm hesaplamalardan sonra elde edilen anlamlı genler işlevsel olarak yorumlanmıştır.

Sonuç olarak, doku özgünlüğü için yeni bir yaklaşım kullanarak genler özgün olarak ifade edildikleri birden fazla dokuyla başarıyla eşleştirildi. Daha sonra bir kanser dokusundaki ifadesi değişen genlerin dokuya özgü genlerle çok az örtüştüğü ortaya konuldu. En önemlisi, farklı bir dokuyla sınırlı ifadesi olmasına rağmen kanserde de ifade edilen ve

heterojenite için aday biyolojik belirteçler tespit edildi. Çalışmamızda belirlenen genlerin, tümor heterojenitesi veya kanser hücrelerinin genetik kararsızlığıyla ilişkili olduğu çeşitli kanser türleri için literatürde gösterilmiştir. Elde edilen sonuçlar, dokuya spesifik genleri gündeme getirerek uzun yıllardır kanseri anlamaya çalışan bir çok araştırmacının çalışmalarına katkıda bulunacaktır.

Anahtar Kelimeler: Dokuya spesifik genler, RNA dizileme, tau skoru, istatistiksel mesafe, tumor heterojenitesi, kanser hücresindeki genetik kararsızlık.

CHAPTER 1

INTRODUCTION

1.1 Literature Review

Cancer is one of the common cause of death in the world. The poor diagnoses, insufficient therapies, side effects, prognoses and pathogenesis of various cancers are very important to identify mechanism and improve new effective treatment methods. Complicated mechanisms of cancer cells can be mainly affected by stages, locations, tumor microenvironment and heterogenic structure of cancerous tissues, cell differentiation and origin. Bioinformatics can be considered one of critical scientific area for clinically relevant challenges in early diagnosis, efficient and targeted therapies, and predictive prognosis of patients with various cancer types and also other diseases. Developing methods specific for cancer bioinformatics or generating new and advanced tools to answer the specific questions about cancer and identifying molecular mechanisms and related pathways are needed in cancer research [1]. To sum up, cancer is a fatal disease in the world and development of effective therapies, early diagnosis and targeted treatment can be possible using bioinformatics and computational approaches in the near future.

Healthy tissues/organs like liver, lung, kidney, esophagus, heart and brain in human body have heterogenic structure and so do solid tumor tissues. Because tissues and organs have several different layers like epithelial layers and moreover they have various cell types such as epithelial cells, endothelial cells, immune cells, dendritic cells, fibroblasts and tissue related cell types for instance, hepatocyte for liver tissue etc. Although, heterogeneity in both healthy tissues/organs and cancerous tissues is a previously known concept, its impacts on the carcinogen processes and treatment are

poorly understood [2]. There are two main definitions about tumor heterogeneity, first one is “inter-tumoral heterogeneity” that is found among different tumors in individual patient, and/or among different patients’ cancer tissues. Second one is “intra-tumorol heterogeneity” that is within a single cancerous tissue [3]. Intra-tumoral heterogeneity has important implications for personalized medicine because it can limit therapeutic efficacy and lead to resistance to cancer treatment. Hence, tumor heterogeneity is one of the major problem limiting the effectiveness of therapies, inconvenient treatment outcomes, increasing number of metastatic tumors and also high mortality rate from cancer [4]. The molecular complexity of tissues – other terms of this is heterogeneity of tissues as structure - and solid tumors limit the discovery of specific targets for tissue-specific delivery of therapeutic molecules [5]. Determination of tissue-specific genes might be useful to understand intra-tumoral heterogeneity of tumors and to find out new effective biomarkers for both early diagnosis and targeted treatment especially for solid tumors. Tissue-specific genes which are expressed only one or several tissues or cell types specifically can be calculated using analysis of biological data approaches. Determination of tissue-specific genes for solid tumors can be critical in order to understand their molecular mechanisms and biological processes in cancerous tissue, identify tumor heterogeneity and develop targeted treatment and diagnostic panels using tissue-specific markers.

1.2 Objective of the Thesis

In this thesis, we aim to generate list of genes expressed in tissue-specific manner after robust and rigorous analysis of 96 different tissue types using a new computational approach that is described as “extended tau” specificity calculation. And then, we aim to compare and overlay differentially expressed gene data with cancer expression profiles for different solid tumors. After revealing tissue-specific gene lists for each tissue, differentially expressed genes and expression profiles of all genes in cancer patients will be examined in the context of tissue-specific genes using some strict criteria. We aim to find out solid tumor specific biomarkers, understanding cancer cell mechanisms and showing intra-tumoral heterogeneity in terms of tumor microenvironment. For this purpose, intersection of differentially expressed genes in a single tumor and tissue-specific genes for corresponding tissue will be investigated

because overlapping genes might give information to us about tumor specific targets. Expression of all protein coding genes in various cancer types is analyzed to compare with tissue-specific genes in related tissues and also irrelevant tissues as an interesting approach for identifying intra-tumoral heterogeneity and features of microenvironment. Another objective of this thesis study is to attach annotated data to selected genes in the previous step, about their functionality and their role in biological processes. The results originating from this study have potential to provide crucial knowledge about intra-tumoral heterogeneity, comprehensive and complex mechanisms of cancer cells for each tumor. Statistical analysis was conducted using R statistical programming language. Briefly, objectives of the thesis,

- To calculate robust and rigorous classification of tissue-specific genes for related tissues by computational approaches using R programming language,
- To determine differentially expressed genes obtained from a database for each cancer type by computational approaches using R programming language,
- To intersect these two group of genes for each cancer,
- Analyze the cancer expression of all protein coding genes in terms of tissue specificity by using stringent criteria,
- To annotate genes after all calculations and analysis with some bioinformatic tools,
- To discuss intra-tumoral heterogeneity related to tumor microenvironment, cancer cell behaviors and tissue-specific targets for each cancer types.

1.3 Hypothesis

By using RNA-Seq data which is proven to be more effective than microarray data, from various tissue expression studies, we can identify genes expressed specifically in a tissue by improving tau score calculation. After having identified tissue specific genes, complex alterations in gene expression of cancer cells can be elucidated in the context of tissue-specific genes. Finally, we hypothesize that tumor heterogeneity of a solid tumor can be tackled with analyzing the gene expression of that tumor sample. A gene expressed in a tumor sample despite being specific or restricted to another tissue should give clue about source of tissues in heterogeneous tumor environment.

CHAPTER 2

GENERAL INFORMATION ABOUT BIOINFORMATICS AND CANCER RESEARCH

2.1 Bioinformatics and Human Genome

Bioinformatics term was first used by Paulien Hogeweg and Ben Hesper in 1970 as "the study of informatic processes in biological systems" [6]. Bioinformatics can be simply described as application of computer science algorithms to biological problems pertaining to any organism. Rapid growth of biological data increased the necessity for bioinformatics approaches in many fields of research. Bioinformatics generally deals with,

- Genome(s)
- Chromosome(s)
- Gene(s) (DNA or RNA)
- Protein(s)
- Metabolite(s)

Other definition of bioinformatics is "application of computational techniques to understand and establish the information associated with biological molecules like DNA, RNA, protein, enzymes and other macromolecules" as a new discipline in various areas, and encompasses a wide range of subject areas from structural biology, genomics to gene expression studies [7]. Bioinformatics develops new methods or tools for not only analysis but also recovery, storage and organization of biological data.

Bioinformatics analyses generally focus on three primary data sources:

- DNA, RNA or protein sequences,
- Macromolecular structures,
- Functional genomics experiments results.

As the rate at which of biological data accumulates, computational methods have become indispensable to biological investigations.

Bioinformatics is an interdisciplinary area which is related to lots of different sciences shown in Figure 2.1:

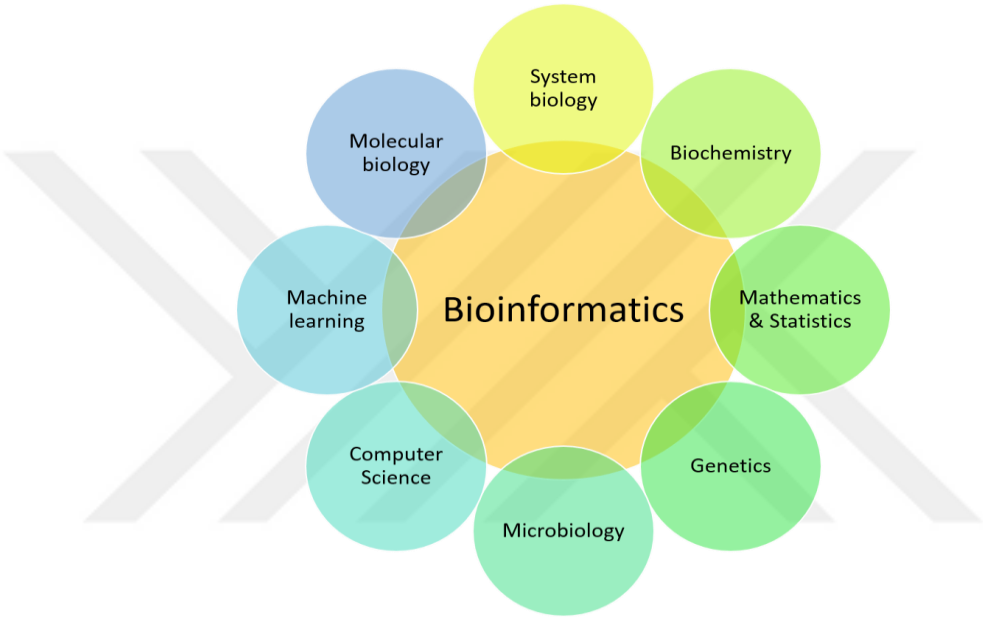


Figure 2.1 Bioinformatics and related sciences

There are some definitions which are important for bioinformatic studies:

Omics: Omics refers to files of studies such genomics, transcriptomics, proteomics, epigenomics, microbiomics and metabolomics.

Multi-Omics: Overlap and interaction between all omic sciences such as microbiomics, metabolomics, proteomics, epigenomics, transcriptomics and also genomics.

Genomics: All the genes in the human genome together, including their interaction with each other, the environment, and the influence of other factors. This definition can be extended for other omics fields such as proteomics referring to all proteins within a cell and interactions among them.

Database: Database is simply defined as storage of biological data as computer language. They are variously classified on varying basis like data type, source, organism, experiment type, location etc.

Tools: Tool is a software for biological system developed to perform various tasks over the stored data such as searches, analysis, submission, annotations, alignment etc. [8].

2.1.1 Bioinformatics Applications

There are lots of different studies or areas using bioinformatics, the list below is a small portion of continuously extending list of applications:

- **Gene expression studies:** Gene expression experiments quantify the expression levels of each gene in genome. These experiments measure the amount of mRNA or protein products quantitatively that are produced by living cells in an organism. Gene expression levels give useful and comparative information between two conditions [9].
- **Sequence data (Nucleotide, DNA, RNA, Proteins etc.):** Genomic studies in bioinformatics have generally deal with DNA sequence of model organisms for microorganism, plants and animals [7].
- **Molecular networks:** A network is defined as graphs with annotated nodes and edges. Pathways or networks are important for the discovery of biological processes and molecular mechanisms of the living cells and organisms [10].
- **Molecular structures:** Analyses of structures of molecules have provided insight into the stereochemical principles of binding, bonds in molecules, 3-D structure of molecules like protein, DNA or RNA [7].
- **Protein homologues:** Proteins with similar amino acid sequences and functions in different species is defined as homologous protein. Identification of protein homologues between closely related or distant species is a routine procedure in genetic applications in order to understand protein functions and have insight about disease mechanisms [11].
- **Drug Design:** One of the important medical applications of bioinformatics has been in improving rational drug design for various human diseases [7].

- **Large-scale censuses:** Although databases can effectively and rigorously store all the genome information, structures and expression datasets, it is useful to condense all these information into understandable trends so that users can comprehend and use large datasets easily [7].

2.1.2 Databases and Tools

There are lots of databases and tools about bioinformatics applications. The important ones are shown in the Table 2.1 below:

Table 2.1 General bioinformatics databases and tools

Databases and Tools	Web site
NCBI (National Center for Biotechnology Information Search database)	https://www.ncbi.nlm.nih.gov/
The Human Protein Atlas	https://www.proteinatlas.org/
UniProt	http://www.uniprot.org/
OMIM (Online Mendelian Inheritance in Man)	https://www.omim.org/
PDBe (Protein Data Bank in Europa)	https://www.ebi.ac.uk/pdbe/
EMBL (The European Bioinformatics Institute)	https://www.ebi.ac.uk/
KEGG (Kyoto Encyclopedia of Genes and Genomes)	http://www.genome.jp/kegg/
Ensembl	https://www.ensembl.org/index.html
Gene Ontology Consortium	http://www.geneontology.org/
DrugBank	https://www.drugbank.ca/

Table 2.1 General bioinformatics databases and tools (cont'd)

SNPs3D	http://www.snps3d.org/
Tumorscape	http://portals.broadinstitute.org/tumorscape/pages/portalHome.jsf
TSGene (Tumor Suppressor Gene Database)	https://bioinfo.uth.edu/TSGene/
TCGA (The Cancer Genome Atlas)	https://cancergenome.nih.gov/
ArrayExpress	http://www.ebi.ac.uk/arrayexpress
Expression Atlas	http://www.ebi.ac.uk/gxa

2.1.3 Bioinformatics and Human Genome

The Human Genome Project (HGP) was officially initiated in the United States under the direction of the National Institutes of Health and the United States Department of Energy with a 15-year, using $3 billion for completing and understanding the human genome sequence in 1990. HGP was the international, collaborative study whose goals were completing the human gene mappings, understanding and identifying all the genes of human genome. The genome sequencing was performed by a whole-genome random shotgun method. After sequencing of a batch is completed, analyzes after the sequencing were made with bioinformatic methods and tools. These analyses include external data processing, pre-assembly, proto I/O file generation, genome assembly and characterization, gene prediction and annotation, genome structure, genome evolution, sequence variations, prediction of protein-coding genes and other gene types such as transcription factors and RNA types. Assembly strategies are very important for HGP. Thus, bioinformatic analysis play important role to map, determine and understand human genome [12].

As a summary, 2.91 billion base pair (bp) of the euchromatic human genome was generated by the whole-genome shotgun sequencing method in HGP. In addition to

HGP, bioinformatics played crucial role in many other genome projects and for public access to their output.

Bioinformatics is not only essential for analysis and discovery but also storage of data and access to it. Collection, analysis, annotation and storage of the ever-increasing amounts of sequenced, mapped and calculated expression data in publicly accessible, user-friendly databases is critical and also significant for genome project's success. In addition, the researchers need computational approaches and methods that will allow scientists to keep, extract, view, annotate and analyze genomic information efficiently and easily [13]. Bioinformatic analyses are also needed in a lot of experimental studies. Therefore, bioinformatics is very crucial for genome projects, understanding the human genome and diseases.

2.2 Bioinformatics in Cancer Research

Huge biological data collected during biological research are publicly available, thanks to powerful research technologies like bioinformatics. Bioinformatics is a new multidisciplinary field and uses advanced biology, genetics, computer science, statistics, chemistry, biochemistry, mathematics and different technological areas to store, manage, analyze and understand the biological data effectively and dynamically [14]. Cancer, the leading cause of deaths all over the world, is a complex disease occurring in multiple organs per system or both organs and systems in human body. There are more than 100 types of cancer in human. Poor diagnosis, weak activity of treatment methods, hard side effects of therapies, tumoral heterogeneity within only one single tissue in a human, molecular mechanisms are not yet discovered completely. A better understanding of the expression of genes and other regulatory RNAs and the networks between them is necessary for the investigation and early detection of the molecular mechanism of cancer. At this point bioinformatics enters the cancer research shown in Figure 2.2 and it is very crucial for cancer studies. Bioinformatics offers many unique and useful approaches to the use for genomic data. The need for bioinformatics is becoming more apparent as the current cancer data rapidly increases from day to day, as various technologies develop [1].

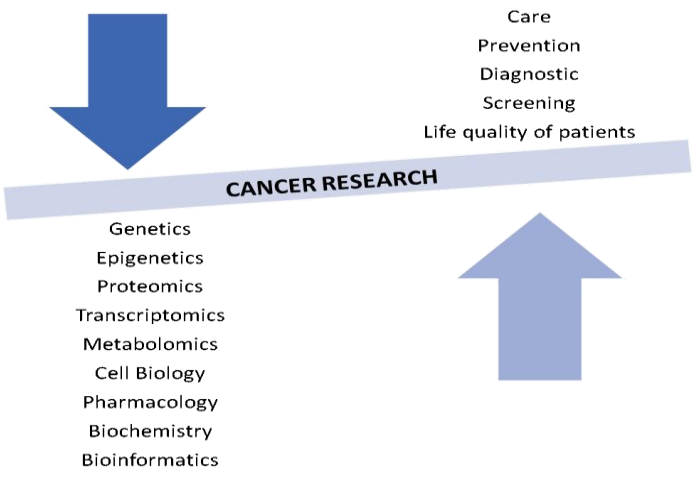


Figure 2.2 Relationship between cancer research and bioinformatics

National Cancer Institute (NCI) [15] has played an important role in improvements of genomics, proteomics, transcriptomics, metabolomics and, imaging to increase our understanding of the molecular pathways and of all cancer types. Cancer bioinformatics has played a vital role in the identification and validation of biomarkers, specific to clinical phenotypes related to early diagnoses, measurements to monitor the progress of the disease and response to therapy, and predictors for the improvement of life quality [1]. Completion of HGP in 2003 by collaborative research had allowed bioinformatics to be applied in the cancer diagnosis and treatment [16]. There are many studies according to PubMed [17] when we count results of the "cancer and bioinformatics" query, as shown in Figure 2.3 below:

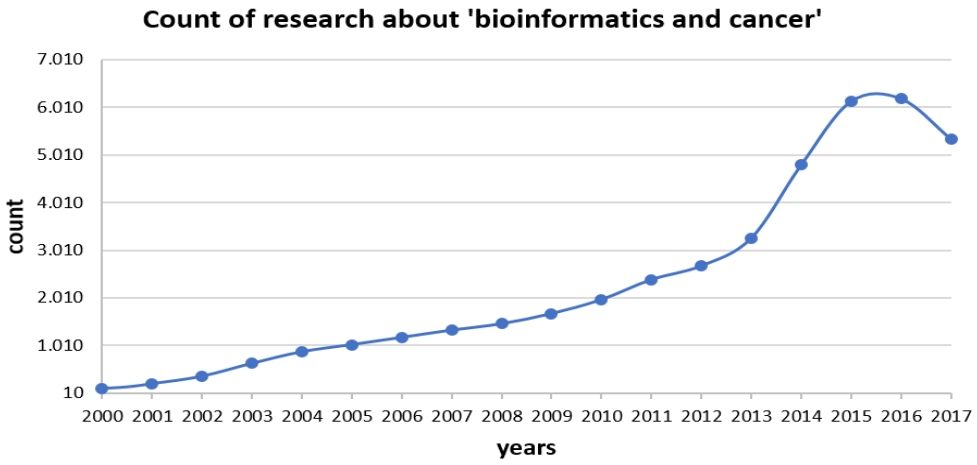


Figure 2.3 Number of cancer and bioinformatics research based on Pubmed query

If we give some example applications of bioinformatics for cancer research:

- One of these applications is to use the computational models that represent biological data and information to know about the quantity of cancer cells, tissues in the body or about the biological state and disease level of the patient [16].
- Many studies have indicated that gene expression studies and classification of expression level of genes in cancer cells is obligatory, and this will ensure efficient results after the treatment. Identification of gene expression associated with disease permit fast and reliable diagnosis of conditions in each cancer level and subtypes. Gene expression data generated from bioinformatic applications is also important for drug response and tumor response also [16].
- To analyze immune response of therapy is controlled by therapists easily using bioinformatics for cancer patients [16].
- Therapists can detect the absence of a gene or mutation using bioinformatic tools, it is important to understand cancer risks for all people [16].
- Gene expression patterns of cancer cells are compared with normal cells or other subtypes of the corresponding cancer type, and genes up regulated or down regulated in related cancerous tissue are organized and clustered using bioinformatics analysis [18].
- Identification of signatures of metastasis, angiogenesis or prediction of clinical outcome can be detected using bioinformatics approaches, also [18].
- Proteomics is valuable in the discovery of biomarkers thanks to bioinformatic analysis. For instance, proteomics can be used to understand biomarkers for cancer diagnosis, to monitor disease progression, and to exhibit therapeutic targets. Thus, bioinformatic tools are needed at all levels of proteomic analysis, especially for cancer patients [19].
- Bioinformatics are associated with the identification of drug targets and are crucial for better drug design as a targeted effective treatment method in cancer therapy [20].

To sum up, molecular mechanisms of more than 100 cancer types, effective treatment methods, quick and accurate diagnosis, identification and validation of novel biomarkers, understanding networks and molecular pathways, and personalized medicine in cancer can be achieved using bioinformatics or computational approaches *in silico.* Integration of methodologies, softwares, computational tools and databases applied specifically to cancer data have great potential to facilitate cure for cancer [1].

2.3 Next-Generation Sequencing Technology

Next-generation sequencing (NGS), also called as high-throughput sequencing, is an umbrella term used to describe many different modern sequencing technologies including:

- Roche 454 sequencing [21],
- Solexa/Illumina sequencing platform [22],
- Sequencing by Oligo Ligation Detection (SOLiD) [23],
- Ion Torrent Personal Genome Machine (PGM) [24],
- Qiagen- intelligent bio-systems sequencing by synthesis [25],
- Polony sequencing [26],
- Single molecule detection system (Helicos BioSciences) [27],
- Pacific Biosciences (the PacBio) [28].

These recent technologies allow us to sequence DNA and RNA much more quickly and cheaply than Sanger technology. Therefore, NGS technology has revolutionized genomic and molecular biology research.

2.3.1 What is NGS?

Sanger and colleagues [29] and Maxam and Gilbert [30] had developed methods to sequence DNA by chain termination and fragmentation techniques, respectively, in 1977. Initially, HGP was started using Sanger technique in 1990 [31] and was completed using NGS technologies instead of Sanger.

Big number of DNA fragments or reads generated by NGS technology enabled the sequencing of entire genomes rapidly [32]. NGS technology paved the way for too many significant discoveries in life sciences by supplying a comprehensive overview of the genomic, transcriptomic and epigenomic state of various cells and organisms. It became more and more attractive and effective as a platform to investigate gene expression. In this context, NGS measures gene expression by generating short reads or fragments which are of 35–300 base pairs sequences [33].

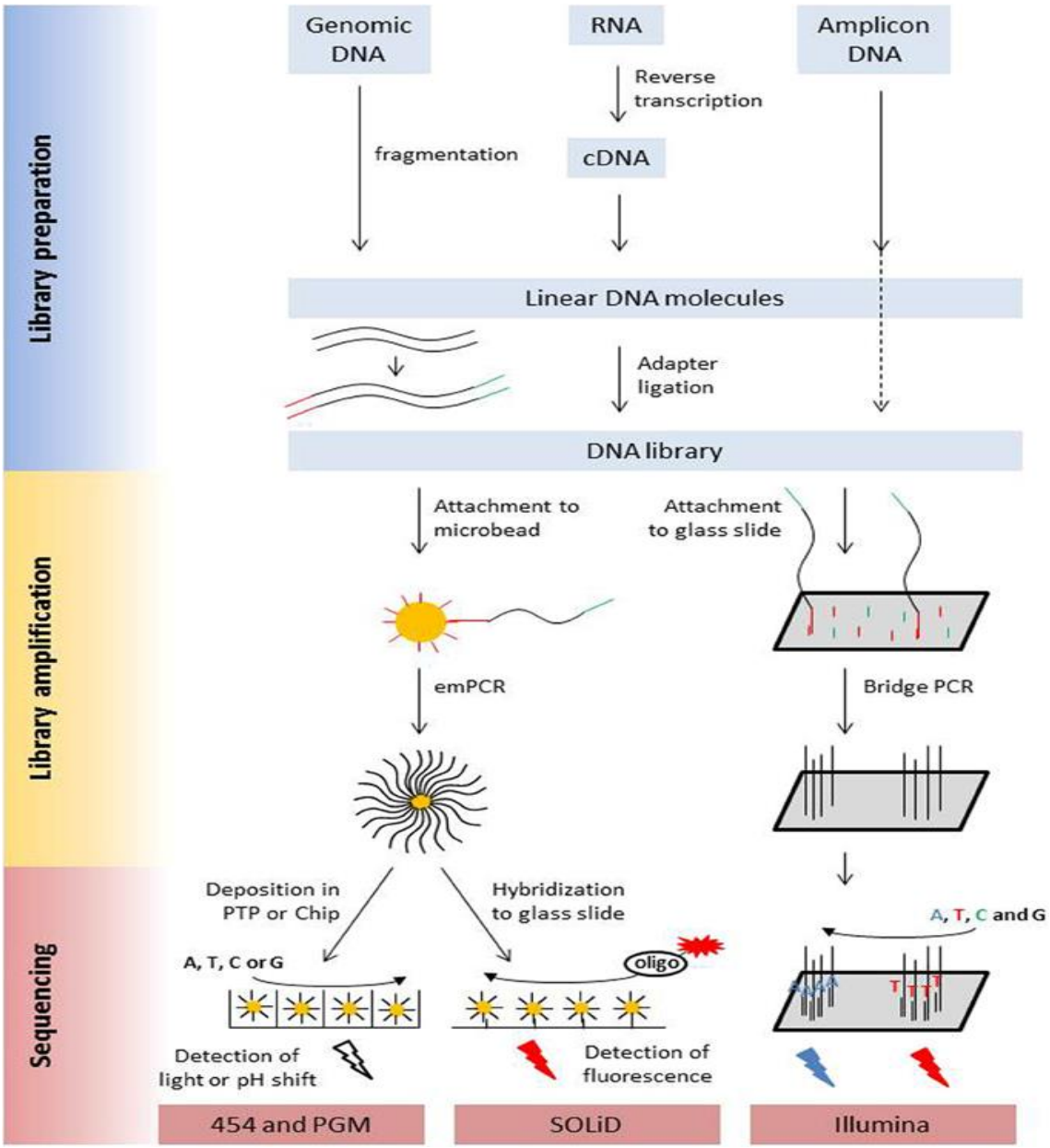


Figure 2.4 Next generation sequencing technology steps in brief [34]

General steps of NGS including **library preparation, clonal amplification, sequencing and validation** are demonstrated in Figure 2.4, briefly and also explained below.

2.3.1.1 Library Preparation

Library preparation is the first step of NGS:

- Fragmentation (sonication, nebulization, shearing)
- End-pair (size selection, blunt-end or A overhang)
- Adaptor ligation (ligation, purification)

2.3.1.2 Clonal Amplification

Clonal amplification is the second step of NGS:

- Bridge amplification (by Illumina)
- Emulsion PCR (by Roche)

2.3.1.3 Sequencing

Sequencing is the third step of NGS:

- Sequencing by synthesis (by Ion Torrent, Illumina)
- Sequencing by oligo ligation
- Other sequencing types involve single molecule sequencing real time sequencing, Nanopore technology, etc.

2.3.1.4 Validation

Validation is the last step of NGS:

- To ensure analytical sensitivity
- To ensure analytical specificity
- To ensure correct region of interest
- To ensure adequate coverage- sequence depth of more than 30 reads

2.3.2 Advantages and Disadvantages of NGS Technology

NGS technology has dramatically reduced both sequence cost and time to decode an entire human transcriptome by sequencing RNA or DNA in a massively parallel fashion. NGS is used for RNA sequencing which is a feasible way for obtaining global transcriptome information to determine potential side effects of drugs and chemicals on public health. Before NGS microarray technology was used mainly. Microarray experiment has been de facto method to find out expression profile of genes, in other words, by quantifying thousands of expression levels at the same time [35]. Comparison of microarray and NGS technologies, as well as different data analysis strategies, shows us their advantages and limitations or disadvantages and moreover, research shows that there is a migration from microarray to NGS thanks to the advantages of NGS like low cost, high efficiency, in high accuracy results and short experimental time [36]. Correlation between RNA-seq and microarray data sets is compatible with each other, although the detection of lowly expressed genes is better via RNA-seq than with microarray experiment [37], [38].On the other hand, there are some disadvantages of NGS technology. NGS is difficult to be standardized and it is also required for meaningful statistical validation of sequencing [39]. Comparison of microarray and NGS technology was carried out in Table 2.2:

Table 2.2 Comparison of microarray and NGS technology [40]

Microarray Technology	NGS Technology
Design requires a priori knowledge of genomic features. It might be incomplete, incorrect when newer genome annotations are released.	Knowledge of genome can be helpful, but not required. In fact, data from NGS platform has recently been assembled into a genome de novo.
High signal-to-noise ratios are a relative measure, limits the dynamic range of high-confidence data. This makes the detection of low-abundance sequences difficult and quantitative.	NGS technology has unlimited full quantitative signal range.
Micrograms of DNA are needed to hybridize to arrays. Moreover, PCR-based amplification of DNA or RNA material can introduce bias into samples.	Nanograms of DNA or RNA material are sufficient for NGS.

Table 2.2 Comparison of microarray and NGS technology [40] (cont'd)

Microarray formats, preparative methodologies for experiment and analytical approaches may limit the reproducibility of data.	NGS has the same data output, therefore reproducibility of experiment is higher, and bioinformatics analysis of data is simple.

2.3.3 NGS Technology Applications

NGS is mainly used for DNA and RNA sequencing of a living cell in an organism.

Genomic DNA sequencing: NGS is used for whole human DNA sequencing and the first large-scale human genetic variation study, the 1000 Genomes Project [41], involves the sequencing of thousands of individuals. After that some tools are generated for analysis of genomic DNA sequencing [42].

RNA sequencing: RNA sequencing is used mainly to determine gene expression levels for healthy and disease conditions and there are many tools for the analysis of RNA sequencing results [43]. At first, RNA sequencing was often used to generate protocols that did not preserve strand information. However, the eukaryotic transcriptome is much more complex than previously thought and many genes produce antisense transcripts [32].

Location-based techniques: Originally, ChIP-seq was developed to determine *in vivo* protein–DNA interactions as a proteomics data [32].

2.3.4 RNA Sequencing Technology

RNA-Sequencing (RNA-Seq) gives extensive details about the RNA landscape within a living cell [44]. Processes of RNA-Seq data is fundamentally different from microarray data. One of the main advantage of RNA-Seq is that it can capture transcriptome dynamics across different tissues or conditions without sophisticated normalization of data sets [45].

2.3.4.1 RNA Sequencing Applications

RNA-Seq not only quantifies gene expression levels but also provides the measurement of levels of transcripts and their isoforms for each gene. Differentially expressed genes

between two different conditions, summarized at the genomic level of interest, such as genes or exons are confirmed are discovered using RNA-Seq [46]. General steps of RNA-seq was defined in Figure 2.5 simply. RNA-Seq is used in many different areas such as to find out novel genes [45], single nucleotide polymorphisms [47], different mutations [48], transcript assembly [49], allele specific expression [50] and splicing and other forms of alternative isoform-specific expression [51].

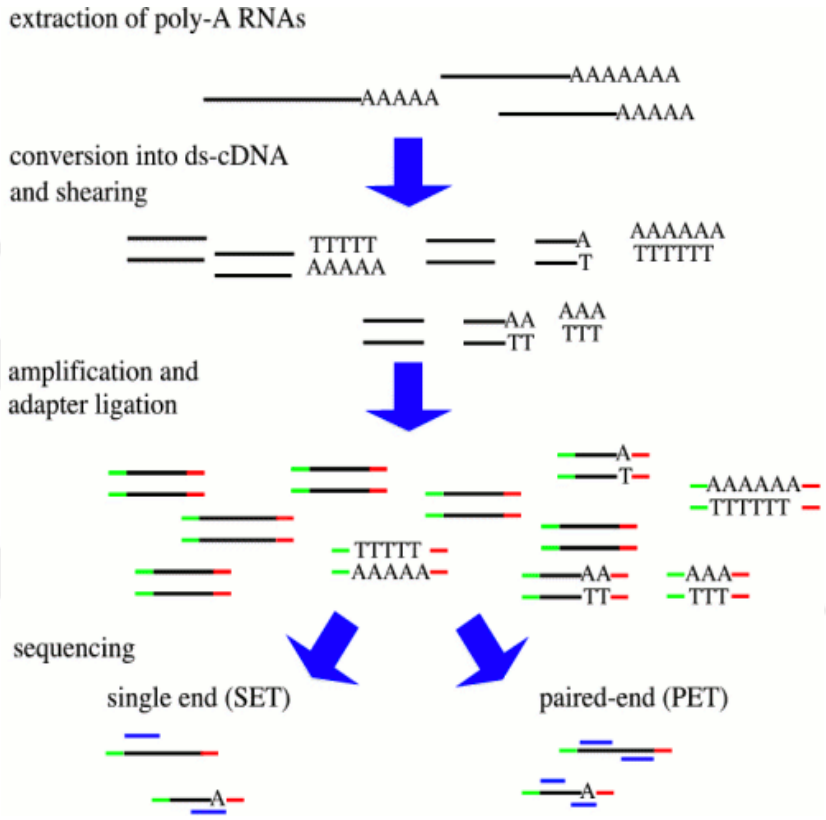


Figure 2.5 Basic RNA sequencing steps [52]

2.3.5 NGS Technology and Human Genome

NGS technologies are currently the popular topic in genomics research. Since 2005, NGS technologies have been used in human and animal genome research. Some applications of NGS include; the analysis of chromatin immunoprecipitation coupled to DNA microarray, ChIPchip or sequencing ChIP-seq , RNA-seq, whole genome genotyping, genome wide structural variation, *de novo* assembly, mutation detection, determination of inherited disorders and complicated diseases, DNA or RNA library preparation, single or paired ends and genomic captures, mitochondrial genome sequencing [53]. There are

various advanced projects about human genome using NGS technology summarized in Table 2.3 below.

Table 2.3 Some applications and studies of NGS in human genome

Application	Description	References
Mutation detection and carrier screening	Detection of structural variations in the human genome using NGS	[54]
	Functional genomic fingerprinting for various mutations detection	[54]
	Using NGS in disease mutation detection	[55]
Detection of inherited disorders in human	Detection of monogenic inherited disorders	[56]
	Genome-wide association studies (GWAS)	[57]
	Identification of causal variants of human disease	[58]
Cancer researches	Understanding cancer molecular mechanisms	[59]
RNA sequencing	MicroRNA expression profiling for diseases	[60], [61]
	RNA-seq application	[62]
	Functional genomics	[63]
	Annotation and mining of biological data	[64]
Library preparation	Paired-end sequencing in NGS	[65]
	Library preparation: Cap analysis of gene expression (CAGE) and serial analysis of gene expression (SAGE) for tissue specific genes	[66]
Genome wide structural variation detection in human population	1000 Genome Project	[41]

Bioinformatics and computational approaches are useful areas for cancer research. Identification of novel, rapid and easy diagnostic methods, development of new effective treatment methods for cancer patients can be achieved by computational methods. Using computational approaches and also artificial intelligence we can generate fast, effective, robust and dynamic methods for cancer therapies and we can make significant contributions to in vitro and in vivo experiments accelerating research and developments in cancer.

CHAPTER 3

HUMAN GENOME AND GENE TYPES

3.1 Human Genome

Genome is defined as complete set of genetic instructions of an organism. Each genome contains DNA material, which is information about building, development, biological processes of organisms. Lots of diseases like cancer, cardiovascular diseases and neurodegenerative diseases are associated with genome, and mostly the diseases are due to changes in genome like mutations, single nucleotide polymorphisms. Therefore, understanding of human genome is significant for resolving the genetic mechanisms of diseases and developing new treatments and diagnosis methods. In this chapter, we examine human genome, genetic material and genes according to their expression level particularly in the context of cancer.

3.1.1 Gene Definition

A nucleotide sequence that constitutes a specific and certain part of a chromosome is called gene. Gene may be a genomic sequence directly encoding functional product molecules such as proteins [67]. DNA is a nucleic acid sequence that carries the genetic instructions necessary for the vital functions and biological evolution of all organisms and some viruses. The main task of DNA, as the polymeric molecule shown in Figure 3.1, is storage of hereditary information. DNA consists of monomers or nucleotides containing a phosphate group, a sugar group and a nitrogen base which comes together to make a linear chain. There are four different monomers in DNA:

- the purine bases: Adenine (A) and Guanine (G),

- the pyrimidine bases: Thymine (T) and Cytosine (C).

DNA molecules are generally double-stranded and consist of two complementary strands in which A always matches with T and C always matches with G. This complementarity of the two strands provides the strict adherence to DNA molecule.

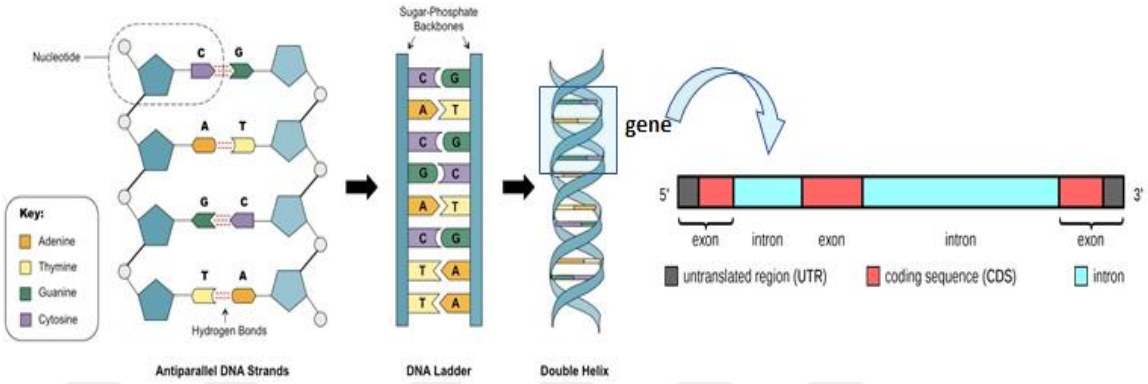


Figure 3.1 Molecular structure of DNA strain and a simple gene [68]

Hereditary information is stored in DNA in nucleus for eukaryotic cells and in cytoplasm for prokaryotic cells by means of the genetic code. DNA template is first converted to RNA through a process known as "transcription mechanism". Messenger RNA (mRNA) is then transported from the nucleus to the cytoplasm in eukaryotes, where the RNA sequence is translated to amino acid that is called "translation mechanism" to generate meaningful protein molecules, briefly. Protein synthesis occurs on ribosomes which are cytoplasmic organelles. Ribosomes are made up specialized type of RNA known as ribosomal RNA (rRNA) and many different structural proteins themselves. Translation process also involves transfer RNA (tRNA). tRNA provides the molecular link between coded mRNA sequences and amino acid sequence of newly produced protein on ribosomes [69]. While some genes are only a few kilobases (kb, 1 kb = 1000 base pairs) in length, others can be hundreds of kb. Protein coding sequences of genes are called "exons" and many genes can be interrupted by one or more noncoding regions are called "introns". Although introns are initially transcribed into RNA in the nucleus, they are not found in mature mRNA and subsequently in amino acid sequence of translated protein. Interestingly, despite their importance, exons constitute only 1.2% of the human genome. Nucleotide sequences provide the molecular "start" and "stop" signals for mRNA synthesis transcribed from exon. At the 5' end of the gene is a "promoter

region" which includes sequence responsible for the proper initiation of transcription [69] described briefly in Figure 3.2.

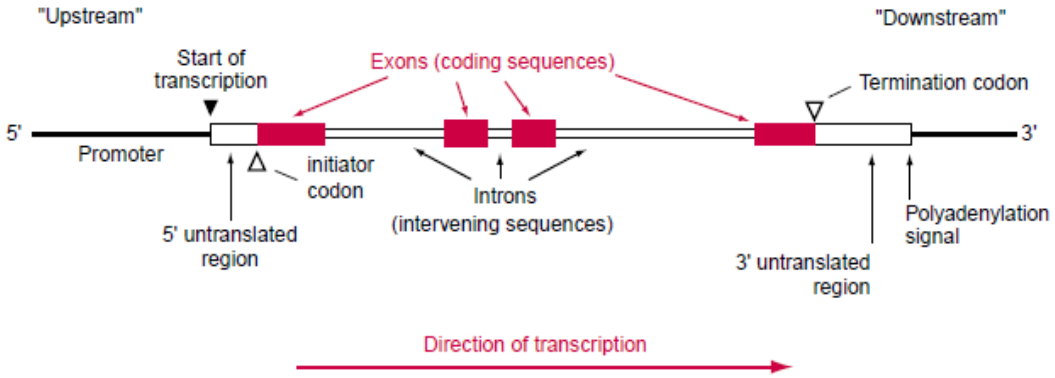


Figure 3.2 Protein coding gene structure [69]

Genes work in concert in multiple ways to generate products. Focusing on the functional products of genes is important to understand molecular mechanism and pathways. Thus, studying functional properties of genes is primary focus of genomic research [67], [69].

The ENCODE consortium completed its characterization of 1% of the human genome by various high-throughput experimental and computational techniques and identified functional elements of genome [70]. ENCODE project represents a major information for the characterization of the human gene functions, and the recently discovered findings reveal a complex molecular activity. There are approximately 21.000 protein-coding genes according to the latest estimation of ENCODE project. The project is at the center of attention of many researchers that investigate complexity of transcripts and their functional roles [67].

3.2 Gene Types

There are many gene classes according to their different properties. We give some information about gene types and classification of genes according to their expression levels in living cells or tissues. Classification of genes according to their expression level is significant for this thesis study, because we tried to discover specifically expressed genes in tissues using computational approaches, rigorously.

3.2.1 Pseudogenes

Pseudogenes are defined as genes that are related to functional genes but not translated to an active protein. Some researchers consider that pseudogenes are nonfunctional copies of functional genes like illustrated in Figure 3.3. Pseudogenes represent genes that were functional before however are not functional now since they were inactivated by mutations in protein coding genes. They are wide-spread in the human genome. For instance, disruption in signal bases that give rise to the start of transcription affect mRNA synthesis and consequently translation negatively, and expression of the gene cease to exist [69].

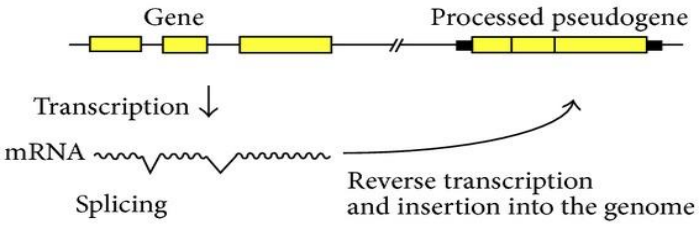


Figure 3.3 Formation of pseudogenes in genome [71]

3.2.2 Protein Coding Genes

Protein coding and non-protein coding genes are two main classes of genes due to their transcription processes. RNA sequences generally originate from protein coding genes in central dogma process. mRNAs are generated from protein coding genes as transcription step and then they are converted into protein chains in ribosomes during translation. Proteins are still assumed to be the main functional and structural players in the cells, tissues and organism even though noncoding RNAs have come to an importance place over the past ten years [72], [73]. Introns do not participate in mRNA and protein coding processes. Therefore, protein production occurs only in the exons. Despite the fact that protein coding genes are very important, there are new studies about non-protein coding genes and their functions and significance. Roles of introns in disorders are unclear and recent studies are interested in this issue [74]. Main structure of protein-coding genes was summarized using a simple illustration aa following figure, Figure 3.4.

Figure 3.4 Structure of protein-coding genes [75]

3.2.3 Non-Protein Coding Genes

Function of non-coding genes are becoming increasingly elucidated thanks to the development of NGS characterization methods and other improvements in gene annotations, bioinformatic analysis and new tools [76].

RNAs that are longer than 200 nucleotides are called long RNAs (lRNAs) and RNAs of whole cells that are less than 200 nucleotides are called short RNAs (sRNAs). There has been growing interest in studying functions of non-coding transcriptome because non-coding sequences might be associated with mechanism of human diseases and understanding of their roles in genome are crucial. Many of these RNAs have been classified into novel categories like long non-coding RNAs (lncRNAs), short non-coding RNAs, microRNAs, endogenous small interfering RNAs and PIWI-associated RNAs based on their function, sequence lengths, biogenesis, structural features, protein-binding partners and other features [76], [77].

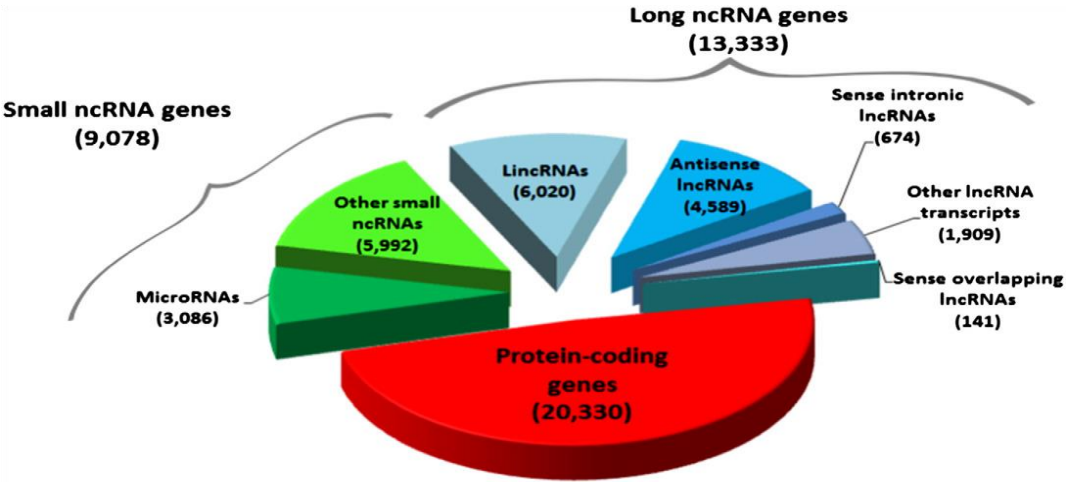


Figure 3.5 Protein coding and non-coding genes in the human genome. The numbers are based on Gencode V17 [78]

According to Figure 3.5, there are about 20.000-21.000 protein coding genes. Remaining genes is composed of non-protein coding RNAs which include long non-coding and short non-coding RNAs. Understanding the roles of long non-coding RNAs is important for the analysis of molecular mechanisms in cells. microRNAs, making up one third of small non-coding RNAs, have important role in the regulation of genes and associated with various mechanisms important for both healthy and disease conditions.

3.2.4 Long Non-Coding RNAs

Long non-coding RNAs (lncRNAs) have the largest portion of the mammalian non-coding transcriptome. The biological roles and significance of lncRNAs are contradictive. More lncRNAs have been identified day by day thanks to new comprehensive studies and development of bioinformatic methods. lncRNAs have been shown to fulfill a diverse range of regulatory roles. On the other hand, functions of the clear majority remains unknown and untested completely [77]. Therefore, exact prevalence of lncRNAs should be examined and interpreted.

3.2.5 Short Non-Coding RNAs an Example of Small Nucleolar RNA (snoRNA)

Small non-coding RNAs (sncRNA) have been shown to act as significant regulators for the expression of several genes in a variety of living organisms. SnoRNAs are a large family of sncRNA, generally ranging from 70 to 140 nucleotides as sequence length. SnoRNAs have vital role in ribosome biogenesis by serving as a guide for the site specific modification of rRNA [76].

3.2.6 Micro RNAs (miRNAs)

miRNAs are large family of small, endogenous and evolutionarily-conserved regulatory non-coding RNA, typically have sequence length of 21–23 nucleotides. They were found mostly in plants, animals and also human genome. miRNAs regulate the translation of up to 60% of protein coding genes, in general showing in Figure 3.6. Moreover, some protein coding genes are regulated by only a single miRNA, while others are regulated by many miRNAs. In addition, miRNAs serve as guides in RNA silencing and post-transcriptional regulation according to different studies [79]. There is a miRNA database

which includes human miRNAs that is called “miRbase” [80], total number of them is about 2500 in this database. miRNAs as fine-regulators of gene expression play role in different processes like cell differentiation, proliferation, development and also cell death [81].

miRNAs have become the center of interest in cancer studies either as oncogenes or tumor suppressor genes. They can be responsible for causing various cancers according to recent studies. There are 14 different online databases that address every aspect of miRNAs for cancer research. These databases focus on miRNAs and related particular cancer types, otherwise some other databases focus on the behavior of miRNAs in different malignancies at the same time [82]. Deregulated expression of miRNA in human cancers has been observed since 2002 [83].

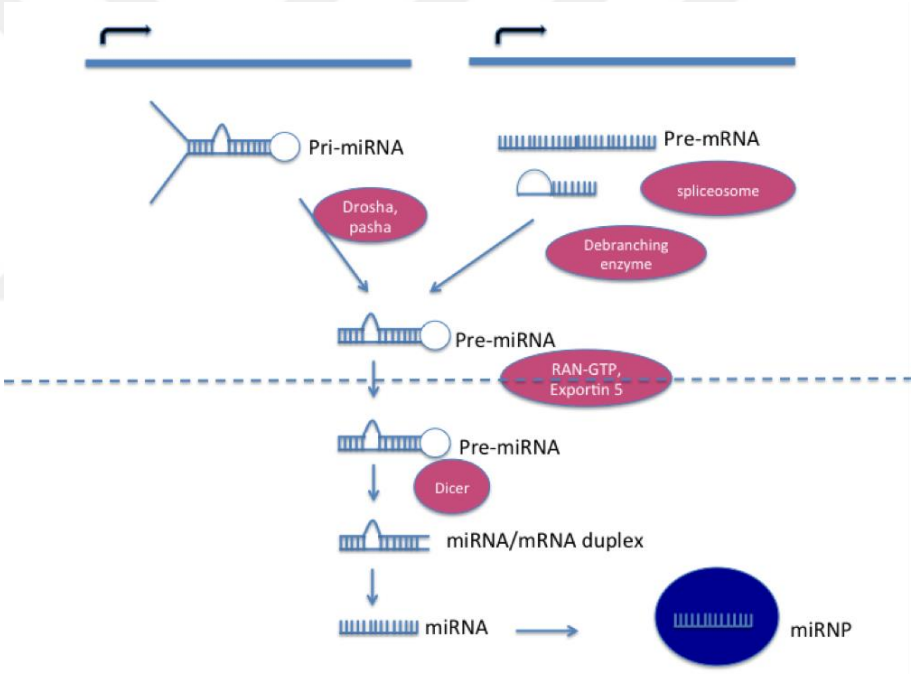


Figure 3.6 miRNA function and formation [84]

3.2.7 Transcription Factors

Transcription factors (TFs) are molecules that have significant role in regulating gene expression. TFs are usually proteins, or they can also consist of sncRNA. TFs can act individually or as a single group or as complexes over one gene. They can form multiple interactions that allow for varying degrees of control over rates of gene transcription mechanism. Lots of TFs have been identified which function in a wide variety of

processes. Some TFs are SRF, NF-KB, Fos and Jun, C/EBP, and related factors [85], [86]. Specifically, TFs have role in the pathogenesis of human cancers [87].

3.3 Classification of Genes According to Their Functions

Human genes can be examined according to their functions using computational tools. Table 3.1 below lists classification of genes, their functional properties and number of genes in each group. Data were obtained from study of D'Onofrio et al [88].

Table 3.1 Number of genes for human according to their functional properties

Functional properties	Number of genes
TRANSCRIPTION and TRANSLATION	
RNA processing and modification	497
Chromatin structure and dynamics	233
Translation, ribosomal structure and biogenesis	1203
Transcription mechanism	1046
Replication, recombination and repair	278
CELLULAR PROCESSES and SIGNALING	
Cell cycle and division control	245
Cell wall, membrane, envelope biogenesis	62
Cell motility	28
Post translational modifications and protein turnover	1427
Signal transduction	1723
Intracellular trafficking, secretion and transport	639
Defense mechanisms of cell	164
Extracellular matrix structure	272
Nuclear structure	18
Cytoskeleton	531
METABOLISM	
Energy production and conversion	336
Amino acids transportation and metabolism	367
Nucleotide transportation and metabolism	167
Carbohydrate transportation and metabolism	386
Lipid transportation and metabolism	93
Coenzyme transportation and metabolism	379
Inorganic ion transportation and metabolism	343

Table 3.1 Number of genes for human according to their functional properties (cont'd)

Secondary metabolites biosynthesis, transport and metabolism	217
POORLY CHARACTERIZED	
Only general function prediction	3271
Unknown functions	1187

3.4 Classification of Genes According to Their Expression Level

Different transcripts are expressed in diverse organs, tissues or cell types, as well as in different developmental stages or diseases. With the development of NGS technology, researchers can measure simply gene expression in various tissues at genome-wide scale. Cancer pathology and prognosis caused by defects in human transcripts are usually related to tissue-specific transcripts [89]. If we examine studies about gene types according to expression levels, genes can be classified as marker, specific or selective and housekeeping according to their expression in various tissues or cells in an organism. There are some explanations about these gene types according to their expression level in the tissues. Researchers established the FANTOM [90] and H-Invitational [86] consortiums which annotated comprehensive full-length cDNA collections in mouse and human, respectively [91].

3.4.1 Marker Genes

There are some genes which are used for cell identification that are called marker genes. Marker genes can be used to characterize cell types *in vitro* such as membrane specific markers. They have great importance for determining tissue or cell identity, and for understanding tissue-specific gene functions and the molecular mechanisms underlying complex diseases. Furthermore, marker genes of healthy tissues could be used to understand the molecular mechanisms and biological processes that are necessary for the illumination of developmental stages of all living organisms [92].

3.4.2 Specific Genes

There are more than 25.000 genes in the human genome, and they demonstrate dramatic diversity in terms of expression levels and tissue expression patterns [93].

Tissue-specific genes are a group of genes whose functions and expressions are preferred in one or several tissues or cell types [94]. Tissue or cell-type specificity of gene expression is central point and it is crucial approach to human biology and biomedicine. Genes with tissue-specific expression play significant roles in the physiology of multicellular organisms and associate with human diseases [95]. In general, about 100 to 200 signature genes are expressed in a specific tissue. Tissue-specific transcripts can indicate novel functions of known and unknown genes. Tissue specificity is also associated with many significant results including expression-quantitative trait loci [96] evolution [97] and various diseases [89], [95], [98] which are most important of all in this thesis study. Additionally, tissue-specific pattern of gene expression can help solving and identifying the molecular mechanisms and developmental processes of tissues, gene functions, transcriptional regulation of biological processes, prognosis of diseases, etiology and discovery of novel tissue-specific drug targets [99]. By the way, tissue-specific pathways related with tissue-specific expression for each gene in cell, both known or unknown, may be disrupted in disease. Thanks to these information in literature, we can understand that the expression of tissue-specific genes can also be used as an indicator for many complex diseases like solid tumors, in particular. It impacts fundamental problems such as tissue ontogenesis and carcinogenesis. Many related biomarkers and targeted molecules for drugs are discovered by means of tissue-specific expression [95]. In this context, there have been many studies that examine and try to find out tissue-specific expression and their relationship with cancer and also other diseases [95], [100], [101], [102], [103]. Because of importance of tissue- specific genes and tissue-specific networks, many databases have developed to identify and store information of tissue-specific genes and their relations, regulations and disease association in a variety of human. As a summary, tissue-specific expression of genes can be associated with pathogenic mechanism, diagnosis, and therapeutic applications to discover tissue-specific and effective therapeutic targets. Therefore, determination, examination and observation of specific genes in each human tissue can provide inside about cancer research.

3.4.3 Selective Genes

The selective expression of genes suggests their possible roles of molecular functions. Thus, they may be potential drug targets or diagnostic indicator for possible tissues [104]. There is a confusion among the gene definitions in here. Actually, some researchers have identified specific genes as selective genes. On the other hand, some researchers have also identified specific genes as marker genes. In this thesis study, we used "marker genes" and "specific genes" terms with following definition, if gene is expressed only one tissue or cell type it is called marker gene and if gene is expressed in one or several tissues or cell types it is called specific gene.

3.4.4 Housekeeping Genes

Housekeeping (HK) genes are expressed in almost every tissue to maintain basal cellular functions or cell viability such as developmental stage and cell cycle under normal conditions. HK genes are described as universally expressed genes in all tissues. They have higher centrality properties and may play important roles in the comprehensive biological networks [104], [105].

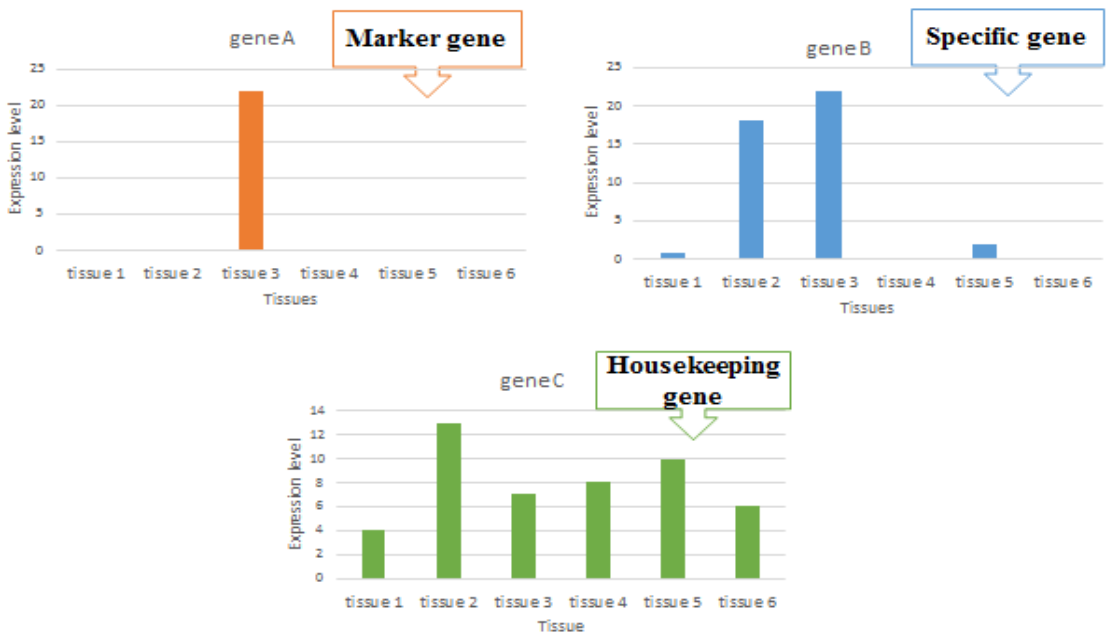


Figure 3.7 Gene types according to their expression level

Figure 3.7 illustrates the marker, specific and housekeeping genes with hypothetical expression level plots. A gene expressed only in one cell type or tissue is called "marker

gene". Secondly, genes expressed in one or several cell types or tissues specifically and significantly are called "specific gene". Other definition is "predominantly expressed in one or a few biologically relevant tissue types" in literature [100]. Thirdly, genes expressed in all cell types or tissues in organisms are called "housekeeping gene."

3.5 Human Tissue Structure

Tissues are collection of cells with a common structure including basic layers and functions. There are four main tissues in the human body. These are epithelium, muscle, connective tissue and nervous tissue. Solid organs are formed by the cellular complexity and structural integrity of tissues and they are also comprehensive and complex structures in human body. Structure, functions, histogenesis, morphogenesis, organogenesis and roles of normal human tissues and organs are very crucial for the understanding of human biology and diseases [106]. Formation of tissues possesses multiple components such as various cell types, extracellular matrix, scaffold and complex geometric information of all. Hence, tissues are multilayered structures. For instance, liver tissue is multilayered and has heterogeneous structure in an organ microstructure and chemical composition because it is highly vascularized organ [107]. Organs are made up of two or more different tissues organized to carry out specific functions in the corresponding organ. Therefore, organs have several different tissue layers. Moreover, organs and tissues also have various cell types. This structure is called **"heterogeneous"**. The heterogeneous structure of stomach is given below as an example in Figure 3.8.

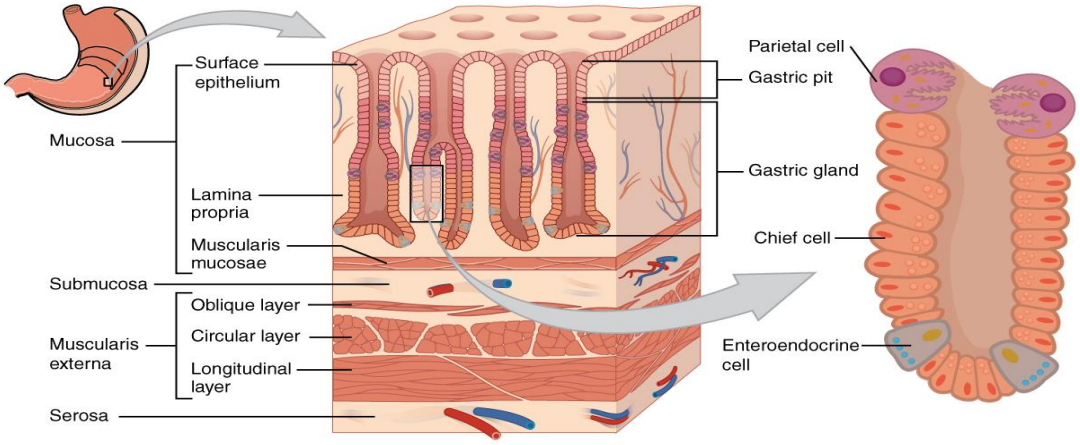


Figure 3.8 An example of tissue heterogeneity. Stomach tissue has several layers and lots of cell types [108]

3.6 Determination of Tissue Specific Genes via Computational Approaches

In this chapter, we discussed that tissue-specific genes play important roles in developmental stages of healthy tissues and formation of many diseases at molecular levels and in context of pathways. Besides, importance of tissue-specific genes and their roles were emphasized in Section 3.4.2. Thus, identification of tissue-specific genes is a necessity for cancer research. However, examining all of genes in laboratory takes a lot of time with high cost and it is possible that experimental errors can affect results seriously. If we calculate tissue-specific genes using bioinformatic approaches we can determine genes expressed specifically for one or several tissues or cell types for human quickly. After that we can validate our results using experimental approaches. For these purposes, we try to calculate and obtain a list of tissue-specific genes using computational approaches to accelerate research in this area. In the thesis, we have concentrated on the calculation of the tissue-specificity of expression as a first aim.

There are several methods for the classification of tissue-specific genes, each of them differs according to their assumptions, scale, organism and also related gene types. These methods can be divided into two main groups. First group summarizes in a single number whether a gene is tissue-specific or wide-expressed. These methods contain calculation of some scores such as tau, Gini, Tissue Similarity Index (TSI), entropy (Hg), and count. The second group calculates, for each tissue separately, how specific the gene is to that tissue. These methods are called z-score, Specificity Measure (SPM), Expression Enrichment (EE) and Preferential Expression Measure (PEM) [38]. In addition to these, there are some tools which using methods mentioned above via different experiments like microarray, EST data or sequencing. These methods and tools are explained in this section. All scores and calculation methods indicate in how many tissues a gene is expressed specifically, and whether it has large differences of expression level among genes in each tissue.

Tau, Gini, TSI, Counts and Hg are poor scores that calculates the specificity per gene per tissue and this provides coarse results and are not robust and rigorous. The second group (z-score, SPM, EE and PEM) make efforts in broad and weak expression conditions. Different methods can conflict among each other to obtain tissue-specific gene lists for each tissue or organ. Thus, there are some disadvantages and deficiencies in each

method. Tau appears consistently to be the most robust method according to Kryuchkova-Mostacci and Robinson-Rechavi [38] and more significant than other methods.

Even if tissue specificity is often used in various studies and it is very significant, there is no clear robust and rigorous method until now. Moreover, several databases were developed to establish a knowledge base of large-scale data regarding tissue-specific gene expression in a variety of human tissues. These are:

- TiSGeD: A database for tissue-specific genes [94]
- TiGER: Tissue-specific gene expression and regulation [109]
- PaGenBase: A pattern gene database for the global and dynamic understanding of gene function [110]
- GENEVESTIGATOR: The world's expression database [111]

However, there is no correlation among their results. Tissue specificity is of interest to molecular evolution, as well as a functional feature which may evolve itself in tissue developmental stages. While many different methods for measuring specific expression of genes have been proposed and used for many studies, there is not a strong correlation among them because of using different data types and experimental methods. Most methods were established for ESTs and microarrays, and several methods have used RNA-seq data [38]. In recent years, most expression studies utilize RNA-seq due to its advantages, which provides more reliable and accurate results compared to microarray and ESTs.

3.6.1 Different Methods and Tools for Tissue Specificity Calculation

Even if tissue-specific expression is often used in several studies, there is no clear explanation for why one method is more accurate and detailed over another. There are several methods to measure tissue-specific genes and some tools to produce tissue-specific gene lists. In this section we mentioned about tissue specificity calculation scores and also bioinformatic tools.

3.6.1.1 Expression Enrichment (EE)

Expression enrichment method was used in TiGER database [93], [109]. TiGER contains three types of data, these are tissue-specific gene expression profiles, combinational gene regulations, and cis-regulatory module (CRM) detections. ESTs data were used when TiGER was established. The database include 7261 tissue-specific genes in 30 different human tissues based on the expression enrichment (EE) calculation [109]. EE can be calculated as follows:

$$\mathrm{EE} = \frac{\mathrm{xi}}{\sum_{i=1}^{n} \mathrm{xi} * \frac{\mathrm{Si}}{\sum_{i=1}^{n} \mathrm{Si}}} = \frac{\sum_{i=1}^{n} \mathrm{Si}}{\mathrm{Si}} * \frac{\mathrm{xi}}{\sum_{i=1}^{n} \mathrm{xi}} \tag{3.1}$$

- xi is defined as expression of a gene in tissue i
- n is number of tissues
- Si is average of the expression of all genes in tissue i

This database is outdated because of used data type and calculation method compared to newly developed technologies like NGS.

3.6.1.2 Tissue Similarity Index (TSI)

Tissue similarity index is a calculation method ranging from 0 to 1. The meaning of 0 is that there is no expression in related tissue. Otherwise, the meaning of 1 is that gene is expressed only in one tissue. If TSI score is equal and bigger than 0.75, gene is specific to corresponding tissue. TSI had revealed to investigate pharmaceutical targets, and researchers believe that it will become an important tool for experimental and biomedical research, however it is not widely used according to literature. TSI can be calculated as follows:

$$\mathrm{TSI} = \frac{\max_{1 \le i \le n} \mathrm{xi}}{\sum_{i=1}^{n} \mathrm{xi}} \tag{3.2}$$

- xi is defined as expression of the gene in tissue i
- n is number of tissues

3.6.1.3 Shannon Entropy (Hg)

Shannon entropy can be used for the calculation of tissue-specific genes as a computational approach. Schug et al. [112] used firstly Shannon entropy for the determination of tissue specificity of human genes according to their expression in multicellular processes. They used both microarray and ESTs data for human genes, and mouse genes were used for validation in their study. Entropy measures the degree of overall tissue specificity of a gene, on the other hand, it does not demonstrate that it is specific to a particular tissue or not. Hg has been used by some other studies to identify potential drug targets, variation of expression levels and clustering of microarray data. Briefly, Hg is used to indicate tissue specificity with transcriptome diversity as entropy of its frequency distribution [112], [113].

Hg specificity index are based on the adaptation of Shannon's information theory [114] to the transcriptome data:

$$\mathrm{Hg} = -\sum_{i=1}^{n} \mathrm{pi} * \log_2 \mathrm{pi} \tag{3.3}$$

$$\mathrm{pi} = \frac{\mathrm{xi}}{\sum_{i=1}^{n} \mathrm{xi}} \tag{3.4}$$

- xi is defined as expression of the gene in tissue i
- n is number of tissues

3.6.1.4 Z-score

Z-score is a statistical parameter that is number of standard deviations from the mean of a variable as a simple definition. It is often used in bioinformatic analysis, for instance microarray data analysis [115]. Z-score was used for specific expression of genes and according to Z-score either only overexpressed genes might be defined as tissue-specific or the absolute distance from the mean is used for tissue specificity. Hence, under-expressed genes are also defined as tissue-specific via Z-score calculation. Z-score generates imprecise or undetailed results, which is a major drawback. For instance, when we examined Z-score calculation for tissue-specificity in the context of our thesis study, we observed that, Z-score sometimes concluded a gene is specific although it has

housekeeping gene pattern. Thus, Z-score is not reliable for tissue specificity [38], [116]. Z-score can be calculated as follows:

$$z = \frac{xi-\mu}{\sigma} \tag{3.5}$$

- μ is the mean of gene expression
- σ is the standard deviation

3.6.1.5 Specificity Measure (SPM)

Specificity measure is used in the database TiSGeD with microarray data. SPM ranges from 0 to 1. If a SPM value is closer to 1, this gene has high tissue specificity. In practice, we can rely on the SPM value to quantitatively estimate tissue specificity of genes. However, gene profiles are highly selective in several similar tissues, SPM value may not give accurate results, in this case it is not successful for the calculation of tissue specificity [94]. For SPM score, each value is indicated by the sum of squared gene expression for all tissues:

$$SPM = \frac{xi^2}{\sum_{i=1}^{n} xi^2} \tag{3.6}$$

- xi is defined as expression of the gene in tissue i
- n is number of tissues

3.6.1.6 Preferential Expression Measure (PEM)

PEM was suggested by Huminiecki et al. using ESTs data [117]. The PEM score is a simple form of the EE score. Because of the fact that these scores are normalized by either maximum expression of a gene or by sum of expression of a gene, two of them are not sensitive to its absolute expression level [38]. PEM score was used in different databases like SAGEmap, dbEST, Gene Expression Atlas and TissueInfo [117]. PEM can be calculated as follows:

$$PEM = \log_{10}\left(\frac{\sum_{i=1}^{n} Si}{Si} * \frac{xi}{\sum_{i=1}^{n} xi}\right) \tag{3.7}$$

- xi is defined as expression of the gene in tissue i

- n is number of tissues
- Si is average of the expression of all genes in tissue i

3.6.1.7 Gini Coefficient

Gini is widely used in economics for measurement of inequality [38].

$$Gini = \frac{n+1}{n} - \frac{2\sum_{i=1}^{n}(n+1-i)xi}{n\sum_{i=1}^{n}xi} \quad (3.8)$$

- xi is defined as expression of the gene in tissue i
- xi has to be ordered from least to greatest
- n is the number of tissues

3.6.1.8 Tau Score

Tissue specificity index, tau (τ) specificity score was defined in 2004 by Yanai et al. as a gene characterization score. Tau varies from 0 to 1. τ is equal to 0 means this gene is housekeeping, and 1 means this gene is strictly expressed in a specific tissue with midrange profiles having $0.15 < \tau < 0.85$ [118]. τ is a quantitative, graded scalar measurement of specificity of a gene expression. Yanai et al. described that tau score is a robust method by giving an example from Su et al. [119]. When they calculated tissue specificity using τ, they had found a high correlation between τ index of genes between two different datasets. Hence, they showed strength of τ index.

Tau score has been used in various studies for the measurement of tissue specificity of genes in various tissues [38], [118], [120], [121], [122], [123], [124]. As we can see, there are lots of studies that used tau score for the calculation of tissue-specific genes. However, some limitations exist in this method. For instance, tau gives single number between 0 and 1, then user assumes that the tissue in which gene shows maximum expression is the tissue of specific expression. So, tau can assign specific expression of a gene to a single tissue only, not multiple tissues. It is obvious that there are many genes which are specifically expressed in several tissues or cell types in multicellular organisms.

Tau specificity score is defined as follows,

$$\tau = \frac{\sum_{i=1}^{n}(1-Xi)}{n-1} \tag{3.9}$$

$$Xi = \frac{xi}{\max\limits_{1\leq i\leq n} xi} \tag{3.10}$$

- xi is defined as expression of the gene in tissue i
- n is number of tissues

In this section, it is showed that there are many specificity scores as calculational methods. Kryuchkova-Mostacci and Robinson-Rechavi did a study comparing different tissue specificity calculation methods and their reliability. There are two main groups as mentioned in Section 3.6. As the benchmark of these two groups, the first group which contains tau, gini, TSI and Hg is more successful than second group. First group gives only a single score that belongs to a single tissue. Otherwise, second group gives scores as many as the number of tissues and we have to decide which gene is specific to which tissue using threshold value. There are some disadvantages and some deficiencies of all methods. However, according to Kryuchkova-Mostacci and Robinson-Rechavi the best specificity score is tau score [38]. Even though it is an effective method and giving the accurate results, it should be improved by additional statistical procedures.

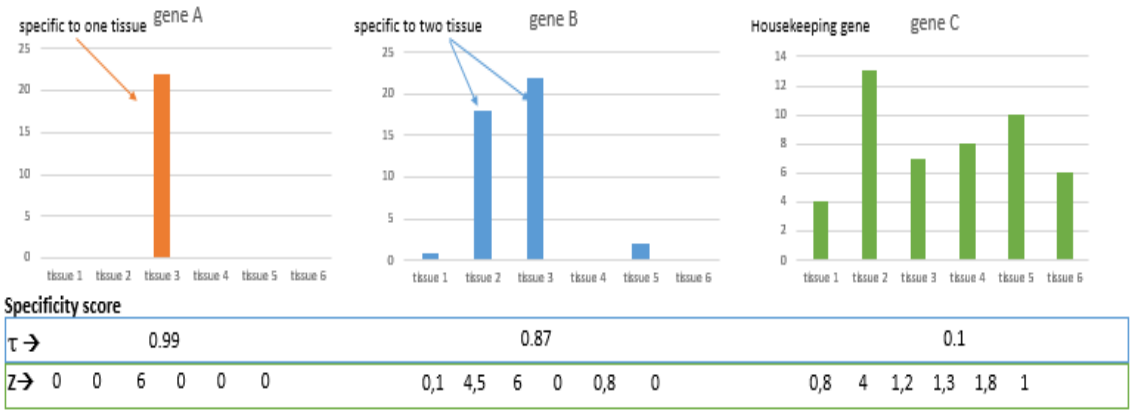


Figure 3.9 Examples of tissue specificity score calculations (one from each group to illustrate disadvantages of them)

According to Figure 3.9, disadvantages of each scoring scheme can be seen. Firstly, tau included in the first group gives only one score to each gene for 6 tissues in three examples. In the first example, tau gives one specificity score (0.99) for one single tissue

that is specifically expressed in related tissue. This is the expected correct result. In the second example, tau points out only one tissue (τ=0.87, for tissue3). In reality, this gene is specific to two tissues (tissue2 and tissue3). There is a deficiency in the determination of related tissues. Last gene is housekeeping and its tau score is very low (0.1) as expected. Tau score is robust, but not rigorous method for tissue specificity.

Secondly, Z-score included in the second group of methods has a threshold for specificity. If Z-score is equal and greater than 3 for each tissue, a single gene is specific to corresponding tissues like in Figure 3.9. Gene A is specific to tissue3 (z-score=6), Gene B is specific to tissue2 and tissue3 (z-score=4.5 and 6, respectively). These are better results. However, z- score indicates that the last gene is specific to tissue2 (z-score=4) actually, it is housekeeping gene because of its wide expression in all tissues. Z-score does not give precise results, it is not reliable. Briefly, there are some deficiencies in each group of methods. Therefore, the best method according to literature, the tau score, is selected, and was improved to produce more precise results in this thesis.

3.6.1.9 Other Computational Approaches and Tools for Tissue Specificity

Other methods and bioinformatic tools for the estimation of tissue-specific genes can be summarized as follows:

CellMapper: It is a sensitive approach to predict genes expressed selectively in single cell type and it is improved to reveal cell-specific markers. CellMapper can make accurate estimations for only human brain cell types and it should be rapidly applied to diverse cell types from many tissues [125]. This case is insufficiency for CellMapper and it will be developed.

Digital Sorting Algorithm (DSA): DSA is a specific and sensitive algorithm for extracting cell-type specific gene expression profiles and used deconvolution processes of gene expression. It is an algorithm developed for only cell type-specific genes. It does not require prior knowledge of cell type frequencies [126].

Marker Gene Finder in Microarray data (MGFM): This new bioinformatic tool was developed to predict marker genes from microarray data. It is a useful tool to estimate tissue or cell type marker genes [92]. However, it is limited to only microarrays whereas

RNA-seq have a lot of advantages compared to microarray in the context of speed, cost and accuracy.

A Pattern Gene Database (PaGenBase): PaGenBase is a freely accessible database giving information about specific genes, selective genes, housekeeping genes and also repressed genes. SPM measurement was used in PaGenBase. If we examine this database, we can see that there is no correlation among data [110].

ROKU: ROKU was developed by Kadota et al. [127] in 2006 that is a tool for selection of tissue-specific patterns from expression data. Microarray data were used for many tissues and thousands of genes. ROKU is based on Shannon entropy.

In this chapter, some features of human genome particularly gene types according to expression levels were summarized. Tissue-specific genes are crucial to understand human biology and to solve biological problems. Therefore, identification of these genes is a significant requirement. Computational approaches can be used for this purpose. Here, we have to choose a robust and rigorous method as broad sense. According to our literature review, although there are many different methods that serve the same purpose, there is no correlation among them. This situation can be due to using different types of data. It is also of interest that the RNA-seq data have not yet been widely used for tissue specificity calculation. When we consider all of them, we decided to generate a new method for tissue specificity using RNA-seq data.

CHAPTER 4

CANCER TYPES AND TUMORAL HETEROGENITY

4.1 Definition of Cancer

Cancer is a multi-step process in which cells undergo metabolic and behavioral changes leading to excessive, increased proliferation and migration. Uncontrolled cell proliferation causes fatal processes in body such as epithelial-mesenchymal transition, migration from origin site of cancerous cells to other part of body via circulatory and lymph system, invasion and then metastasis resulting in cancer. At a molecular level, cancer is a genetic disease and it typically involves changes in gene expression or function due to various mutations. Any cancer causing genetic alteration typically results in uncontrolled cell growth [128].

Cancers are named according to the tissues and organs in which they originate from. Symptoms, signs and treatments also vary according to the cancer types and subtypes. The most common type of cancer in the world are lung, breast and colorectal cancers. Lung cancer and its subtypes are by far the leading cause of cancer related deaths because its prognosis is very low [129]. Early diagnosis is very important in an effective cancer control. Next important factor is targeted treatment which is popular for cancer therapy because of specificity provided for individual patients. A good diagnosis have to classify the type of cancer for each patient, cured by some targeted molecules, understand tumor surface markers, describe morphology of cancer [130], [131].

Cancer is a disease that develops due to multiple effects rather than only one single cause. Risk factors for cancer vary depending on the individual's lifestyle, age, sex, and family history because of hereditary information. There are also environmental factors

such as smoking, alcohol usage, prolonged exposure to ultraviolet light, excessive exposure to X-ray radiation, and some chemicals (tar, gasoline, dyes, asbestos, etc.), some viruses, air pollution, bad nutrition habits [132], [133], [134]. Free radicals or oxidants are on the rise as a factor playing role in human carcinogenesis. Oxidants play an important role in the development and progression of cancer. It is thought that most types of cancer are the result of reactions between free radicals and DNA molecule, resulting in mutations that adversely affect the cell cycle and it potentially leads to cancer malignancy [135], [136].

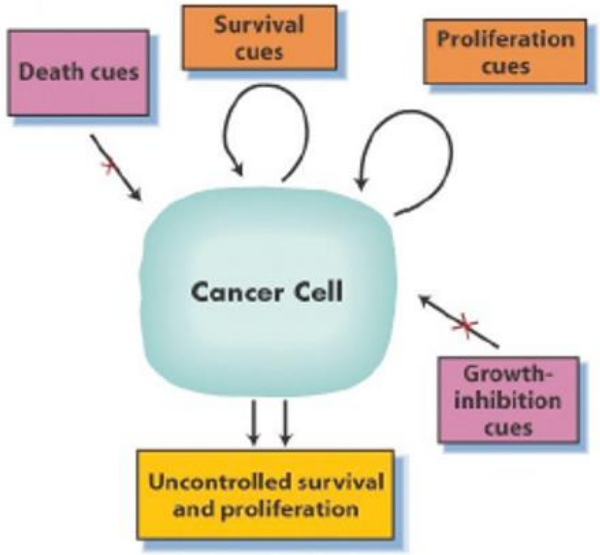


Figure 4.1 Simple definition of cancer and basic properties of cancer cells [130]

Tumors basically are divided into two groups; benign and malignant. Benign tumors cannot spread by invasion or metastasis; hence, they only grow locally, and their treatment is easier compared to the treatment of malignant tumors. Malignant tumors can spread by invasion and metastasis via cancerous cell migration amd cancer cell main properties are defined in Figure 4.1. In normal tissues, rates of new cell growth and cell death is approximately equal. In cancer, this balance is disrupted. This disruption can result from uncontrolled cell growth [131].

When a normal cell is converted into a cancer cell, multiple changes occur during the conversion processes in organism. Some properties of cancer cells:

- Exhibit several characteristics that are far from normal cells.
- Autocrine stimulation; grow hardly in lack of growth factors,

- Lack of gap junctions and contact inhibition
- Resistance to cell death and persistent telomerase activity
- Migrate to other tissues, growing rapidly, overtake population and invasion
- Angiogenesis as a fatal process
- Genomic instability, accumulation of a lot of mutations in genome that perturb gene functions [130], [131].

4.2 Cancer Incidence

Cancer is one of the most important social health problems in Turkey as well as all over the world. Moreover, cancer is the second most common cause of death in the world. Until 2030, it is predicted to increase rapidly and settle in the first place as leading cause of death. When the 2010 data is evaluated; over 159,000 new cases of cancer patients have been recorded in Turkey within a year. Frequency of incidence is rapidly increasing over time, both locally in Turkey and globally. Complete and effective control of this disease will only be possible with a scientific, multidisciplinary and cost-effective program. We can see in Figure 4.2, rapidly increased incidence of cancer for both male and female:

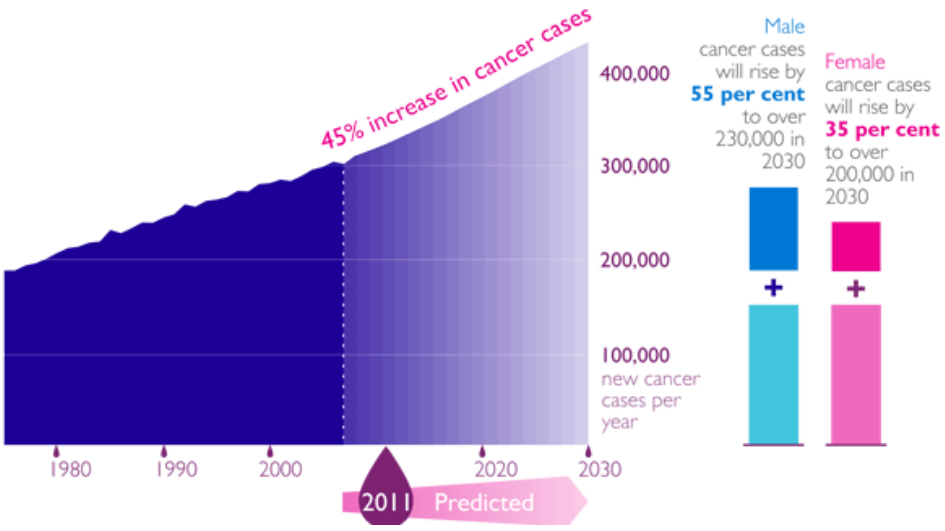


Figure 4.2 Rapidly increased incidence of cancer [137]

4.3 Cancer Treatment

The main traditional therapy used in cancer patients is a series of interventions, including psychosocial support, surgery, radiotherapy, chemotherapy and immunotherapy that is aimed at curing the cancer, prolonging life and increasing patients' quality of life. Different cures can be used together when necessary. Tumor type, histopathology, progresses or stage and performance status of the patient are important for effective therapy. Some of the most common cancer types, such as breast, cervical, oral cancer and colorectal cancer have higher cure rates when detected early and treated easily compared to other cancer types. Some cancer types such as leukemias and lymphomas in children, and testis cancer have high cure rates if appropriate treatment is provided [131], [138], [139].

Surgery: Surgery may be appropriate in early stages of cancer however, getting tumorous tissue is often insufficient for treatment [140].

Chemotherapy: Chemotherapy is a form of treatment with natural or synthetic chemical, biological agents and hormones, with selective lethal effects against non-controlled proliferating cells in cancer patient. It is used frequently; however, it has many side effects to human body. Chemotherapy is sometimes applied before surgery. This practice is called "neoadjuvant therapy". The aim of this is to make it easier to remove by shrinking the large cancer tissue. On the other side, it can also be given after surgery. This practice is called "adjuvant therapy". The goal is to remove the cancer cells that are too small to be seen. Chemotherapy is defined as "chemoradiotherapy" when it is given in conjunction with radiotherapy [141], [142].

Radiotherapy: Radiotherapy is a long-acting method that kills cancer cells. The positive effects of radiotherapy are transient and are usually limited to the area of radiation. In addition, radiotherapy can cause some adverse side effects on the healthy tissues of patients' body. It uses high energy gamma rays to kill cancer cells and shrink tumors. Gamma radiation can cause single or double chain breaks, mutations, and chromosomal abnormalities in DNA strands. Besides, it can also cause inactivation by providing oxidation of organic molecules such as enzymes and proteins, or by causing damage to their chemical chains [143].

Immunotherapy: Immunotherapy is especially suitable for the treatment of metastatic and recurrent tumors. Efficacy of therapy depends on the immunological tolerance of cancer patients. Cancer-specific T cells are found around the immunosuppressive tumor. However, these studies have not given very good results yet and researchers pursue this option as shown in literature [144].

If we review the literature about cancer treatment, we can see that available therapy methods are insufficient or have several deficiencies. This is because cancer is a very complex disorder and have effect on broad spectrum of molecular processes.

4.4 Cancer Types in The Context of Thesis Study

Cancers are named according to the tissues, organs or systems in which they originate from. Symptoms, markers and treatments also vary according to the cancer types and subtypes.

4.4.1 Breast Cancer

Breast cancer is the most common malign cancer among women in the world and it is an important health problem that accounts for about 30% of all cancers in women. Fortunately, breast cancer deaths have been decreasing because of a combination of increased awareness, early detection through screening, and modern improvement in cancer therapy day by day. Typically, early-stage breast cancer is treated with breast conservation therapy (BCT) contains lumpectomy and axillary staging with sentinel lymph node biopsy. Breast cancer is an epithelial-based adenocarcinoma most commonly located at the junction of the lobule and the terminal ductus [145], [146].

4.4.2 Cervical Cancer

Cervical cancer is one of the major health problem and important cancer type, which caused 4,020 women deaths in United States in 2014 [147]. The high mortality rate of cervical cancer can be reduced by applying the integrated strategy like combination of prevention, screening and treatment of the disease. Epidemiological research shows that certain types of human papillomaviruses are the central cause of cervical cancer [148], [149].

4.4.3 Esophageal Cancer

Esophageal cancers are divided into two main subtypes which are squamous cell carcinoma (SCC) and adenocarcinoma (ADCA) as histological observation. Primary site of esophageal cancer is often identified as a 'non-target lesion'. When stratified by anatomical location, the incidence of adenocarcinoma of esophagus and gastroesophageal junction (GEJ) continues to increase rapidly due to Barrett's esophagus. Development of chemoradiotherapy can provide effective treatment of esophageal cancer [150], [151].

4.4.4 Kidney Cancer

Adrenal gland has usually cortical adenomas without sectoral features and they are diagnosed incidentally as a result of radiological examinations. Kidney cancer represents about 5% of all cancer cases in overall. Subtypes of kidney cancer contains clear cell renal cell carcinoma, papillary tumors, chromophobe renal cell carcinoma, papillary type I renal cell carcinoma and papillary type II renal cell carcinoma. The most common form of kidney cancer develops from renal epithelium that is called renal cell carcinoma (RCC). Specifically, the most frequent RCC subtypes involve clear cell, papillary and chromophobe tumors. The therapeutic strategy for advanced kidney cancer subtypes is based on cytokines. There is some remarkable improvement in treatment outcomes and several targeted agents, thus half of the patients who has advanced kidney cancer are likely to survive more than two years [152], [153] .

4.4.5 Liver Cancer

Liver cancer ranks as the fourth cancer type among the most common malignancies, globally. Hepatocellular carcinoma (HCC) is one of the most common malignant tumor subtypes of liver cancers with a poor prognosis worldwide. Although early stage can be effectively treated with curative approaches which include liver transplantation or surgery, these are only suitable for a small number of patients. Advanced stage of liver cancer is resistant to the conventional chemotherapy or radiotherapy. Liver cancer generally repeats and relapse after therapy, and drug resistance is the critical for poor

prognosis. Thus, it is a type of cancer that is important for survival and life quality of patients [154], [155].

4.4.6 Lung Cancer

Lung cancer is the leading cause of cancer deaths and it has the highest cancer-associated mortality rate globally. It is a fatal disease that affects both men and women all over the world because premature mortality occurs when cells are out of control in the lung. The cancerous tissue formed here grows primarily in the original environment, and in later stages it can cause damage to the surrounding tissues or to the remote organs such as liver, bone, brain by circulation. Typical risk factors include smoking and exposure to arsenic, chromium, radon and air pollution. Therefore, multifactor and multistage processes, with genetic instability are considered to be the key cause of lung cancers. Histological types of lung cancer are non-small cell lung cancer (NSCLC), small cell lung cancer (SCLC, primary lung cancer) and other malignant tumors. Adenocarcinoma is the most common type of lung cancer in women, while squamous cell type is the most common type of lung cancer in men. All of them have different metastatic features and their treatment strategy must be different from each other [156], [157].

4.4.7 Pancreatic Cancer

Pancreatic cancer is the 10th common cancer type according to 2014 statistics and fall into 3% of all cancers. Pancreatic cancer may start in the head, body or tail of the pancreas. There are different types of cells in the pancreas because of heterogeneity of tumor. Knowing the type of cell and where in the pancreas it starts helps deciding about which treatment is suitable [158]. Pancreatic ductal adenocarcinoma (PDAC) of pancreatic cancer that is also called exocrine tumor is common type that is the most lethal human malignancy with the worst 5-year overall survival compared to other types of cancer. Pancreatic adenocarcinoma can be thought to be sporadic, however approximately 5% to 10% of it occurs in the presence of a family history of the disease, so genetic information is important for its development. If pancreatic cancer hasn't spread outside the pancreas tissue or there is no metastasis, surgery is possible and

suitable for treatment whereas advanced stage of it is very fatal because the median survival time is about 6 to 11 months for patients [158], [159].

4.4.8 Prostate Cancer

Prostate cancer is the most common type of cancer that typically affects men over 50 years of age in worldwide [160]. Molecular mechanism of prostate cancer is discovered to a limited content. This is due to the cellular heterogeneity of the prostate tissues and a lack of genetic information, related genes, pathways and markers from systematic analysis. Prostate cancer-specific genes or other molecules like enzymes can be used as markers for screening, diagnosis, prognosis, therapeutic monitoring and early detection of cancer [161]. For this reason, Myers et al. explained that there are several studies about gene and protein expression differences between normal and malignant prostate tissues and their usage in putative diagnosis, prognosis for the purpose of obtaining excellent biomarkers [162].

4.4.9 Rectum Cancer

Colorectal cancer covers different tissues which are colon, rectum or appendix while rectal cancer is the growth of abnormal cancerous cells in the rectum which is the lower part of the colon that connects the anus to the large bowel. Rectal cancer develops slowly over years. Its risk factors include increasing age, smoking, hereditary information or family history, high-fat diet and a history of polyps or inflammatory bowel disease in patients' life. After surgery, rectal cancer may relapse locally [163]. Nowadays, gene expression profiling has shown great promise in diagnosis as well as targeting rectal cancers like other cancers.

4.4.10 Stomach Cancer

Stomach cancer, also known as gastric cancer, starts in the stomach. Stomach cancers tend to improve slowly and quietly over many years. Pre-cancerous changes are often observed in the inner layer mucosa of the stomach after which cancer can merge. These early changes can show us symptoms and therefore its detection can be easy. There are various types of stomach cancer including adenocarcinoma, lymphoma, gastrointestinal

stromal tumor, carcinoid tumor. Rare types of this cancer are squamous cell carcinoma, small cell carcinoma, and leiomyosarcoma which can also start in the stomach [164].

4.4.11 Thyroid Cancer

The thyroid gland is below the thyroid cartilage in the front part of the neck anatomically. Thyroid nodules are a common important clinical problem for metabolic activity of human body. Their differentiation to thyroid cancer is becoming increasingly prevalent. It develops from thyroid follicular cells especially. Although thyroid cancers can spread and cause death, for many people thyroid cancer can be cured using surgery and chemotherapy. However, the incidence of thyroid cancer is increasing, thus to learn and understand risk factors of thyroid cancer is crucial for early diagnosis and effective treatment [165], [166].

4.4.12 Urinary Bladder Cancer

Bladder carcinoma is a heterogeneous tumor, presenting as either superficial disease or muscle-invasive disease and it shows diversity for histopathology. Subtypes contain urothelial carcinoma (pure transitional cell carcinoma or focal squamous or glandular differentiation) and pure squamous cell carcinoma [167]. Urinary bladder cancer (UBC) is a known cancer in the world. 2.7 million people have a history of UBC. Its incidence varies over the world and incidence is higher in developed communities than others [168]. To investigate gene expression profiles, molecular mechanisms, metabolic pathways and networks of bladder cancer, network strategy is necessary to find biomarkers for early detection, diagnosis and accurate targeted treatment like other cancers.

4.4.13 Uterine Cancer

The uterus is part of a woman's reproductive system. It is one of the 3 common cancers seen in women (breast, uterine, colorectal) and it is in second place after breast cancer in women. The median age at diagnosis is 61 for uterus cancer, and obesity is a known reason for this type of cancer. Uterus tumor can be benign or malignant. Malignant tumors are important for survival and patients' life quality. Malign uterus cancer may be

a threat to patient, they can be removed by surgery. However, it can relapse again and grow back and then invasion or metastasis can occur to other nearby tissues and organs like other reproductive organs or other parts of body. It is a serious cancer type. Radiation therapy, hormone therapy and/or chemotherapy can be used out of surgery depending on the stage of cancer, histological type, patient's situation [169].

4.5 Differentially Expressed Genes in Cancer

Development, growth and malignancy of any cancer type is a multistep process including initiation, progression, invasion and metastasis. Each of these steps involves multiple genetic alterations that give cancer cells a selective advantage over normal cells [170]. Gene expression profile can change in cancerous tissues, some genes can be up-regulated, or some genes can be down-regulated. In this case, when the genes whose expression is changed are observed in disease condition, the disease process can be interpreted. These genes are named differentially expressed genes (DEG). Finding or discovering compactly described groups of DEG and their better characterization in each specific cancer is very important for understanding cancer molecular mechanism, related pathways and developing new targeted treatment. Differentially expressed genes are a group of genes related to one cancer type or its subtype and their expression level can significantly change between healthy and cancerous conditions. Expression level of one gene can significantly increase in cancer compared to healthy condition, on the other hand its expression can significantly decrease in related cancer. The features that describe the differentially expressed genes in terms of their functionality and interactions with other genes are crucial for all cancer types [171] . DEGs give us important and certain clues on the process of disease formation and treatment. We can observe DEG of a liver is explained in Figure 4.3.

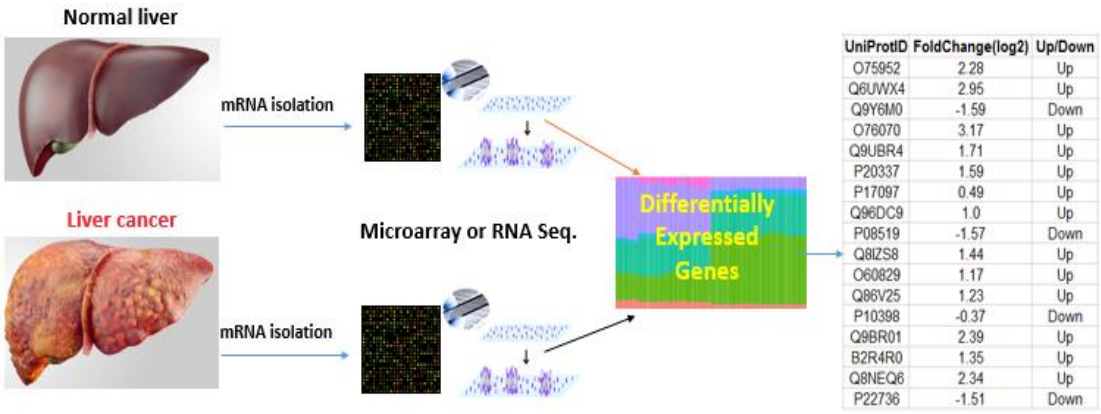


Figure 4.3 Example of DEG between healthy liver tissue and liver cancer

DEG lists for various diseases can easily be calculated via bioinformatic tools or algorithms using microarray and RNA-seq data. There are many studies to examine and execute DEGs for various conditions like cancerous tissues, specific subtypes of cancers, other diseases (neurodegenerative disease, genetic disorders, infections etc.). Xu et al. presented gene expression profile for lung adenocarcinoma using microarray data. They identified significantly differentially expressed genes via t-test. Fold change was also used such that $\log_2$(fold change) is higher than 0.5. If expression of a gene increases or decreases 0.5 and higher fold compared to normal expression, it is denoted as differentially expressed gene [157]. Yang et al. determined DEG using RNA-seq data of hepatocellular carcinoma. The aim of their study was to analyze and to identify critical genes related to the pathogenesis and prognosis through differentially expressed genes in cancer [172]. Another research group carried out studies to obtain genes which had DEG ($\log_2$(fold-change) greater than 1.2) in liver cancer treated with berberine molecule which is a pharmacological botanical agent to understand its impact over the liver cancer [173]. There are plenty of examples of DEG studies from literature.

4.6 Solid Tumor Heterogeneity

Cancers emerge from one abnormal cell that goes through multiple transformations from normal to malignant features as a basic definition. Cancer is a complicated disease. Its development and outcomes vary from one patient to another in terms of heterogeneity. The heterogeneity and diversity are also present at the cellular and molecular level. Cancer is a multi-step process in which cells undergo metabolic and

behavioral changes leading to excessive and uncontrolled proliferation, escape from the surveillance of the immune system, and ultimately invade distant tissues to form metastases. These changes arise by accumulation of disturbance in cell proliferation and modifications in genetic programs that control its life span, relationship to neighboring cells that is called microenvironment, and its escape capability from the immune system. Such tumors can be asymptomatic for a long time. However, it will eventually lead to changes in physiological functions, the replacement of the immune system and the spread of a large number of symptomatic cells depending on its size [174], [175].

Solid tumors are defined as tumors that form discrete masses, such as carcinomas or sarcomas, which are leading cause of death in all over the world. The molecular complexity of tissues and the inaccessibility of most cells within a tissue hinder the discovery of key targets for tissue-specific targeted delivery of therapeutic agents to solid tumors. Most solid tumors have series of mutations; however, as with point mutations, the majority of them are translocations [176].

Tumor heterogeneity [3] has been studied previously but its impact on the carcinogenesis processes and treatment are poorly understood. There are two classes about tumor heterogeneity, first one is ***'inter-tumor heterogeneity'*** that describes spatial heterogeneity found between different tumors in individual patients. The second one is ***'intra-tumor heterogeneity'*** that describes heterogeneity within each lesion [4]. Tumoral heterogeneity is a major challenge in mapping of cancers, diagnosis and therapy approaches [177]. Figure 4.4 is demonstrated that progression of cancer heterogeneity from inter-patient to intra-genomic heterogeneity.

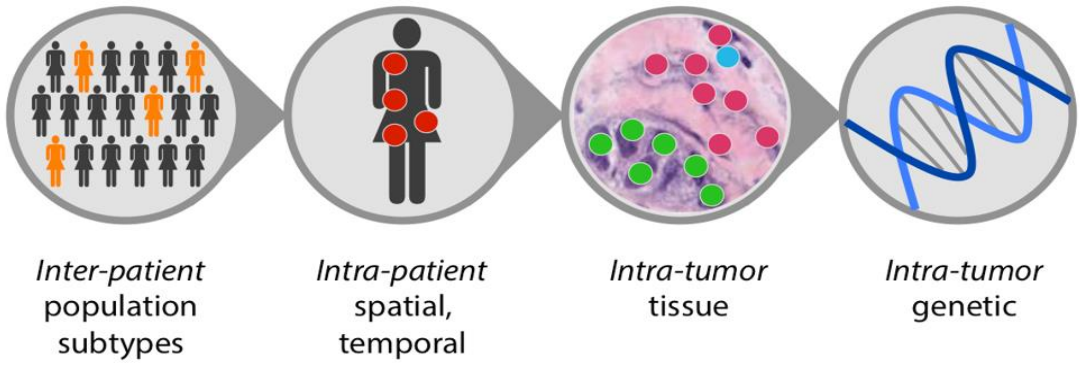


Figure 4.4 Tumoral heterogeneity in tumorigenesis process [178]

Cancer heterogeneity has deep implications for effective drug therapy particularly for solid tumors. Targeted therapy is known to be dependent on solid tumor types, thus, patient outcomes fall in a wide range, depending on tumor lesion. Targeted therapeutic molecules do not help all patients, and even when they show clinical improvements, it is often limited. Thanks to deep sequencing methods, the extent and prevalence of intra and inter-tumoral heterogeneity is increasingly uncovered according to various studies in literature [179].

4.6.1 Intra-Tumoral Heterogeneity

Intra-tumoral heterogeneity is defined mainly as diversity among the cells of only one tumor tissue. Intra-tumoral heterogeneity has important implications for personalized medicine approaches since it can limit therapeutic efficacy and lead to resistance to therapy. Thus, tumor heterogeneity is one of the major problems limiting the efficacy of therapies and compromising treatment outcomes. Even though it was discovered previously, it has not been taken into consideration in regard to therapy, it is now being investigated in detail to improve new therapies for each patient. Intra-tumoral heterogeneity phenomenon more recently has been observed throughout the genome by many research groups [4], [176], [180], [181]. We can see that there is increasing evidence about intra-tumor heterogeneity. Solid tumors may comprise of several tissue layers and subpopulations of cells with distinct genomic alterations within the same cancerous tissue.

In intra-tumoral heterogeneity, cells at the opposite ends of large solid cancerous tissue will be spatially distinct and, have more differences than neighboring or same type of cells, it is also associated with tumor microenvironment. As a summary, many solid tumors involve phenotypically and functionally heterogeneous cancer cells depending on the cell types and tissue layers they contain [176], [182].

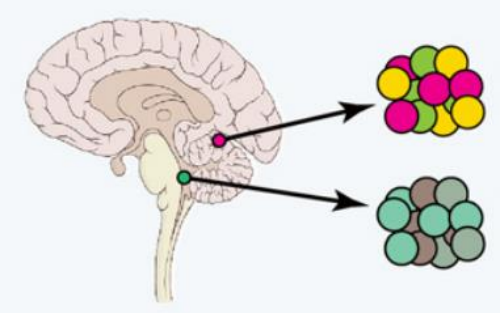


Figure 4.5 Intra-tumoral heterogeneity of brain tumor [182]

Giving an example, brain has a heterogenic structure indicating in Figure 4.5, because same organ has different tissue samples, and various cell types resulting in genetic diversity within a lesion.

Human solid tumors frequently have pronounced heterogeneity of both neoplastic and normal cells on the histological and genetic levels and they can be composed of several tissue layers which are a variety of clones or subpopulations of cells [183]. While such marked intra-tumoral heterogeneity is the result of an inherent interactivity among tumor cells, genomic instability of transcription, translation, or post-translational modifications remains unclear. According to this, solid tumors are heterogeneous in their tissue and cell composition like in Figure 4.6 [184].

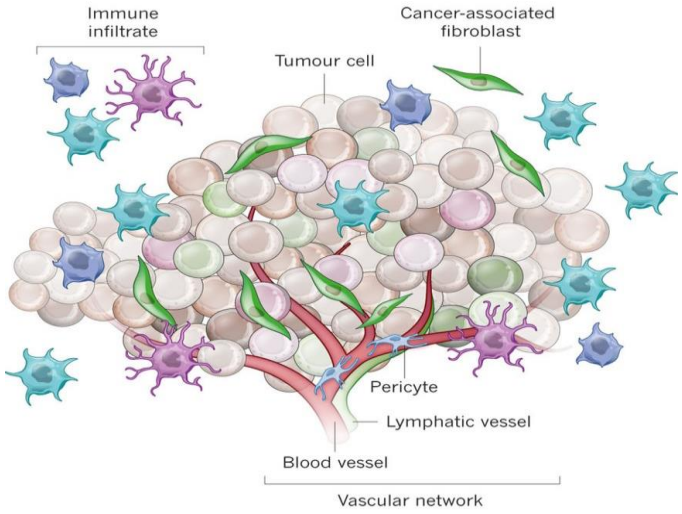


Figure 4.6 Intra-tumor heterogeneity points out variety of tissue layer and cell types

To sum up, the molecular complexity and comprehensive structure of tissues or organs and the inaccessibility of most cells within a tissue restrict the discovery of key targets for tissue-specific delivery of drugs [5]. Therefore, it is very important to identify tissue

specific genes by using bioinformatic approaches to understand intra-tumoral heterogeneity. Furthermore, the pathogenesis of solid tumors is associated with related tissue or primary tissue of origin. Determination of tissue specific genes for solid tumor is essential to understand their mechanism, identify tumor heterogeneity and develop targeted treatment using tissue specific markers.

4.6.2 Metastatic Heterogeneity

Metastatic heterogeneity can be classified as inter-metastatic and intra-metastatic heterogeneity. Each metastasis is confirmed by a single cell or small group of same cells with a set of fundamental mutations that is called intra-metastatic heterogeneity. Inter-metastatic heterogeneity occurs among different metastatic lesions of the same patient. Patients who have recurrence with a single metastatic lesion can often still be cured by surgery or radiotherapy. On the contrary, patients that have more than one metastatic lesions, could not be visualized by imaging. In this case, therapy is hard since the elimination of a subset of the metastatic lesions in patient will not be enough for long-term survival. For this reason, to understand inter-metastatic heterogeneity or a variety of metastatic tissues in same human body is crucial for accurate diagnosis and exact treatment [176].

4.6.3 Inter-Tumoral Heterogeneity (Interpatient Heterogeneity)

Tumor heterogeneity is one of the major problems inhibiting the efficacy of targeted therapies and yields conflicting treatment outcomes. Inter-tumor heterogeneity in other words inter-patient heterogeneity is the heterogeneity among the tumors of different cancer patients that have been observed and cataloged by every medical oncologist. Some of these differences might be related to host factors, such as somatic mutations, germline variants and some other nongenetic factors. As a result, there is significant inter-tumoral heterogeneity based on genomic heterogeneity, which makes impossible to develop a universally applicable effective therapy. This is a clinical challenge that must be overcome in near future. Hence, understanding the mechanism behind heterogeneity is crucial for tumorigenesis research [4], [176].

Tumoral heterogeneity is summarized with a figure as follows, Figure 4.7:

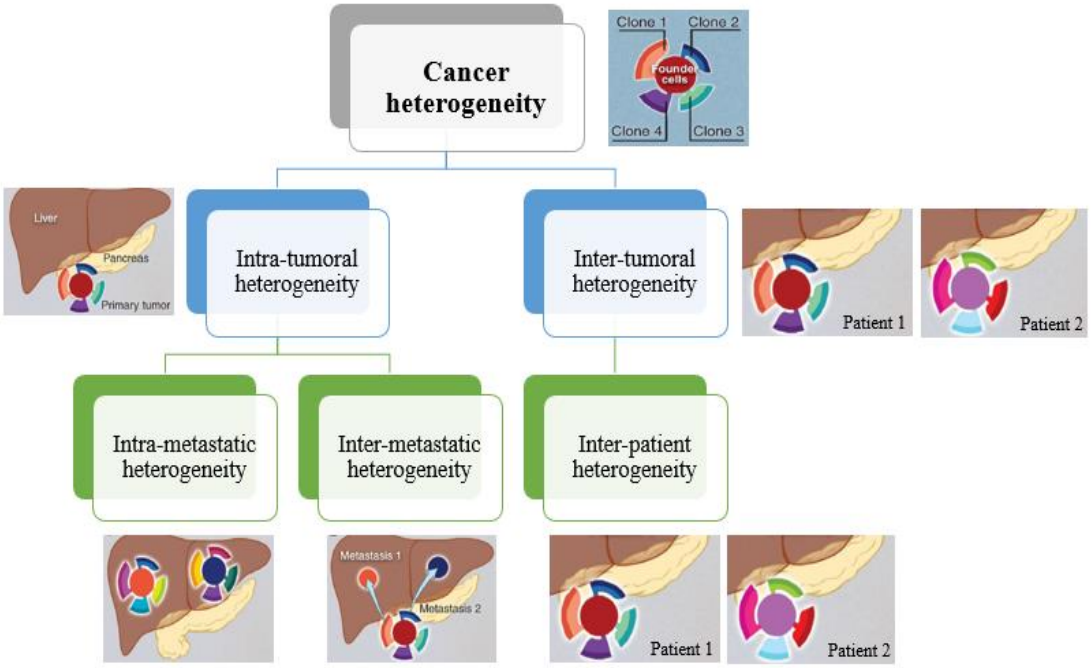


Figure 4.7 Four types of heterogeneity in tumor tissue in liver example [176]

Heterogeneity of tumors was summarized in Figure 4.7 givinfg liver cancer example. In Figure 4.7, intra-tumoral heterogeneity is defined simply as heterogenetiy within a single tumor; intra-metastatic is defined as heterogenetiy within a single metastatic lesion; inter-tumoral or inter-patient is defined as heterogenetiy between two or more different patients. Tumoral heterogenetiy is also associated with cancer cell migration and metastasis processes. Therefore, metastasis, primary and secondary tumor definitions were explained in new section.

4.7 Cancer Metastasis

Cancer is a very deadly disease because of its ability to spread to other body parts from origin of it. This is basically called “metastasis”. There are three different metastasis processes:

- Cancer cells can spread **locally** migrating within the same tissue.
- Cancer cells can spread **regionally** migrating into the lymph nodes, other tissues and organs.
- Cancer cells can spread to **distant parts of the human body** [185].

Cancer which spreads to other parts of the body is called “metastatic cancer” or “stage IV cancer”. Other two important definitions for cancer migration are “primary” and

"secondary". Cancer originating from a tissue is called **primary tumor** of that tissues. Cancer cells from a primary tumor may spread to other parts of the body and it causes a new **secondary tumor** in other tissues or organs.

Metastatic cancer cells have features similar to primary tumor cells. So, they are not like the cells in secondary tissue or organ. In this way, medical doctor can determine origin of a secondary tumor. Sometimes when people are diagnosed with metastatic cancer, medical oncologist cannot tell which tissue it originated from. This type of cancer is called "cancer of unknown primary origin". Hence, to find origin of the metastatic cancer is important for effective therapy [185], [186]. Metastatic cancer cell has some invasive properties to migrate the other part of the body. Metastasis process was summarized in Figure 4.8:

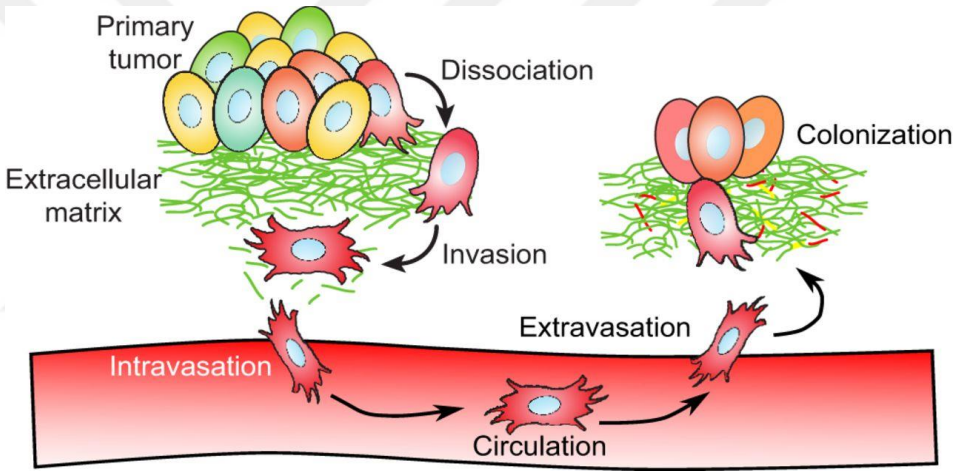


Figure 4.8 General progression of metastasis [187]

In progression of metastasis (Figure 4.8), circulation is key for migration of cells and metastasis process. Tumor microenvironment is heterogeneus structure in the context of diverse composition surrounding the cancerous tissue.

The most common sites where cancer cells migrate are bones, liver, and lungs in the body. Table 4.1 below summarizes the most common sites of metastasis per cancer type for primary tumors:

Table 4.1 Cancer types and their metastatic tissues or organs in the body

Cancer Type	**Main Sites of Metastasis**	**References**
Breast cancer	Bone, brain, liver, lung Rarely: esophagus, parotid,	[185], [188], [189] [190], [191]
Cervical cancer	Lymph node, liver, lung, peritoneum, bone, brain	[185], [192]
Esophageal Cancer	Lymph node, head and neck, brain, lung, trachea, aorta, heart, breast, liver, skin, thyroid Rarely: pancreas, spleen, small intestine	[193], [194],[195]
Kidney cancer	Lymph node, adrenal gland, bone, brain, liver, lung Rarely, spleen, colon, small intestine, urinary system, ovary, paranasal sinuses, heart	[185], [196], [197], [198], [199], [200],[201]
Lung cancer	Lymph node, adrenal gland, bone, brain, liver, another lung Rarely, stomach, small and large intestines, pancreas, skin, kidney, breast, testis	[185], [202], [203], [204] [205]
Liver cancer (hepatocellular carcinoma)	Lymph node, bone, lung, distant organs, testis	[206], [207], [208], [209] [205]
Pancreatic Cancer	Liver, lung, peritoneum, adrenal gland, bone, lymph node Rarely: esophagus, heart	[185], [210], [211],[212], [213]
Prostate cancer	Lymph node, brain, bone, liver, lung; rarely, adrenal gland, breast, kidney, muscle, pancreas, salivary gland, spleen, testis	[185], [214], [215]

Table 4.1 Cancer types and their metastatic tissues or organs in the body (cont'd)

Colorectal cancer	Lung, liver, brain, peritoneum, ovary Rarely: brain, bone, adrenal, spleen, testis, pancreas	[216], [217],[218], [219]
Stomach cancer	Liver, lung, peritoneum, bone	[185], [220]
Thyroid cancer	Bone, liver, lung Rarely: testis, pancreas	[185], [221], [222], [223]
Uterine cancer	Bone, brain, liver, lung, peritoneum, vagina, pelvic, muscle Rarely: pancreas, heart	[185], [224], [225], [226]
Urinary Bladder	Lymph node, bone, liver, lung, peritoneum, pleura, pancreas Rarely; brain, adrenal, urethra, penis, penile intestine	[185], [227], [228]

When cells of a single tumor migrate to other tissues or organs, it can be very hard to control. Approximately 90% of mortality in cancer patients is due to metastatic cancer. Although some types of metastatic cancer can be cured with current therapy methods, most types cannot be cured. There is an urgent need to develop better and more effective therapies to overcome metastatic cancer. For this purpose, the detection of origin of secondary tumor is a necessity for correct treatment. In addition, it is also very important to know where primary cancer will migrate, and consequently predicting the tissue or organ at which secondary cancer will start.

Many research groups focus on molecular mechanisms of metastatic tumors. Especially, genetic instability and molecular mechanisms of intra-tumoral heterogeneity are most important factors to be elucidated in metastasis [229]. In general, primary tumors initially display a high degree of heterogeneity and they have cells with various metastatic potentials, after migration of them, secondary tumors have also genetic heterogeneity [230]. As a summary, metastasis has been defined as the migration of tumor cells from the primary tissue, angiogenesis, survival, extravasation of the circulatory system, and progressive colonization in the other tissue/organ of body,

respectively. Besides, genetic instability has been supported by metastasis, and the varying microenvironments of metastasis contribute to the intra-tumoral heterogeneity [231].

In general, the main objective of my thesis is to explore the possibility of using gene expression data and tissue specific gene information to gain insight about intra-tumoral heterogeneity. Therefore, in this chapter, cancer definition, related cancer types, differentially expressed genes between healthy and tumor tissues, classes of tumoral heterogeneity and metastasis are explained. Interestingly, tumor heterogeneity is related to tissue-specific expression profiles of genes and tissue/organ specific metastatic features of cancer cells, thus we examined intra-tumoral heterogeneity in the context of tissue specificity by generating a new robust, rigorous and comprehensive computational approach.

CHAPTER 5

METHODS

PART 1: DETERMINATION OF TISSUE-SPECIFIC GENES THROUGH A NEW APPROACH

5.1 Raw Data Accession and Preparation

The emergence of high-throughput sequencing technologies has enabled thousands of genes to be checked simultaneously under multiple conditions [110]. According to Kryuchkova-Mostacci and Robinson-Rechavi [38] correlation between tissue specificity methods is higher in RNA-seq than microarray. Thus, we used RNA-Seq data to determine the tissue-specific genes.

R programming language [232] is a free software environment for applying statistical computing and graphics which is used in variety of fields. In this thesis study, R version 3.4.3 has been used. R Studio [233] is an integrated development environment (IDE) for R which is commonly used by R users. There are many useful packages in R to manipulate data and calculate statistical measurements (tidyverse package including tidyr, dplyr, purr, readr), to import-export data (readr, readxl), to plot aesthetic graphs (ggplot), to generate and visualize networks (ggraph) and exercise reproducible research where code and report are knitted together (rmarkdown and knitr). R programming language is also being used extensively for bioinformatic analyses, because many bioinformatic tools and databases cooperate with R (Bioconductor). The analysis steps of this thesis study were carried out with R programming.

5.1.1 Raw Data Information

In this study, 5 different big datasets about gene expression of healthy tissue samples were used as raw data. Detailed information about the data is explained in the table on next page.

Data were downloaded from **ArrayExpress** [234] and **ExpressionAtlas** [235]. ArrayExpress is a database, established by EMBL-EBI of functional genomics data and it serves the data to the research community. Similarly, ExpressionAtlas is established by EMBL-EBI and is an open access science resource that gives us an effective way to find information about gene and protein expression for many species and biological conditions such as different tissues, cell types, developmental stages and diseases.

The following human gene expression profiling study data were downloaded as plain text files from ArrayExpress and ExpressionAtlas:

- **E-MTAB-1733:** RNA-seq of 95 human individuals representing 27 different tissues
- **E-MTAB-2836:** RNA-seq of 122 human individuals representing 32 different tissues
- **E-MTAB-5214:** RNA-seq from 53 human tissue samples from the Genotype-Tissue Expression (GTEx) Project
- **E-MTAB-3358:** RNA-Seq CAGE Cap Analysis of gene expression analysis of 56 human tissues in RIKEN FANTOM5 Project
- **E-MTAB-4344:** Strand-specific RNA-seq of 13 human tissues from Michael Snyder's lab for the ENCODE Project (Encyclopedia of DNA Elements Consortium)

Some properties like number of tissues types, number of samples, sex, experimental process, data type, and normalization method were summarized in Table 5.1.

Table 5.1 Information about five datasets used in this study

Data	Number of tissue	Number of samples	Sex	Developmental stage	Experiment type Platform/Model	Data type	Expression level	Citations	Raw Data Provider
E-MTAB-2836	32	122 individuals	Male and female	Adult	Illumina HiSeq 2000 paired end sequencing	HPA, RNA-seq	FPKM	[236], [237], [238]	Uhlan's Lab
E-MTAB-1733	27	95 individuals	Male and female	Adult	Illumina HiSeq 2000 paired end sequencing	HPA, RNA-seq	FPKM	[239], [240], [241], [242], [243]	-
E-MTAB-5214	53	1641 samples 175 individuals	Male and female	Adult	Illumina HiSeq 2000 paired end sequencing	RNA-seq	FPKM	[244]	GTEx Project
E-MTAB-3358	56	-	Male and female	Adult	HELICOS, Helicos HeliScope	RNA-Seq, CAGE	TPM	-	FANTOM5 Project
E-MTAB-4344	13	348 samples	Male and female	Adult	Different platforms	RNA-seq	FPKM	[245]	ENCODE Project

The studies contained RNA-Seq data conducted on lots of different cell types and also various tissue types. 5 different datasets have contained 98 different unique types of healthy tissue. 96 of them were selected for use in the thesis study. However, when we investigated tissue samples by phylogenetic analysis via **Brenda Tissue Ontology** (BTO) [246], a structured controlled vocabulary for the source of tissues in phylogenetic tree approach, some of the tissues were found to be the part of the other tissues. Therefore, all tissues were examined by Brenda Tissue Ontology and grouped together with BTO numbers. At this stage, 96 different tissue subtypes were considered as **"child tissue"** and 39 tissues were defined as main tissues and they were called **"parent tissue"**. All tissue types and their BTO accession ID are listed in the following table, Table 5.2:

Table 5.2 Tissue list and their BTO accession number

Group BTO ID	Group name	Tissue BTO ID	Tissue name
BTO_0001487	adipose tissue	BTO_0001487	adipose tissue
BTO_0001487	adipose tissue	BTO_0004042	subcutaneous adipose tissue
BTO_0001487	adipose tissue		greater omentum
BTO_0000047	adrenal gland	BTO_0000047	adrenal gland
BTO_0000084	appendix	BTO_0000084	appendix
BTO_0000123	bladder	BTO_0000123	bladder
BTO_0000123	bladder	BTO_0000493	gall bladder
BTO_0000123	bladder	BTO_0001418	urinary bladder
BTO_0000141	bone marrow	BTO_0000141	bone marrow
BTO_0000142	brain	BTO_0000142	brain
BTO_0000142	brain	BTO_0001042	amygdala
BTO_0000142	brain	BTO_0000211	caudate nucleus
BTO_0000142	brain	BTO_0000231	cerebellar hemisphere
BTO_0000142	brain	BTO_0000232	cerebellum
BTO_0000142	brain	BTO_0000233	cerebral cortex
BTO_0000142	brain		cerebral meninges
BTO_0000142	brain	BTO_0000342	diencephalon
BTO_0000142	brain	BTO_0001637	dura mater
BTO_0000142	brain	BTO_0002246	globus pallidus
BTO_0000142	brain	BTO_0000601	hippocampus
BTO_0000142	brain	BTO_0000614	hypothalamus
BTO_0000142	brain	BTO_0001408	locus ceruleus
BTO_0000142	brain	BTO_0000041	medulla oblongata
BTO_0000142	brain	BTO_0000144	brain meninx
BTO_0000142	brain	BTO_0004834	middle frontal gyrus
BTO_0000142	brain		middle temporal gyrus
BTO_0000142	brain	BTO_0001862	nucleus accumbens
BTO_0000142	brain		occipital cortex

Table 5.2 Tissue list and their BTO accession number (cont'd)

BTO_0000142	brain	BTO_0000293	occipital lobe
BTO_0000142	brain	BTO_0001001	parietal lobe
BTO_0000142	brain	BTO_0001067	pineal gland
BTO_0000142	brain	BTO_0000427	pituitary gland
BTO_0000142	brain	BTO_0000212	putamen
BTO_0000142	brain	BTO_0000143	substantia nigra
BTO_0000142	brain	BTO_0001365	thalamus
BTO_0000142	brain	BTO_0001362	olfactory apparatus
BTO_0000149	breast	BTO_0000149	breast
BTO_0000269	colon	BTO_0000269	colon
BTO_0000269	colon	BTO_0000645	sigmoid colon
BTO_0000269	colon	BTO_0000272	transverse colon
BTO_0000408	epididymis	BTO_0000408	epididymis
BTO_0000959	esophagus	BTO_0000959	esophagus
BTO_0000959	esophagus		esophagus muscularis mucosa
BTO_0000959	esophagus	BTO_0004364	gastroesophageal junction
BTO_0000959	esophagus	BTO_0002859	esophagus mucosa
BTO_0000562	heart	BTO_0000562	heart
BTO_0000562	heart	BTO_0000903	atrium
BTO_0000562	heart	BTO_0001702	left atrium
BTO_0000562	heart	BTO_0001629	left ventricle
BTO_0000671	kidney	BTO_0000671	kidney
BTO_0000671	kidney	BTO_0001166	cortex of kidney
BTO_0000759	liver	BTO_0000759	liver
BTO_0000763	lung	BTO_0000763	lung
BTO_0000784	lymph node	BTO_0000784	lymph node
BTO_0000784	lymph node	BTO_0001387	tonsil
BTO_0000887	muscle	BTO_0001260	smooth muscle
BTO_0000887	muscle	BTO_0001103	skeletal muscle
BTO_0000975	ovary	BTO_0000975	ovary
BTO_0000980	oviduct	BTO_0000980	oviduct
BTO_0000988	pancreas	BTO_0000988	pancreas
BTO_0000405	penis	BTO_0000405	penis
BTO_0001078	placenta	BTO_0001078	placenta
BTO_0001129	prostate	BTO_0001129	prostate
BTO_0001158	rectum	BTO_0001158	rectum
BTO_0001203	salivary gland	BTO_0001203	salivary gland
BTO_0001203	salivary gland		minor salivary gland
BTO_0001203	salivary gland	BTO_0001004	parotid gland
BTO_0001203	salivary gland	BTO_0001316	submandibular gland
BTO_0001234	seminal vesicle	BTO_0001234	seminal vesicle
BTO_0001253	skin	BTO_0001253	skin
BTO_0001253	skin		lower leg skin
BTO_0001253	skin		suprapubic skin

Table 5.2 Tissue list and their BTO accession number (cont'd)

BTO_0001253	skin		zone of skin
BTO_0000651	small intestine	BTO_0000651	small intestine
BTO_0000651	small intestine	BTO_0000365	duodenum
BTO_0000651	small intestine	BTO_0001784	small intestine Peyers patch
BTO_0001281	spleen	BTO_0001281	spleen
BTO_0001279	spinal cord	BTO_0001279	spinal cord
BTO_0001307	stomach	BTO_0001307	stomach
BTO_0001363	testis	BTO_0001363	testis
BTO_0001379	thyroid	BTO_0001379	thyroid
BTO_0001385	tongue	BTO_0001385	tongue
BTO_0001424	uterus	BTO_0001424	uterus
BTO_0001424	uterus	BTO_0001422	endometrium
BTO_0001424	uterus	BTO_0001421	cervix
BTO_0001424	uterus		ectocervix
BTO_0001424	uterus	BTO_0003002	endocervix
BTO_0000243	vagina	BTO_0000243	vagina
BTO_0000564	valve		mitral valve
BTO_0000564	valve		pulmonary valve
BTO_0000564	valve		tricuspid valve
BTO_0001427	vas deferens	BTO_0001427	vas deferens
BTO_0001102	vessel	BTO_0000573	artery
BTO_0001102	vessel	BTO_0000290	coronary artery
BTO_0001102	vessel	BTO_0000135	aorta
BTO_0001102	vessel		tibial artery
BTO_0000089	whole blood	BTO_0000089	whole blood

5.1.2 Filtration of Protein Coding Genes

Each RNA sequencing dataset of human tissues retrieved from the human body included different number of genes. The genes from each dataset were inspected via **BioMart** [247] and found to be categorized into various gene types such as, protein coding, non-protein coding, transcription factors, miRNA or lncRNAs. In this study, we focused on only **protein coding genes**. For this reason, we filtered the datasets according to the gene type at BioMart.

5.1.3 Normalization Methods After Sequencing

RNA-seq has a wide variety of applications. The most common application of RNA-seq is to estimate gene and transcript expression. However, no single analysis pipeline can be used in all cases. Analysis method is primarily based on the number of reads. Count of

raw reads is not sufficient to compare expression levels among samples because these values are affected by factors such as transcript length, total number of reads, sequencing biases, experimental errors and laboratory conditions [248]. The following normalization methods are used when working with the RNa-seq data:

Transcripts per million (TPM): TPM is a measurement of the proportion of transcripts in RNA-seq data. TPM counts per length of transcript (kb) per million reads mapped. It takes the transcript length and sequencing depth into consideration; a natural measurement is the rate. The rate is also dependent on the total number of fragments. To adjust for this, measurement is divided by the sum of all rates and this calculation gives the proportion of transcript i in sample. TPM has a powerful interpretation if researchers are looking at transcript abundances [249]. TPM formula:

$$\mathrm{TPMi} = \frac{\mathrm{Xi}}{\hat{\mathrm{l}}\mathrm{i}} * \left(\frac{1}{\sum_{\mathrm{j}} \frac{\mathrm{Xi}}{\hat{\mathrm{l}}\mathrm{i}}}\right) * 10^{6} \tag{5.1}$$

- Xi is defined as number of reads mapping to transcript i
- li is defined as length of transcript i

Reads per kilobase of exon per million mapped reads (RPKM) and fragments per kilobase of exon model per million mapped reads (FPKM): RPKM and FPKM are essentially the same normalization methods. FPKM is equal to RPKM if data is single-end reads. RPKM is a sample normalization method that will cancel out the read length differences and library size effects. RPKM, FPKM, and TPM are used frequently to normalize RNA-seq gene expression (N→ number of reads) [248]. FPKM formula:

$$\mathrm{FPKMi} = \frac{\mathrm{Xi}}{\left(\frac{\hat{\mathrm{l}}\mathrm{i}}{10^{3}}\right) * \left(\frac{\mathrm{N}}{10^{6}}\right)} = \frac{\mathrm{Xi}}{\hat{\mathrm{l}}\mathrm{i}\mathrm{N}} * 10^{9} \tag{5.2}$$

Relationship between FPKM and TPM: The FPKM and TPM values are calculated using the total number of mapped reads. Both FPKM and TPM methods account for the length of genomic features. FPKM corrects for the number of reads that have been sequenced. In addition to this, TPM measures the average number of mapped bases per read [250]. FPKM is a commonly used normalization measure, whereas TPM values have been

shown to meet the invariant average criterion [251]. FPKM can be converted into TPM using a simple formula [252]:

$$\mathrm{TPMi} = \left(\frac{\mathrm{FPKMi}}{\sum_{\mathrm{j}} \mathrm{FPKMj}}\right) * 10^{6} \tag{5.3}$$

In our datasets, FPKM had been used in four raw datasets including EMTAB-1733, EMTAB-2836, EMTAB-5214 and EMTAB-4344. In these datasets, for each tissue, the average FPKM value of all individual samples was used to estimate the gene expression level. TPM was only used in normalization of EMTAB-3358 data. In this study, FPKM normalized data for four datasets and TPM normalized data for one dataset were used without making any conversion of normalization.

5.1.4 Arrangement of Downloaded Data

Data were downloaded from ArrayExpress and ExpressionAtlas as txt files. When we examined tissue names, we noticed that there were discrepancies between datasets in tissue naming. For instance, one dataset had tissue name “small intestine”, but other dataset named the same tissue as “smint”. In order to remove discrepancies, we generated a look up table of tissue names so that only one unique name has been used for only one tissue and the names of tissues were arranged accordingly.

Having determined all these preliminary operations for editing the raw data, the five datasets with the following code are made ready for analysis by making all these arrangements mentioned above. The code below, Algorithm 1, is used to import and manipulate the all datasets:

Algorithm 1

```
prepareFile <- function(fileName) {
  geneData <- read.table(fileName, sep = "\t", header = TRUE, check.names = TRUE,
    fill = TRUE)
  proteinCodings <- read.csv("protein_coding.csv", sep = "\t", header = TRUE, check.n
ames = TRUE)
  geneData <- geneData[which(geneData[, 1] %in% proteinCodings[, 1]), ]
  tissueMetaData <- read.table("tissue_meta_data.txt", sep = "\t", header = TRUE,
```

```
    check.names = TRUE)
  tissueMetaData <- tissueMetaData[tissueMetaData$fileName == fileName, ]
  tissueMetaData <- tissueMetaData[tissueMetaData$status > 0, ]
  aliasses <- as.matrix(unique(tissueMetaData$alias))
  dimData <- dim(aliasses)
  numRows <- dimData[1]
  result <- data.frame(matrix(0, ncol = 0, nrow = dim(geneData)[1]))
  for (i in 1:numRows) {
    columnList <- tissueMetaData[tissueMetaData$alias == aliasses[i], 1]
    colData <- geneData[, matrix(columnList)]
    status_list <- tissueMetaData[tissueMetaData$alias == aliasses[i], 3]
    if (is.data.frame(colData)) {
      result <- cbind(result, rowMeans(colData, na.rm = TRUE))
    } else {
      result <- cbind(result, colData)
    }
    colnames(result)[i] <- aliasses[i]
  }
  return(result)
}
files <- c("E-MTAB-1733-query-results.txt", "E-MTAB-2836-query-results.txt", "E-MTAB-
5214-query-results.txt", "E-MTAB-4344-query-results.txt", "E-MTAB-3358-query-result
s.txt")
sample_names <- c("data1733", "data2836", "data5214", "data4344", "data3358")
all <- data.frame(files = files, samples = sample_names, stringsAsFactors = FALSE)
geneData <- lapply(all$files, prepareFile)
names(geneData) <- all$samples
head(geneData[[1]][1:5])  #This is an example:

##    EnsemblGeneID    colon    kidney    liver   pancreas
## 1 ENSG00000127720  4.346168  4.447785  2.507183  2.345215
## 2 ENSG00000256574  0.014684  0.040785  0.022350  0.000000
```

```
## 3 ENSG00000109819  8.214000 25.960575 25.385967  2.321475
## 4 ENSG00000161057 45.173380 51.938800 54.337233 16.997100
## 6 ENSG00000051596 26.498260 22.123125 13.683633  5.415120
## 7 ENSG00000172244  1.782478  1.274395  0.327570  0.109125
```

With the above procedure, the five datasets were grouped together as a list and prepared for analysis. At this stage, inappropriate information in the datasets was discarded, the expression levels of the protein coding genes were selected, and the names of the tissues were arranged to be common in all datasets.

Number of genes and tissues before and after filtration is indicated in Table 5.3:

Table 5.3 Number of genes and tissues in each dataset

Number of	E-MTAB-1733	E-MTAB-2836	E-MTAB-5214	E-MTAB-3358	E-MTAB-4344
all tissues	27	32	53	56	13
all genes	20050	57073	57073	21017	57073
filtered tissues	27	32	48	56	13
filtered genes	18863	19676	19676	16438	19676

5.1.5 Apply F-test to Datasets

F test is used when we want to compare two variances. In other words, the test compares the ratio of two variances. If the variances are equal, the ratio of the variances will be equal to 1. F-test can often be considered a refinement of the more general likelihood ratio test (LR) considered as a large sample chi-square test. F-test can be used in the special case that the error term in a regression model is normally distributed [253]. If the p-value obtained from F-test is less than 0.05 ($p < 0.05$), there is a significant difference between two variances. The F-value produced after the F-test is also very important for the interpretation of the results. If the F-value is small, the similarity between the data in this case is high, so there are no significant differences [254]. F-test was applied to obtain information about the distribution of the data used in the thesis study and the compatibility with each other. All datasets were tested in binary combinations with each other and results were recorded using code in Algorithm 2.

F-test among the five data in binary combinations was applied with the following code:

Algorithm 2

```
len_test <- length(geneData)
test_result <- data.frame(matrix(NA, ncol = 4, nrow = 25))
colnames(test_result) = c("F statistics", "P value", "data", "data2")
names(test_result)
cursor = 1
for (i in 1:len_test) {
  for (j in 1:len_test) {
    res <- var.test(as.numeric(unlist(geneData[[i]])), as.numeric(unlist(geneData[[j]])),
      ratio = 1, alternative = c("two.sided", "less", "greater"), conf.level = 0.95)
    test_result[cursor, ] <- c(as.numeric(res["statistic"]), as.numeric(res["p.value"]),
      names(geneData)[i], names(geneData)[j])
    cursor = cursor + 1
  }
}
f_test_result <- write.csv(test_result, file = "F_test_result.csv")
```

We can see F-test results as a table for binary combination of datasets like in Figure 5.4:

Table 5.4 F- test results between datasets for each group

F statistics	P value	Data 1	Data 2
1.0000000	1	data1733	data1733
0.1252055	0	data1733	data2836
0.1811273	0	data1733	data5214
0.0578017	0	data1733	data4344
0.9378366	0	data1733	data3358
7.9868689	0	data2836	data1733
1.0000000	1	data2836	data2836
1.4466403	0	data2836	data5214
0.4616542	0	data2836	data4344
7.4903780	0	data2836	data3358
5.5209779	0	data5214	data1733
0.6912569	0	data5214	data2836
1.0000000	1	data5214	data5214
0.3191217	0	data5214	data4344
5.1777752	0	data5214	data3358
17.3005430	0	data4344	data1733
2.1661233	0	data4344	data2836

Table 5.4 F- test results between datasets for each group (cont'd)

3.1336012	0	data4344	data5214
1.0000000	1	data4344	data4344
16.2250825	0	data4344	data3358
1.0662838	0	data3358	data1733
0.1335046	0	data3358	data2836
0.1931331	0	data3358	data5214
0.0616330	0	data3358	data4344
1.0000000	1	data3358	data3358

According to p-values of F-test, there is no significant difference between each pair of datasets meaning that we can use the determined five datasets for further analysis. If F-value of test is examined, large scores can be shown between some datasets. For instance, F-value of data4344 - data1733 is higher than others. It can be interpreted there is a difference between them. Number of tissues of these datasets is different from each other. Data3358 was normalized via TPM and data4344 was normalized through FPKM and also number of tissues is different from each other. In this test both p-value and F-value are used together to decide on the differences.

Several graphs can be drawn to examine the data after the F-test. In order to show and interpret distribution of the data, box plot and violin plot were plotted using log2 based normalized data. When data were normalized according to log2, values less than 1 are "-". If we use **ifelse (condition, true, false)** function, we can assign new value to problematic expression values. Box and violin plots were plotted using the Algorithm 3:

Algorithm 3

```
b_data1733 <- log2(sapply(unlist(geneData$data1733[, 2:28]), function(x) ifelse(x < 1,1, x)))
b_data2836 <- log2(sapply(unlist(geneData$data2836[, 3:34]), function(x) ifelse(x < 1,1, x)))
b_data5214 <- log2(sapply(unlist(geneData$data5214[, 3:50]), function(x) ifelse(x < 1,1, x)))
b_data4344 <- log2(sapply(unlist(geneData$data4344[, 3:15]), function(x) ifelse(x < 1,1, x)))
b_data3358 <- log2(sapply(unlist(geneData$data3358[, 3:58]), function(x) ifelse(x < 1,1, x)))
```

```
boxplot <- boxplot(b_data1733, b_data2836, b_data5214, b_data4344, b_data3358,
xlab = "Data Labels", ylab = "Expression Level", main = "Distribution of Data in Each
Dataset", col = terrain.colors(5), names = c("EMTAB-1733", "EMTAB-2836", "EMTAB-
5214", "EMTAB-4344", "EMTAB-3358"), cex.lab = 1.5, cex.axis = 0.8, cex.main = 1.5,
cex.sub = 1.5)
legend("topright", inset = 0.01, cex = 0.6, title = "Normalization type", c("FPKM",
  "FPKM", "FPKM", "FPKM", "TPM"), fill = terrain.colors(5), horiz = TRUE)

library(vioplot)
b_data1733<-log2(sapply(unlist(geneData$data1733[, 2:28]), function(x) ifelse(x <
1,20, x)))
b_data2836<-log2(sapply(unlist(geneData$data2836[, 3:34]), function(x) ifelse(x <
1,20, x)))
b_data5214<-log2(sapply(unlist(geneData$data5214[, 3:50]), function(x) ifelse(x <
1,20, x)))
b_data4344<-log2(sapply(unlist(geneData$data4344[, 3:15]), function(x) ifelse(x <1,
20, x)))
b_data3358<-log2(sapply(unlist(geneData$data3358[, 3:58]), function(x) ifelse(x <1,
20, x)))
vioplot <- vioplot(b_data1733, b_data2836, b_data5214, b_data4344, b_data3358,
names = c("1733", "2836", "5214", "3358", "4344"), col = "blue")
title("Distribution of data in each dataset via violin plot")
figs(name = "vioplot", "Distribution of data in each dataset via violin plot")
```

We can see distribution of all datasets as a box plot and a violin plot like in Figure 5.1:

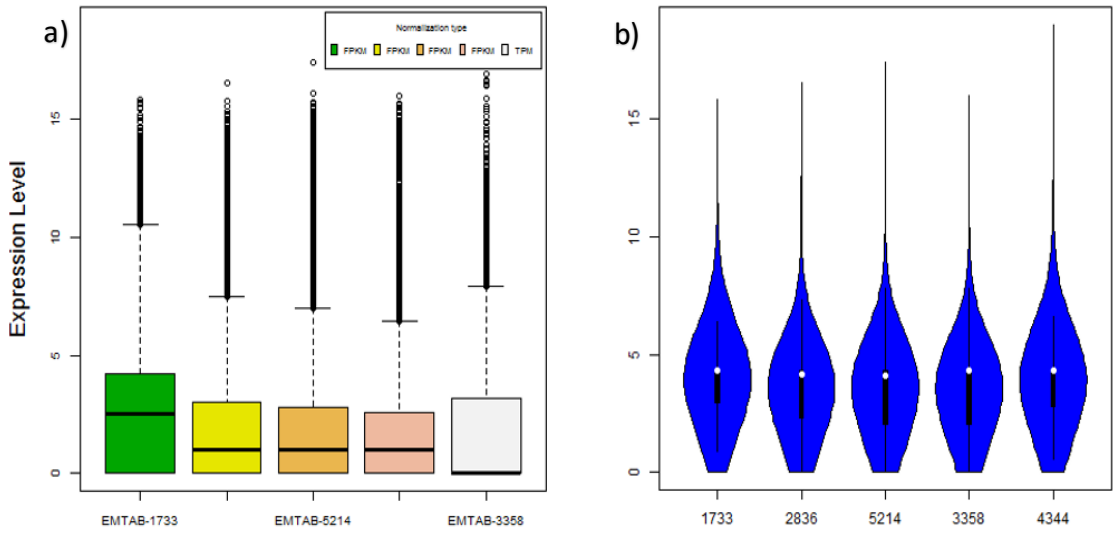


Figure 5.1 Distribution of expression for each group of data as a box (a) and a violin plot (b)

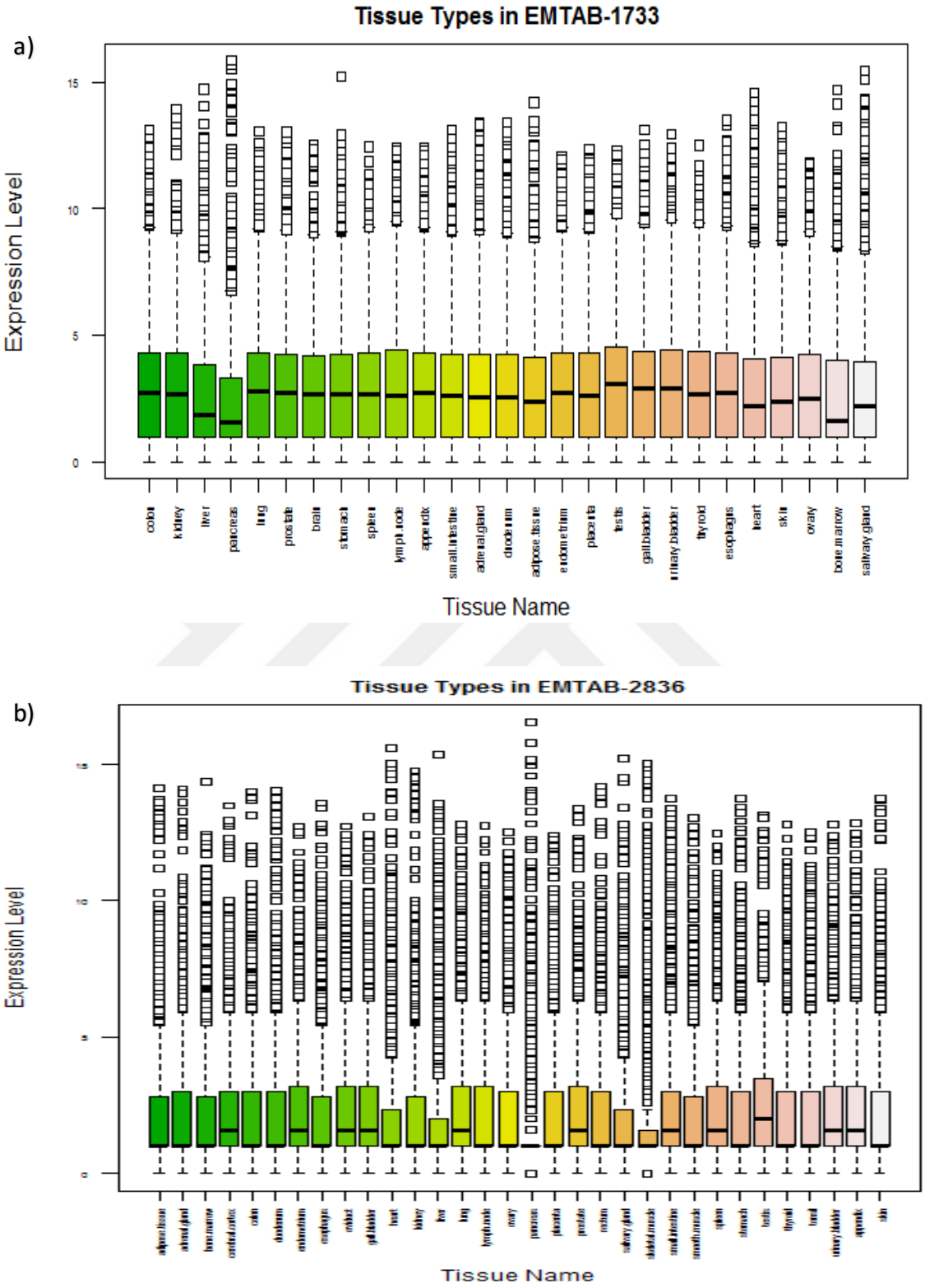
According to boxplot and violin plot, in Figure 5.1, distribution of EMTAB-3358 is a little different from other datasets because of different normalization method. However, distribution of other datasets is suitable and similar to each other.

We can see tissue types and distribution of expression in each dataset using box plots. Here, there is an example of graphs (data 1733) as code, all graphs were plotted using the same code like in Algorithm 4:

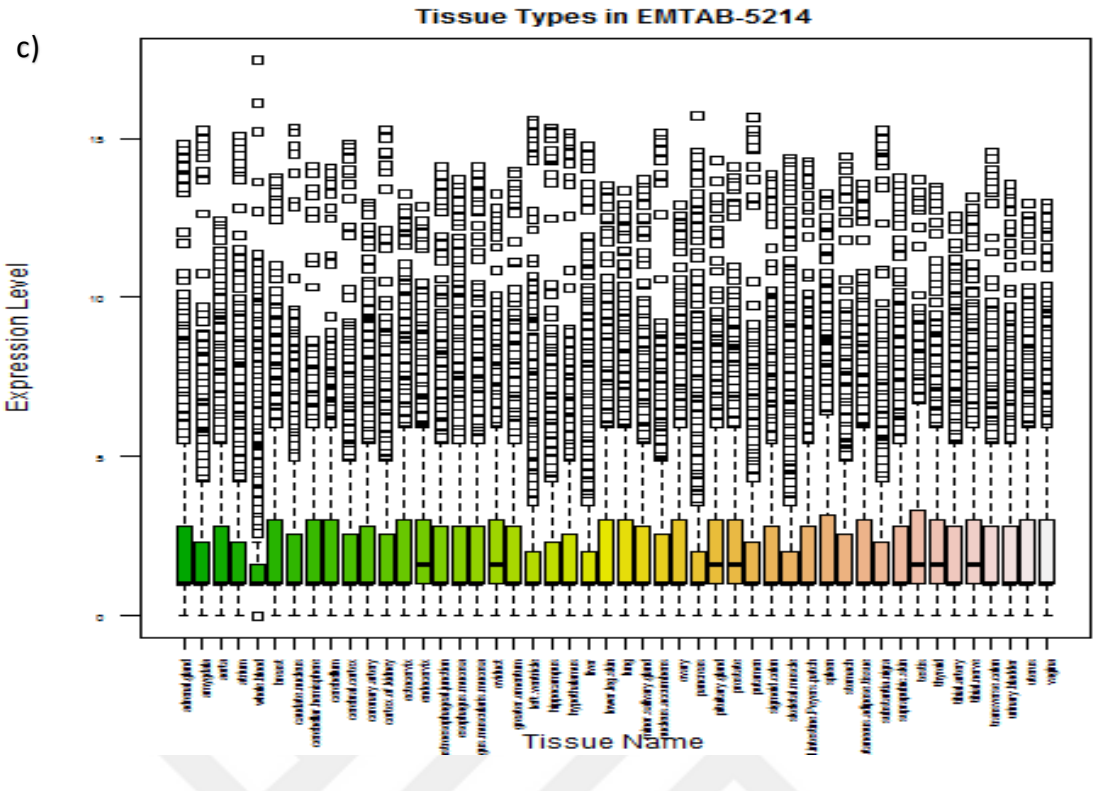
Algorithm 4

```
t_data1733 <- log2(sapply(geneData$data1733[, 2:28], function(x) ifelse(x < 1, 2, x)))
t_data2836 <- log2(sapply(geneData$data2836[, 3:34], function(x) ifelse(x < 1, 2, x)))
t_data5214 <- log2(sapply(geneData$data5214[, 3:50], function(x) ifelse(x < 1, 2, x)))
t_data4344 <- log2(sapply(geneData$data4344[, 3:15], function(x) ifelse(x < 1, 2, x)))
t_data3358 <- log2(sapply(geneData$data3358[, 3:58], function(x) ifelse(x < 1, 2, x)))
boxplot(t_data1733, ylab = "Expression Level", main = "Tissue Types in EMTAB-1733",
col = terrain.colors(ncol(t_data1733)), names = colnames(t_data1733), cex.lab = 1,
cex.axis = 0.5, cex.main = 1, cex.sub = 1, las = 2)
mtext("Tissue Name", side = 1, line = 4
```

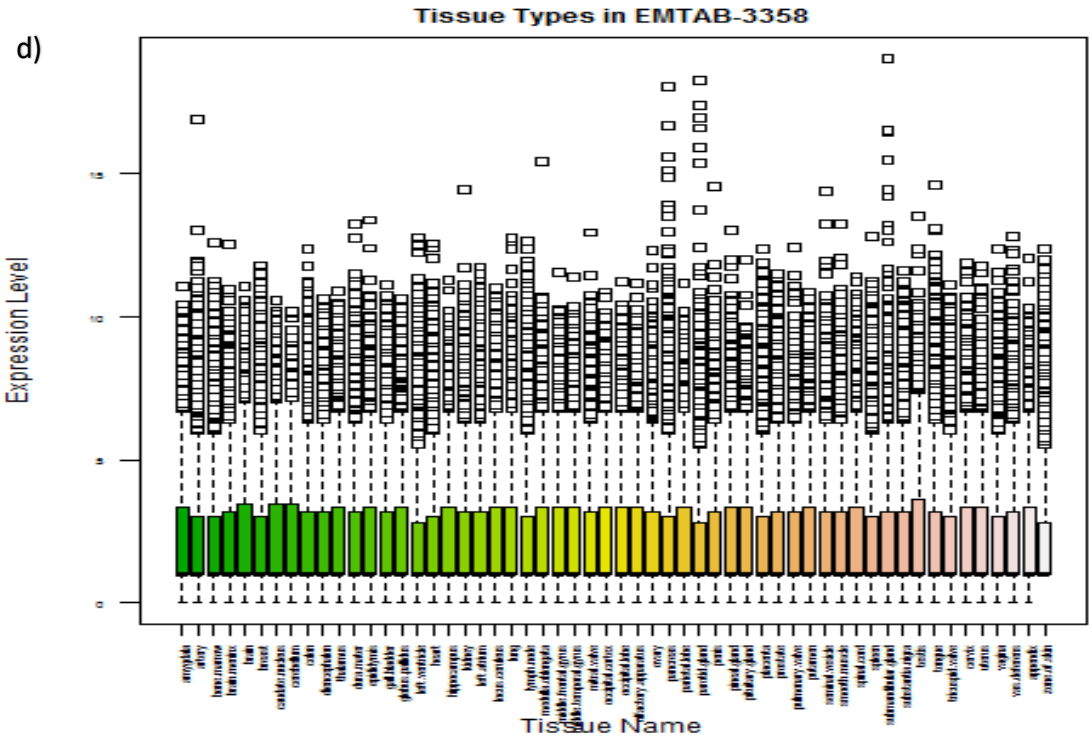
We can see distribution of expression level of all tissues for each dataset in Figure 5.2:



c)


Tissue Types in EMTAB-5214
Expression Level
Tissue Name

d)


Tissue Types in EMTAB-3358
Expression Level
Tissue Name

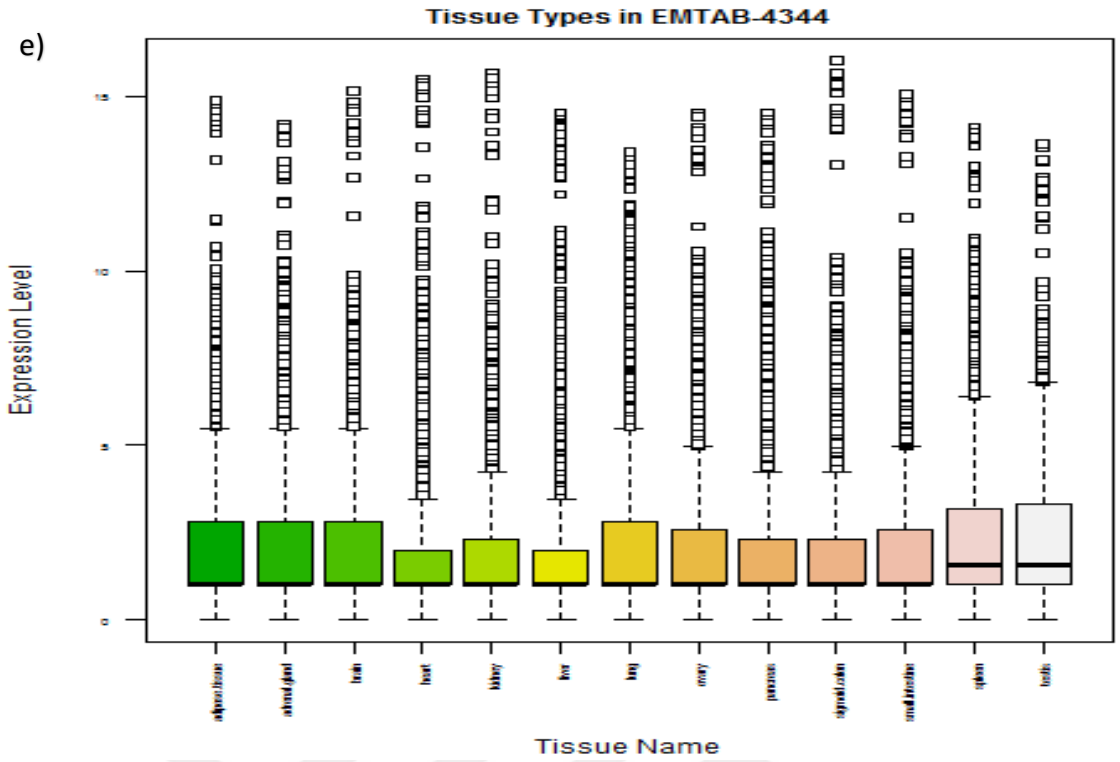


Figure 5.2 Distribution of gene expression levels for tissues in each dataset, (a) E-MTAB-1733, (b) E-MTAB- 2836, (c) E-MTAB-5214, (d) E-MTAB-3358 and (e) E-MTAB-4344, respectively.

Distributions of data according to tissue types are required to be shown like in Figure 5.2 in order to understand data and to analyze accurately.

5.1.6 Assignment of Variables

Filtered data were kept as a list of data frames called “geneData”. In this step, analyzes are carried out for each element of the list separately, and all results are merged afterwards. We identify initial variables like number of rows, number of columns, number of tissues and names of columns. Start column is automated according to the start point of numeric values. The first dataset to be analyzed is E-MTAB-1733. Following this, all datasets are analyzed and the results from each dataset is recorded using Algorithm 5.

Algorithm 5

```
geneData <- geneData$data1733
startColumn <- 0
dimData <- dim(geneData)
```

```
numRows <- dimData[1]
numCols <- dimData[2]
for (i in 1:numCols) {
  if (!is.factor(geneData[1, i])) {
    startColumn <- i
    break
  }
}
numTissue <- numCols - startColumn + 1
columnNames = colnames(geneData, do.NULL = TRUE, prefix = "")
```

5.2 Determination of Thresholds, Filtration of Data and Calculation of Tau Score

Experimental errors can be found in datasets. For this reason, we cleaned the noise in data and adjusted thresholds according to literature. Data were analyzed according to the following steps:

Step 1 - Transcripts which have low expression values are eliminated. FPKM $\leq$ 1.0 was used as first treshold [255]. It was called **"Null expression"**.

Step 2 - Tau as a tissue specificity score was calculated. Tissue specificity index, tau value is in range between 0 and 1, housekeeping genes expressed in all tissues have tau score of 0 and genes expressed strictly in one tissue have tau score of 1 [118]. Besides, it is assumed that genes with a tau score equal and bigger than 0.85 are specific to corresponding tissue. According to Kryuchkova-Mostacci and Robinson-Rechaviand [38], tau is more successful measurement than other methods. They showed that almost all tissue-specific genes found by any method are also found by tau, dramatically. Furthermore, tau appears consistently to be the most robust method and it is the best in recognizing tissue-specific genes. This was showed with the examples of brain and testis-specific genes in their study [38].

In this thesis study, after first filtration of RNA-seq data, expression levels were transformed based on log2 [38]. After log2 transformation, tau score was calculated. If tau score $\geq$ 0.85 for one gene, that gene is specific to the tissue, which has maximum

expression level. This gene was called **"Specific expression"**. Equations (3.9) and (3.10) were used for the calculation of tissue specificity.

Step 3 - If tau score $<$ 0.85, these gene profiles were classified as **"Wide-spread expression"**.

Step 4 - Genes which have expression values lower than 10.00 was described as **"Weak expression"** [95]. Tau score was calculated for these genes. However, they were not included in the specific gene calculation.

NOTE: Log transformation was used only during tau score calculation, other analyses after these steps were carried out using FPKM or TPM values. Genes were filtered using thresholds, classified using tau score. After all calculation status were added to each gene according to four steps mentioned above. All of them were analyzed with the following code, Algorithm 6:

Algorithm 6

```
det_status <- function(row_data) {
  temp_data <- as.numeric(row_data[startColumn:numCols])
  lowerThan1 <- as.numeric(sum(temp_data <= 1))
  lowerThan10 <- as.numeric(sum(temp_data < 10))
  if (identical(lowerThan1, numTissue)) {
    row_data <- cbind(t(row_data), -1, "Null expresssion")
  } else {
    temp_data[temp_data < 1] <- 1
    log2Vals <- log2(temp_data)
    sumVals <- sum(log2Vals)
    tauCount <- length(temp_data)
    maxVal <- max(log2Vals)
    tau <- (tauCount - sumVals/maxVal)/(tauCount - 1)
    if (tau < 0.85) {
      row_data <- cbind(t(row_data), tau, "Wide-spread expression")
    } else {
      if (identical(lowerThan10, numTissue)) {
```

```
            row_data <- cbind(t(row_data), tau, "Weak expression")
        } else {
            row_data <- cbind(t(row_data), tau, "Specific expression")
        }
    }
  }
  return(t(row_data))
}
analysis_result <- data.frame(matrix(0, ncol = numCols + 2, nrow = 0))
analysis_result <- rbind(analysis_result, apply(geneData, 1, det_status))
analysis_result <- data.frame(t(analysis_result))
colnames(analysis_result) <- cbind(t(colnames(geneData)), "Tau.Score", "Status")
head(analysis_result[, c("Ensembl.Gene.ID", "Status", "Tau.Score")])
```

We generated a gene list called **"included list"** according to the tau calculation and used it for further analyzes like in Algorithm 7.

Algorithm 7

```
geneMatrix <- data.frame(matrix(0, ncol = ncol(analysis_result), nrow = 0))
includedList <- as.matrix(analysis_result[analysis_result$Status == "Specific
expression",  ])
geneMatrix <- includedList[, 1:ncol(analysis_result)]
colnames(geneMatrix) <- colnames(includedList)
dimData <- dim(includedList)
numRows <- as.numeric(dimData[1])
```

If we want to save included list or all result as csv file, we can export the list:

```
write.csv(includedList, file = "includedList1733.csv")
write.csv(analysis_result, file = "analysis_result1733.csv")
includedList1733 <- includedList
analysis_result1733 <- analysis_result
```

Up to this section, genes were analyzed depending on their expression profiles. Our results after filtration and tau calculation is seen here using an example, EMTAB-1733, notice the "Status" column which summarizes filtration and tau score calculation results:

```
##   X  EnsemblGeneID   colon    kidney     liver pancreas     lung
## 1 14 ENSG00000214814 7.609656 3.4065550 0.00000000 0.011210 0.128262
## 2 16 ENSG00000180739 0.408624 0.2285900 1.09884333 0.060915 1.880716
## 3 18 ENSG00000172020 3.643640 0.1768150 0.02903667 0.420145 0.175362
## 4 19 ENSG00000100218 0.367730 5.4585200 0.13221667 0.000000 1.395074
## 5 24 ENSG00000111218 0.078320 0.3282125 0.00000000 0.309435 0.875790
## 6 35 ENSG00000136531 0.141494 2.3508125 0.05502000 0.074825 0.208450
##                Status
## 1 Wide-spread expression
## 2      Null expresssion
## 3 Wide-spread expression
## 4 Wide-spread expression
## 5 Wide-spread expression
## 6 Wide-spread expression
```

5.3 Determination of Tissue-Specific Genes

One-dimensional tissue specificity index is limited in its capacity to identify only one tissue and categorize specific classes of expression patterns. For instance, tau as a one-dimensional specificity score shows that a simple gene is specifically expressed only one tissue which has maximum expression. However, gene can have another high expression value except from maximum expression level. It is an important deficiency of tau equation that it does not demonstrate the second or third tissue which has higher expression values among all tissues. Thus, to generate a detailed and rigorous identification of tissue-specific genes, statistically significant interval from maximum expression must be defined to find out second and/or third tissue which has high expression after maximum expression value for each gene in terms of specificity.

To achieve significant interval estimation firstly we examined confidence interval [256], [257] to understand significant interval from average of all data. Confidence interval and

also other statistical interval estimation methods may be useful for generating a newly method to assign genes to several tissues which are specifically expressed. In order to use this approach, data must include independent random variables. Variables are identically distributed, and they are often normally distributed. These are usual assumptions for this approach. Confidence interval of nearly normally distributed data can be estimated using this simple and basic equation [258]:

$$\textbf{Average of data} \pm \textbf{standard deviation}$$

As a detailed description for confidence interval from average of all data is defined as [259]:

$$x \pm z^* \frac{\sigma}{\sqrt{n}} \tag{5.4}$$

- X is the mean of data
- Z represents the z-value from the standard normal distribution
- σ is standard deviation of sample
- n is the sample size

Confidence interval can be at 90%, 95% and 99% for obtaining meaningful and more reliable results. Z-value of some specific confidence are in Table 5.5:

Table 5.5 Z–values for various confidence levels

Confidence Level	**Z-value**
80%	1.28
90%	1.645
95%	1.96
98%	2.33
99%	2.58

For instance, if confidence level is 95% shown in Figure 5.3, equation of reliable interval estimation is: $\textbf{mean of data} \pm \textbf{standard.deviation} * \mathbf{1.96}$

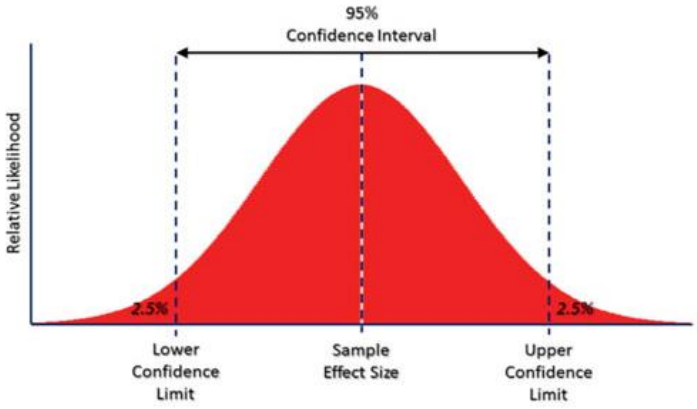


Figure 5.3 95% of confidence interval estimation [260]

Statistical intervals provide invaluable tools for quantifying sampling uncertainty. Intervals such as prediction intervals, tolerance intervals and confidence intervals on distribution quantiles are frequently needed in practice [261]. Improved computational approaches are important for bioinformatic analysis and statistical intervals can be used in bioinformatic. Estimation of statistical interval is crucial for reliability of results. However, we need a reliable significant interval from maximum expression value instead of confidence interval from mean of data for each gene in order to assign genes to multiple tissues in the context of specificity. **If we can determine a statistically significant distance from the maximum expression, we can also show other tissues that related gene is specifically or restrictedly expressed besides the maximum expression.**

Although tissue-specific genes are expressed in one or several tissues/cell types as definition, tau score shows that one gene is specific to only one tissue. If tau is 1, one gene is specific to one tissue, however if $0.85 \leq tau < 1$, gene may be specific to more than one tissue according to our examination. Even in this case, tau shows us only one tissue and it is not comprehensive and rigorous approach. We know that tissue-specific genes are defined as “genes specifically expressed in one or a few tissue or cell types”. To overcome this ambiguity problem, after the examination of genes via tau because of the best specificity score, we tried to develop a new procedure that assign genes to specifically expressed multiple tissues. In other words, we defined a new approach to estimate gene specifically expressed one or more tissues. Important deficiency of tau score and newly generated statistical approach are illustrated in Figure 5.4:

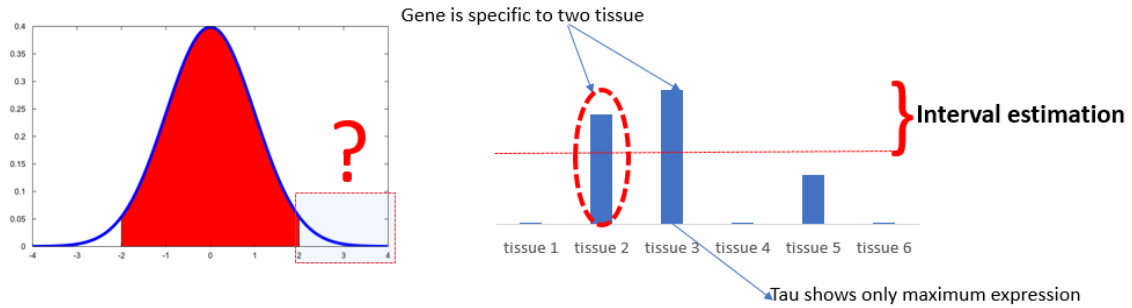


Figure 5.4 Illustration of statistically significantly interval

Tau is a robust specificity score but not rigorous. For this reason, in this part of the study, we aimed to develop a new approach adding some statistical calculations as demonstrated in Figure 5.4. For this purpose, we tried to design a new algorithm to determine tissue-specific genes more accurately and comprehensively.

After all examination and research about statistical interval in literature, our new approach is based on statistically significant interval estimation from maximum value. We should pay attention to that point, it is different from confidence interval because we aim to define statistically estimation of interval from maximum expression of all genes in each dataset. Thereby, after tau calculation we used statistically significant interval estimation from maximum expression values and our purpose is to assign genes to multiple tissues. In this way, all tissues in which genes are specifically expressed can be detected. Firstly, we used a basic equation below to estimate the significant distance from maximum value of expression:

$$\textbf{Maximum expression of a single gene} - \textbf{standard deviation}$$

A threshold value is required to obtain more accurate results in high sensitivity. Z-value was used as sensitive threshold to increase sensitivity of calculation and equation becomes:

$$\textbf{Maximum expression of a single gene} - \textbf{standard deviation} * \textbf{Z\_value}$$

Statistical distance is a broad topic which plays a fundamental role in statistics, machine learning and in the associated scientific disciplines. This method can be invaluable in bioinformatic analyses. Statistical distance has robustness role in analyses and it is also prominent in goodness-of-fit. Hence, statistical distance measures play a ubiquitous role in statistical theory and thinking [262] . It might be a good approach to reveal tissue-

specific genes comprehensively and rigorously. Figure 5.4 illustrates the concept of statistical distance from maximum expression value.

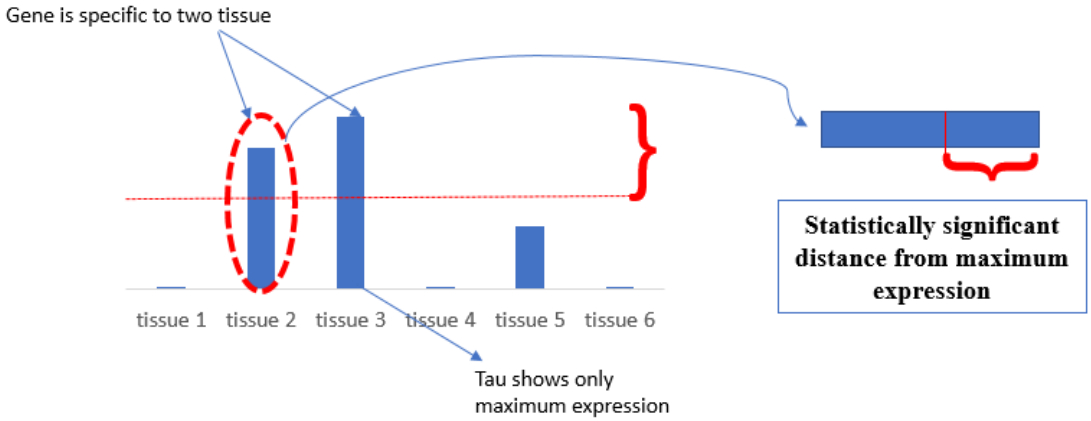


Figure 5.5 Estimation of statistically significant interval from maximum expression

The determination of tissue-specific genes is a process similar to cleaning the noise in data analysis because expression level of genes in a single tissue is clearly higher than others. Tissue-specific genes increase the entropy in the data as the perspective of data analysis. While the usage of Shannon entropy in the detection of tissue-specific genes has also been used [112], developments of dynamic systems such as artificial intelligence (AI) for rapid progression, deep learning, machine learning have become more prevalent, effective and easy, especially in the analyzes performed on biological systems. In other words, new developments in artificial intelligence show that an unsupervised learning mechanism can adapt a dynamic environment. Besides, these methods can help to provide clear and exact results with high accuracy in biological data. Hence, this thesis study aims to develop a new algorithm to calculate tissue-specific genes, and this system is a dynamic system that provides continuity. Tissue specificity calculated using Shannon entropy measures the degree of tissue specificity of a gene; but it does not determine whether it is specific to a particular tissue or specific to which tissue, and it is insufficient to robust and comprehensive calculation of tissue-specific genes [112].

In this study, the second stage of subsequent tau calculation is matching genes to multiple tissues which are specifically expressed in related tissues. Within this purpose, as mentioned above, a statistically significant distance from the maximum expression value was calculated. **Integration of tau calculation and estimation of statistically**

significant interval from maximum expression was described as "extended tau" calculation for tissue specificity in this study.

Determination of tissue-specific genes rigorously after only tau calculation were carried out some significant steps given below:

- To define lower and upper bound via clustering all raw data,
- To calculate upper level threshold as Z-value because our important point in this estimation maximum expression value and significant distance from it for each gene,
- To determine significant distance to identify several tissues for specificity.

5.3.1 Clustering of Raw Data

Firstly, clustering of the data, the centers of the clusters and the visibility of the clusters in the entire data need to be computed in each dataset. Each dataset was basically divided into two clusters. For determination of the boundary between high level and low-level expression, Fuzzy c-means (FCM) clustering [263] has been successfully applied to feature analysis, clustering, and classifier designs in biological data analysis [264]. FCM is a successful method for biological datasets [265]. FCM algorithm classifies the data by grouping similar expression values according to the feature space into clusters. This clustering is achieved by assigning membership to each expression point corresponding to each cluster center based on distance between the cluster center and the data point, iteratively [266], [267].

Because of the determination of the specificity limit and understanding of how many genes within the upper cluster is involved in the specific cluster, data were divided into two clusters using c-means clustering. Centers are midpoints of top and bottom clusters. We obtained as many centers as the number of tissues and genes in each dataset and then calculated the average of centers. This code in Algorithm 8 was used for c- means clustering:

Algorithm 8

```
library(e1071)
clusters<-cmeans(includedList[, startColumn:numCols], 2, iter.max = 100, verbose = FA
LSE,
  dist = "euclidean", method = "cmeans", m = 2, rate.par = e - 10)
centerData <- clusters["centers"]
clusterTreshold <- data.frame(t(as.matrix(colMeans(centerData$centers, na.rm = TRU
E))))
```

5.3.2 Calculation of Upper Level Threshold

To determine the threshold value of the upper cluster that obtained from c-means clustering, the necessary variables were assigned. After that, column names were set to be the same as the raw data.

NOTE: To increase sensitivity the analysis we did not use expression "0". Number of elements in upper side cluster and total non-zero elements (higher than 0) were calculated for each column or each tissue. Then the ratio of upper side cluster memberships was calculated.

$$\textbf{threshold ratio} = \frac{\textbf{number of element in upper cluster}}{\textbf{number of total non} - \textbf{zero elements}}$$

All calculations to obtain threshold ratios for each tissue type Algorithm 9 was performed:

Algorithm 9

```
elementCount <- data.frame(matrix(0, ncol = numTissue, nrow = 1))
tresholdRatio <- data.frame(matrix(0, ncol = numTissue, nrow = 1))
nonZeroElementCount <- data.frame(matrix(0, ncol = numTissue, nrow = 1))

elementCount <- setNames(elementCount, columnNames[startColumn:numCols])
tresholdRatio <- setNames(tresholdRatio, columnNames[startColumn:numCols])
nonZeroElementCount <- setNames(nonZeroElementCount,
columnNames[startColumn:numCols])
```

```
for (i in 1:numRows) {
  for (j in startColumn:numCols) {
    if (as.numeric(includedList[i, j]) >= clusterTreshold[1, columnNames[j]]) {
      elementCount[1, columnNames[j]] <- elementCount[1, columnNames[j]] + 1
    if (as.numeric(includedList[i, j]) > 0) {
      nonZeroElementCount[1, columnNames[j]] <- nonZeroElementCount[1,
columnNames[j]] + 1
    }
  }
}
for (j in startColumn:numCols) {
  tresholdRatio[1, columnNames[j]] <- elementCount[1,
columnNames[j]]/nonZeroElementCount[1, columnNames[j]]
}

data_center <- as.data.frame(unname(centerData))
combine <- rbind(data_center, clusterTreshold, elementCount, nonZeroElementCount,
tresholdRatio)
row.names(combine) <- c("upper cluster", "lower cluster", "average of clusters",
  "number of upper cluster element", "number of non-zero element", "treshold ratio")
options(scipen = 10)
write.csv(combine, "combine1733.csv")
```

The values obtained from above calculation were summarized as tables for each dataset. And resulting tables can be found Appendix-A. There are ratios as much as the number of tissues, thus we used regression analysis to minimize the differences among them.

5.3.3 Regression Analysis

We used Regression analysis (RA) to minimize the differences among all ratios of columns for each dataset. RA observes the relationship between one dependent variable and several independent variables. RA was defined in 1966 as a method producing linear regression equations by constructing new explanatory variables or elements using linear combinations of the original variables. The method can be

effectively applied to the settings where the number of explanatory variables is very large [268].

When a frequency distribution is normally distributed, we can find out the probability of occurrence of a score by standardizing the scores, known as z-scores. A z-score is also known as a standard score and it can be placed on a normal distribution curve. Z-scores range from -3 standard deviations which would fall to the far left of the normal distribution curve, up to +3 standard deviations which would fall to the far right of the normal distribution curve.

Qnorm (inverse of pnorm, **qnorm (p, mean, sd)**) is used to determine what the z-score of the pth quantile of the normal distribution is. The probability in the normal distribution table can be at most 0.9999. To calculate the threshold value as a z-score we used qnorm(optimizedTreshold). It is not enough in order to determine optimal threshold. We wanted to find the expression of top 1% cluster to assign genes to multiple tissues. For this reason, we have to find z-score of 99% cluster. We calculated qnorm (0.9999) and qnorm(optimizedTreshold) and then qnorm(0.9999) must be subtracted from qnorm (optimizedTreshold). By this way, we can find **"threshold value as a z-value"**. We calculated the optimal threshold value with the Algorithm 10:

Algorithm 10

```
x <- as.matrix(1:numTissue)
regressionResult <- lm(unlist(tresholdRatio) ~ x)
coeffs <- coefficients(regressionResult)
optimizedTreshold <- mean(x * coeffs[2] + coeffs[1])
tresholdVal <- qnorm(0.9999) - abs(qnorm(optimizedTreshold))
treshold4344 <- data.frame(c(optimizedTreshold, tresholdVal), row.names =
c("optimized treshold", "treshold value"))
```

All steps to generate sensitive threshold as z-value in order to estimate significant interval from maximum value were summarized in Figure 5.5:

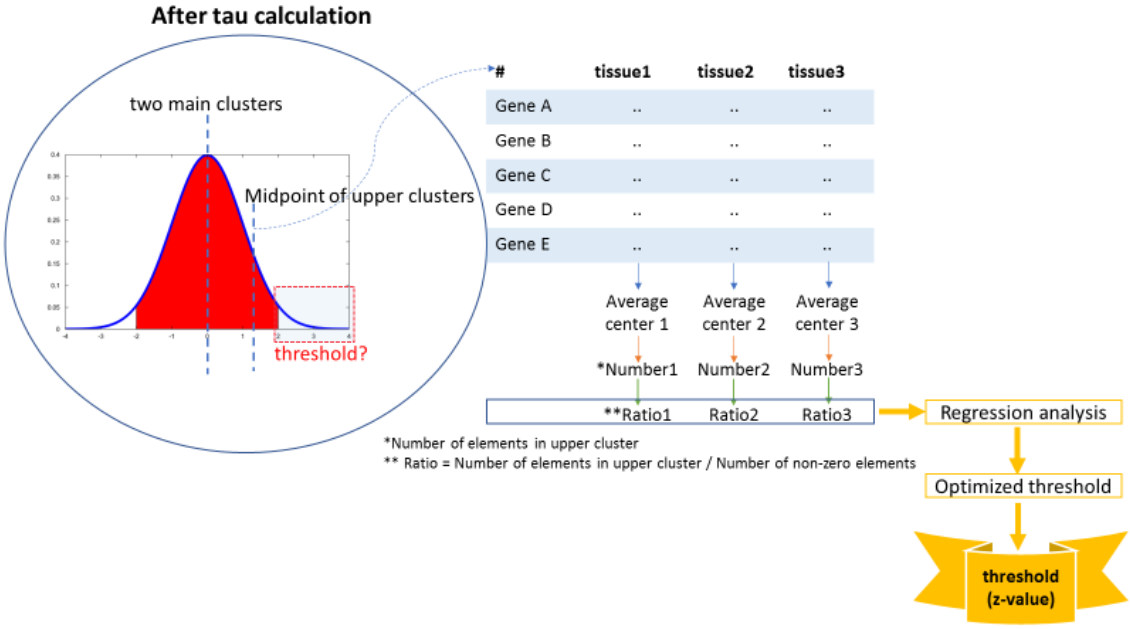


Figure 5.6 Obtaining thresholds for each dataset to calculate significant interval

After all steps like in Figure 5.6, final threshold values for each dataset was obtained as Z-value according to nearly normally distribution of raw data. Optimized threshold after regression analysis and Z-values are shown in Table 5.5:

Table 5.6 All threshold ratios for each dataset

	EMTAB-1733	**EMTAB-2836**	**EMTAB-5214**	**EMTAB-3358**	**EMTAB-4344**
optimized threshold	0.10	0.16	0.16	0.43	0.15
Z- value as threshold	2.46	2.72	2.73	3.54	2.68

In order to establish statistical distance, threshold value is needed for each dataset. Threshold values are comparable, expect only one which is far from others. EMTAB-3358 has threshold value of 3.54 which is far from other datasets and the reason is that the normalization method of EMTAB-3358 is different from others.

5.4 Calculation of Statistical Distance and Assigning Genes to Multiple Tissues

Although tau is the best method, it assigns a single gene to only one tissue. We would like to demonstrate that one gene can be specific to only one or several tissues. Therefore, we used ***statistically significant distance estimation*** term. Figure 5.6 demonstrate that equation of estimation of statistically significant interval.

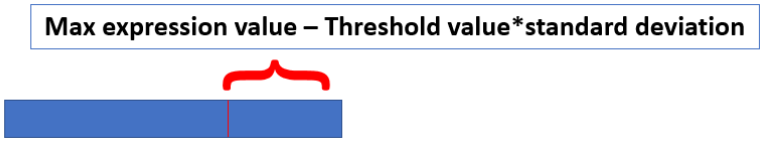


Figure 5.7 Determination of significant distance from maximum expression

We calculated the value of the significant distance and assigned genes to tissues as specifically expressed if gene's expression is above significant distance value using Algorithm 11. In other words, a gene is specific to a tissue if its expression value in that tissue is in the interval between maximum expression and statistically significant distance.

Algorithm 11

```
fileName <- "E-MTAB-1733-query-results.txt"
geneTissuePair <- data.frame(matrix(0, ncol = 4, nrow = 0), row.names = NULL)
colnames(geneTissuePair) <- c("gene", "tissue", "expressionLevel", "file")
for (i in 1:numRows) {
  row_data <- includedList[i, ]
  temp_data <- as.numeric(row_data[startColumn:numCols])
  len_temp <- numCols - startColumn + 1
  stdDev <- sd(temp_data)
  maxVal <- max(temp_data)
  for (j in 1:len_temp) {
    if (as.numeric(temp_data[j]) >= as.numeric(maxVal - stdDev * tresholdVal)) {
      geneTissuePair <- rbind(geneTissuePair, data.frame(row.names = NULL,
        gene = row_data[1], tissue = columnNames[j + startColumn - 1],
expressionLevel = temp_data[j], file = fileName))
    }
  }
}
geneTissuePair1733 <- geneTissuePair
```

We prefer to **assign genes to multiple tissues** as a table. Then, gene tissue pairs and/or multiple tissues can be exported as csv files for each dataset according to Algorithm 12:

Algorithm 12

```
library(reshape)
multiple_tissue <- aggregate(tissue ~ gene, data = geneTissuePair, paste, collapse = ",")
head(multiple_tissue)
multiple_tissue1733 <- multiple_tissue

write.csv(geneTissuePair, file = "geneTissuePair1733.csv")
write.csv(multiple_tissue, file = "multiple_tissue1733.csv")
```

Up to now, tissue-specific genes were determined as lists for each of the five datasets using a novel, robust and rigorous approach. Firstly, we calculate tau score for each gene and we obtain one tissue that has maximum expression of related gene in terms of specificity. Secondly, we generate a new statistical approach to determine significant interval from maximum expression in order to find out several tissues have specific expression corresponding gene. This integrated method was defined as **"extended tau"** calculation for tissue specificity in this thesis study.

5.5 Kolmogrov Symirnov Test and Q-Q Plot

There are several methods of assessing whether data are normally distributed or not. They are divided into two main categories:

- statistical tests
- graphical tests

If we want to know our data fits normal distribution, we can use **Kolmogrov Symirnow test** as a statistical test. And also if we want to visualize the distribution we can use Q-Q plot as graphical test [269], [270].

The Kolmogorov–Smirnov (K–S) [271] is a goodness-of-fit test that compares a hypothetical or fitted cumulative distribution function (cdf, used pnorm) with an empirical cdf to assess the fit. Other definition is that K-S test is based on the empirical distribution function. Given N ordered data points Y1, Y2, ..., YN, the empirical cdf is defined as:

$$E_N = n(i)/N \tag{5.5}$$

- n (i) is defined as the number of points less than Yi (Yi are ordered from smallest to largest value).
- N is defined as ordered data points Y1, Y2, … YN.

The Kolmogorov-Smirnov test statistic is defined as:

$$D = \max_{1 \le i \le N} (F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - (F(Y_i))) \quad (5.6)$$

- D is test statistics,
- F is the theoretical cumulative distribution of the distribution being tested [272], [273].

If D is greater than the critical value obtained from K-S tables, the hypothesis regarding the distributional form is rejected. In addition to D, K-S test statistics, if p value > 0.05, there is no significant differences between actual and theoretical distribution, so data fits the normal distribution.

After that, for each mean and standard deviation, the combinations of theoretical normal distribution can be determined. This theoretical normal distribution is compared to the actual distribution of the data using K-S test like in Algorithm 13:

Algorithm 13

```
dataCount <- sum(geneMatrix > -1, na.rm = TRUE)
allTestData <- data.frame(matrix(0, ncol = 2, nrow = dataCount))
allDataCursor <- 1
testResult <- data.frame(matrix(0, ncol = 2, nrow = numRows))
colnames(testResult) = c("D.value", "P.value")
cursor <- 1
for (i in 1:numRows) {
  rowData <- geneMatrix[i, startColumn:numTissue]
  rowData <- as.numeric(rowData[rowData > -1])
  rMean <- mean(rowData)
  rDev <- sd(rowData)
  tCount <- length(rowData)
```

```
testData <- data.frame(matrix(0, ncol = 2, nrow = tCount))
for (j in 1:tCount) {
  val <- rowData[j]
  actualProbability <- length(rowData[rowData <= val])/tCount
  theoreticalProbability <-pnorm(q = val,mean = rMean, sd = rDev,lower.tail = TRUE)
  testData[j, 1] <- actualProbability
  testData[j, 2] <- theoreticalProbability
  allTestData[allDataCursor, 1] <- actualProbability
  allTestData[allDataCursor, 2] <- theoreticalProbability
  allDataCursor <- allDataCursor + 1 }
res <- ks.test(testData[, 1], testData[, 2])
testResult[cursor, 1] <- as.numeric(res["statistic"])
testResult[cursor, 2] <- as.numeric(res["p.value"])
cursor <- cursor + 1 }
```

K-S test result were saved as csv files for each dataset.

Q-Q plots display the observed values against normally distributed data (represented by the line). If distribution of data is closer to linear line it fits to the normal distribution. Algorithm 14 was used to generate Q-Q plots of all dataset used in this thesis study.

Algorithm 14

```
calculated <- allTestData[, 1]
expected <- allTestData[, 2]
qqplot(calculated, expected, plot.it = TRUE, xlab = "Theoretical Quantiles", ylab =
"Sample Quantiles")
title("Expression Values Distribution")
qqline(expected)
dev.copy(png, "qq_1733.png")
dev.off()
```

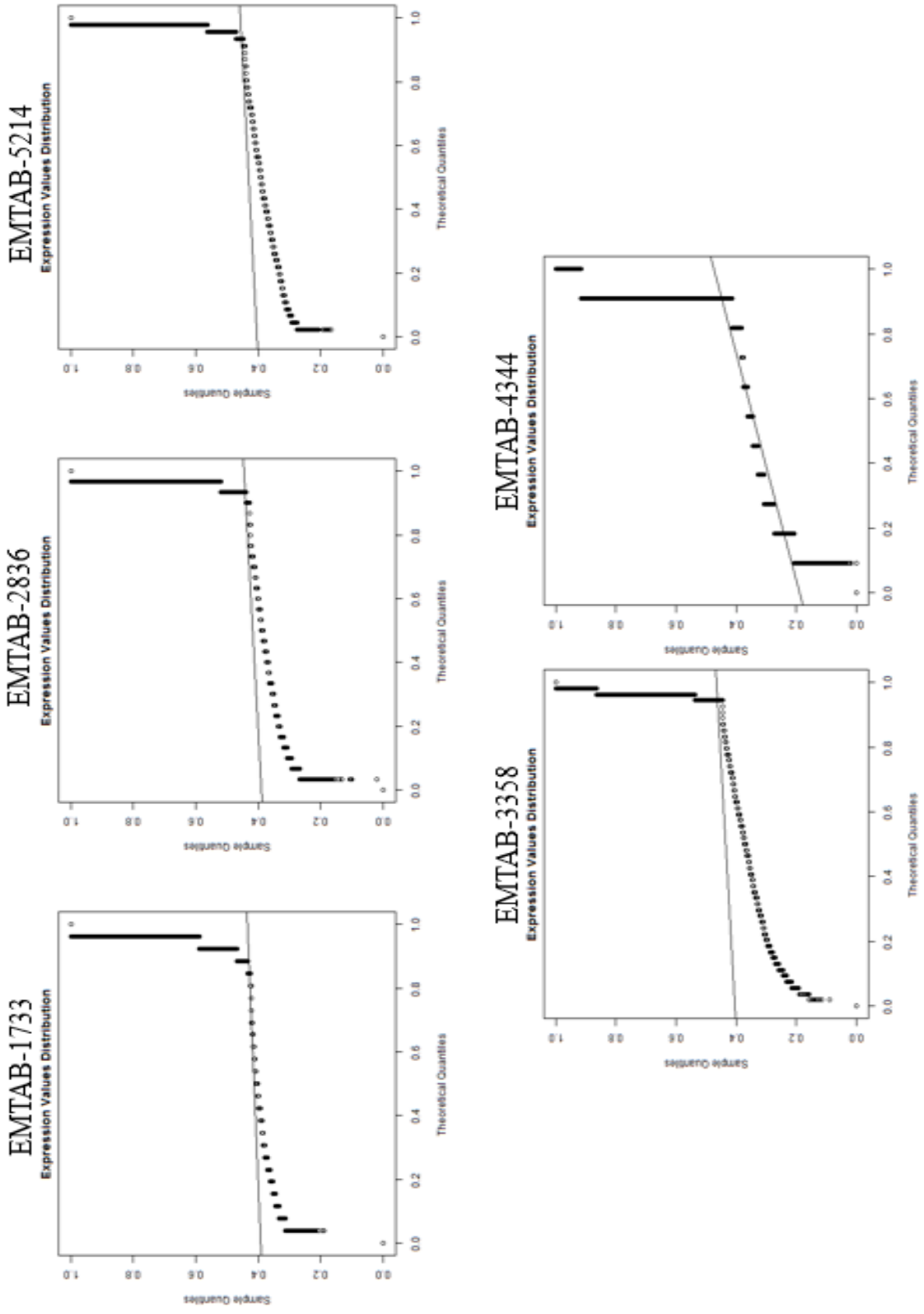


Figure 5.8 Distribution of expressions for all genes in each dataset, separately.

If distribution is near the line, it is normally distributed. According to Figure 5.5, the most appropriate distribution belongs to EMTAB-4344 followed by EMTAB-1733. EMTAB-3358 is less suitable to normal distribution than other datasets.

5.6 Comparison of Results with The Human Protein Atlas Data

The Human Protein Atlas [274] consists of three different datasets. These are Tissue Atlas, Cell Atlas and Pathology Atlas. We selected Tissue Atlas to compare our tissue specific expression results. Tissue Atlas contains information regarding the expression profiles of human genes on the protein level. The protein expression data is derived from antibody-based protein profiling via immunohistochemistry experiments. 76 different cell types corresponding to 44 normal human tissue types have been analyzed to determine their protein levels for each gene and data are presented as knowledge base annotation of protein expression levels.

Protein levels as a tsv file format were obtained from Tissue Atlas database. Then, tissue specific genes assigning with multiple tissues (described as “joined_genepair” in algorithms) file as a csv format was generated during the thesis study after the calculation of tissue-specific genes. It contains genes which are specific to all tissues including child tissues and they are shown according to all five datasets with their expression level and their tau score.

Tissue Atlas contains gene IDs as Ensembl Gene ID, gene names, tissues and cell types within related tissues, protein expression levels as "High", "Medium", "Low" and "Not Detected" and reliability of data as "Approved", "Supported" and "Uncertain". We filtered data according to their reliability which is defined as approved and/or supported. After that, two different data which are protein data and thesis results were joined to generate a new table. After generating the new table, the joined table was integrated parent tissue information to understand each child tissue using Algorithm 15:

Algorithm 15

```
library(tidyverse)
library(stringr)
library(readr)
```

```
protein_atlas<-read_tsv(file="normal_tissue.tsv")
joined_genepair<-read_csv("joined_genepair.csv")
tissues<-read.csv(file="tissue list for database.csv",sep=";",stringsAsFactors = FALSE)
%>%
 rename(parents=Group.name)  %>%
 rename(child=Tissue.name)
 protein_atlas_joined<-protein_atlas %>%
 filter(Reliability!="Uncertain") %>%
 mutate(tissue=ifelse(Tissue=="cervix, uterine","uterus",Tissue),
        tissue=str_replace(tissue," [0-9]","")) %>%
 mutate(tissue=recode(tissue,"caudate"="caudate nucleus","fallopian
tube"="oviduct","gallbladder"="gall  bladder","thyroid  gland"="thyroid"))  %>%
 select(-Tissue) %>%
 select(1,2,6,3,4,5) %>%
 full_join(tissues,by=c("tissue"="child")) %>%
 filter(!is.na(Level),!is.na(parents)) %>%
 inner_join(joined_genepair,by=c("tissue","Gene"="Ensembl.Gene.ID"))
```

Data were counted to show that RNA-seq expression is less than 10 for each parent tissue with Algorithm 16:

Algorithm 16

```
 protein_atlas_joined %>%
 filter(Level!="Not detected") %>%
 select(1,3,5,11,13) %>%
 arrange(Gene)


protein_atlas_joined %>%
 filter(expressionLevel<10) %>%
 select(1,8) %>%
 group_by(parents) %>%
 summarise(n=n_distinct(Gene)) %>%
 arrange(-n)
```

In addition to this comparison, some graphs were generated in this study to understand the correlation of results among five datasets. If similar results are obtained when applying this new method to different data sets, the new approach can be called universal. We exhibited detailed graphs via R programming packages in order to show advantage and disadvantages of our new computational approach in the context of tissue specificity. They were evaluated in Results and Discussion Section. The code of a detailed graph is given in Appendix-A Section as an example. Moreover, obtained specific genes have been examined in terms of their functions using an annotation tool which is going to be explained in Section 6.4.

This part of the thesis study,

- Raw data were obtained from five big datasets which are commonly used for bioinformatic analysis.
- Datasets was examined using F-test and boxplots.
- Tissue specific-genes were confirmed via a new, dynamic, practice, robust and rigorous combined approach that was called “extended tau” calculation. According to extended tau approach, one gene can be specific to more than one tissue.
- K-S test and Q-Q plots were carried out to show the normal distribution of datasets.
- Raw protein data and RNA-seq and also tissue specificity results were compared with The Human Protein Atlas results.
- Advanced graphs were plotted for the comparison of results among datasets.

All steps are explained as a flow chart:

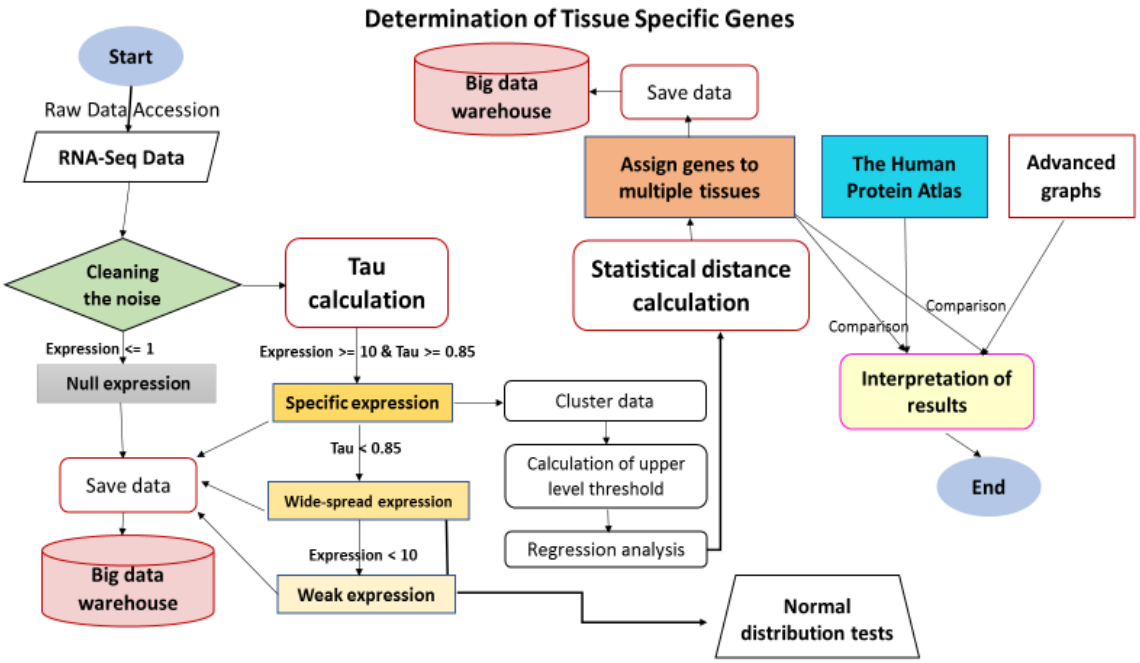


Figure 5.9 Flow chart of the whole algorithm to calculate tissue-specific genes

CHAPTER 6

METHODS

PART 2: PREPARATION OF CANCER EXPRESSION DATA AND INTERPRETATION OF TUMORAL HETEROGENEITY AND CANCER CELL BEHAVIOUR

6.1 Differentially Expressed Genes in Various Cancers

Differentially Expressed Genes (DEG) between healthy tissue and cancerous tissue are important to understand cancer development, metastasis and treatment. DEG lists for various cancers were retrieved from **Bioexpress** [275]. High Performance Integrated Virtual Environment (HIVE) is a tool and database optimized for the storage of NGS data for medical researches. Bioexpress is part of HIVE to store transcript data. We can download differentially expressed genes for various cancer types (source is TCGA, The Cancer Genome Atlas) as a csv file. This file includes a lot of information such as gene Uniprot ID, gene name, references gene, fold change, p-value, significance, cancer type, experiment type. We used DEG in each cancer type generated from Bioexpress through RNA-seq data.

Firstly, this list is imported to R.

```
degList <- read.table("BioXpress_gene_differential_expression_v2.0.txt", header = TRUE, sep = "\t")
```

- **Obtained list from Bioexpress has 374 100 rows for approximately 18 different cancer types and approximately 20 000 different protein coding genes.**

Then, DEG list was filtered according to some parameters. We wanted to throw away the **NA (not available)** from **log2FoldChange** column. Therefore, we used is.na function to eliminate NA as shown below:

```
degList <- degList[!is.na(degList$log2FoldChange), ]
```

- **After filtration of "Not Available" data, 373 976 rows were left for all cancer types.**

We eliminated "**No"** from **Significant** column. Yes and No indicate whether the change in expression is meaningful. If significance of one gene is "Yes", this gene differentially expressed in cancer (upregulated or down regulated).

```
degList_yes <- degList[degList$Significant == "Yes", ]
```

- **If we filtered only genes that change significantly in cancer we had 100 269 gene for all cancers.**

We removed some columns, because there are not necessary to determine differentially expressed genes in cancers.

```
drops <- c("RefSeq", "p_value", "adjusted.p_value", "Patients", "Data.Source", "PMID")
degList_yes_filtercolumn <- degList_yes[, !(names(degList_yes) %in% drops)]
```

Fold change (FC) is very important to identify differentially expressed genes. We examined lots of research articles to adjust fold change value. According to Yang et al. [172], Pan et al. [276] and Corley et al [277], FC is set to equal and higher than 2 as a cut off value to screen out DEG. However, in this list there are log2 fold change values for all genes. Thus, threshold of FC must be higher than 1 because of log2. Another important point is the fact that some genes are upregulated, whereas some of them are down regulated. Thus, cut off values of FC have to be **equal and higher than 1** and **equal and lower than -1**.

Firstly, we change the log2FoldChange column from character to numeric numbers. Secondly, we eliminated the list based on FC values.

```
degList_yes_filtercolumn[, "log2FoldChange"] <- sapply(degList_yes_filtercolumn[,
  "log2FoldChange"], as.numeric)
degList_finish <-
```

```
rbind(degList_yes_filtercolumn[degList_yes_filtercolumn$log2FoldChange > 1, ],
degList_yes_filtercolumn[degList_yes_filtercolumn$log2FoldChange < (-1), ])
```

- **There are 38 532 genes in final DEG list after filtration according to log2(FC) for all cancer types.**

If we want to show unique cancer types according to **Cancer Ontology**:

```
degList_finish_unique <- data.frame(unique(degList_finish$Cancer.Ontology,
incomparables = FALSE))
View(degList_finish_unique)
```

In this list, there are 18 different cancer types, as we can see in Table 6.1 below:

Table 6.1 Cancer types obtained from Bioexpress and their accession ID

Number	Cancer type
1	DOID:5041/ Esophageal cancer [EC]
2	DOID:1612/ Breast cancer [BRCA]
3	DOID:4362/ Cervical cancer [Cerca] & DOID:3744 / Cervical squamous cell carcinoma & DOID:0050940 / Endocervical adenocarcinoma
4	DOID:11934/ Head and neck cancer [H&NC]
5	DOID:263/ Kidney cancer [Kidca] & DOID:4465 / Papillary renal cell carcinoma
6	DOID:263/ Kidney cancer [Kidca] & DOID:4467 / Renal clear cell carcinoma
7	DOID:263/ Kidney cancer [Kidca] & DOID:4471 / Chromophobe adenocarcinoma
8	DOID:3571/ Liver cancer [Livca]
9	DOID:1324/ Lung cancer [Lunca] & DOID:3907 / Lung squamous cell carcinoma
10	DOID:1324/ Lung cancer [Lunca] & DOID:3910 / Lung adenocarcinoma
11	DOID:1793 / Pancreatic cancer [PACA] & DOID:4074 / Pancreas adenocarcinoma
12	DOID:10283/ Prostate cancer [PCa]
13	DOID:1993/ Rectum cancer [Recca] & DOID:1996 / Rectum adenocarcinoma
14	DOID:10534/ Stomach cancer [Stoca]
15	DOID:1781/ Thyroid cancer [Thyca]
16	DOID:11054/ Urinary bladder cancer [UBC]
17	DOID:363/ Uterine cancer [Uteca]
18	DOID:1115/ Sarcoma

The list of healthy tissues compiled from the five RNA-Seq datasets did not contain tissues related to Sarcoma and head and neck cancer, directly, thus were removed from the list.

```
degList_finish<-degList_finish[degList_finish$Cancer.Ontology != "DOID:1115 / Sarcoma, ]degList_finish <- degList_finish[degList_finish$Cancer.Ontology != "DOID:11934/ Head and neck cancer [H&NC]", ]
```

- **Now, there are 37 871 DEG for 16 different cancer types.**

If this gene list is examined in BioMart according to their gene type, we show that they are protein coding genes. Therefore, it is suitable for intersecting with the tissue-specific genes. DEG list can export and save as csv file:

```
write.csv(degList_finish, file = "DEG_list.csv")
```

- We have a unique tissue list and when we examined this list, cancer type can be matched to healthy tissues to interpret tumoral heterogeneity.
- Healthy tissue-specific genes and related cancer type can be intersected to interpret cancer microenvironment and heterogeneity. For instance: liver cancer and healthy liver-specific genes.
- After these analyses, overlapped genes were gathered, only DEG for related cancer and only specific-gene for related healthy tissues.

Cancer type - tissue type matching list is shown below:

Cancer type → Normal tissue

DOID:1612 / Breast cancer [BRCA] → **breast**

DOID:4362 / Cervical cancer [Cerca] & DOID:3744 / Cervical squamous cell carcinoma & DOID:0050940 / Endocervical adenocarcinoma → **cervix, ectocervix, endocervix**

DOID:5041 / Esophageal cancer [EC]→ **esophagus, esophagus muscularis mucosa, gastroesophageal junction, esophagus mucosa**

DOID:263 / Kidney cancer [Kidca] & DOID:4465 / Papillary renal cell carcinoma→ **kidney, cortex of kidney**

DOID:263 / Kidney cancer [Kidca] & DOID:4467 / Renal clear cell carcinoma→ **kidney, cortex of kidney**

DOID:263 / Kidney cancer [Kidca] & DOID:4471 / Chromophobe adenocarcinoma> **kidney, cortex of kidney**

DOID:3571 / Liver cancer [Livca→ **liver**

DOID:1324 / Lung cancer [Lunca] & DOID:3907 / Lung squamous cell carcinoma→ **lung**

DOID:1324 / Lung cancer [Lunca] & DOID:3910 / Lung adenocarcinoma→ **lung**

DOID:1793 / Pancreatic cancer [PACA] & DOID:4074 / Pancreas adenocarcinoma→ **pancreas**

DOID:10283 / Prostate cancer [PCa]→ **prostate**

DOID:1993 / Rectum cancer [Recca] & DOID:1996 / Rectum adenocarcinoma→ **rectum**

DOID:10534 / Stomach cancer [Stoca]→ **stomach**

DOID:1781 / Thyroid cancer [Thyca]→ **thyroid**

DOID:11054 / Urinary bladder cancer [UBC]→ **urinary bladder**

DOID:363 / Uterine cancer [Uteca]→ **uterus, endometrium**

We generated gene tissue pair using 96-tissues that are called as child-tissue. If we want to analysis all tissues, we can use **"all_geneTissuePair"** file. But if we want to analysis only parent tissues, we can use **"only_parent_geneTissuePair"** file. The storage of these two files is important for further analyzes.

6.2 ID Conversion Process

Genes in DEG list has Uniprot ID. On the other side, tissue-specific gene lists have Ensembl gene ID. In that case, we have to convert ID of DEG from Uniprot to Ensembl gene ID. BioMart was used to download the conversion table by "biomaRt" package in R.

Firstly, BioMart was called from R library.

```
# source('https://bioconductor.org/biocLite.R') biocLite('biomaRt')
library("biomaRt")
```

Homo Sapiens GrcH 37 version data was selected.

```
ensembl = useEnsembl(biomart = "ensembl",dataset ="hsapiens_gene_ensembl",
GRCh =37)
```

Attributes were described as a vector. And then genes were filtered according to “gene type”, and “protein coding genes” section was selected. BioMart can give maximum three external IDs via R in a single call.

```
attributes_total <- c("ensembl_gene_id", "gene_biotype", "external_gene_name",
"chromosome_name", "start_position", "end_position", "uniprot_gn",
"uniprotswissprot", "uniprotsptrembl")
conversion <- getBM(attributes_total, filters = "biotype", values = "protein_coding",
mart = ensembl)
knitr::kable(head(conversion))
conversion <- write.csv(conversion, file = "conversion.csv")
```

We received a list that contained ensemble gene id, gene biotype, external gene name, chromosome name, start position, end position, uniport gen id, uniport swissprot ID, uniprotsp trembl ID. Uniprot swissprot and Uniprotsp trembl IDs are parts of UniProt database. The gene ID from Bioexpress corresponded to “Uniprotswissprot ID” column of newly received gene list. ID conversion is an inevitable step when merging data from different sources, in our cases tissue-specific genes and differentially expressed genes. Some preliminary works were done to prepare the files for manipulation and then conversion using with Algorithm 17.

Algorithm 17

```
library(tidyverse)
library(readr)
library(kableExtra)
library(knitr)
all_geneTissuePair <- read_csv(file = "all_geneTissuePair.csv")
only_parent_geneTissuePair <- read_csv(file = "only_parent_geneTissuePair.csv")
all_geneTissuePair <- tbl_df(all_geneTissuePair)
only_parent_geneTissuePair <- tbl_df(only_parent_geneTissuePair)
```

```
DEG_list <- read_csv(file = "DEG_list.csv")
names(DEG_list)[names(DEG_list) == "UniProtKB_AC"] <- "uniprotswissprot"
DEG_list <- tbl_df(DEG_list)
uniqu_deg_list <- DEG_list %>% select(Cancer.Ontology, uniprotswissprot) %>%
group_by(Cancer.Ontology,
   uniprotswissprot) %>% distinct()
deg_table <- uniqu_deg_list %>% group_by(Cancer.Ontology) %>%
summarise(DEG_number = n())
write.csv(deg_table, "deg_table.csv")

deg_table <- read.csv(file = "deg_table.csv", stringsAsFactors = FALSE)
kable(deg_table)
```

Before ID conversion the number of differentially expressed genes in each cancer type is as shown in Table 6.2 below:

Table 6.2 Number of DEG (with UniProt ID) for all cancer types

Cancer Ontology	Number of DEG
DOID:10283 / Prostate cancer [PCa]	1552
DOID:10534 / Stomach cancer [Stoca]	904
DOID:11054 / Urinary bladder cancer [UBC]	3936
DOID:1324 / Lung cancer [Lunca] & DOID:3907 / Lung squamous cell carcinoma	5334
DOID:1324 / Lung cancer [Lunca] & DOID:3910 / Lung adenocarcinoma	21
DOID:1612 / Breast cancer [BRCA]	1780
DOID:1781 / Thyroid cancer [Thyca]	814
DOID:1793 / Pancreatic cancer [PACA] & DOID:4074 / Pancreas adenocarcinoma	177
DOID:1993 / Rectum cancer [Recca] & DOID:1996 / Rectum adenocarcinoma	8
DOID:263 / Kidney cancer [Kidca] & DOID:4465 / Papillary renal cell carcinoma	680
DOID:263 / Kidney cancer [Kidca] & DOID:4471 / Chromophobe adenocarcinoma	10
DOID:263 / Kidney cancer [Kidca] & Kidney renal clear cell carcinoma	4645
DOID:3571 / Liver cancer [Livca]	610
DOID:363 / Uterine cancer [Uteca]	4431
DOID:4362 / Cervical cancer [Cerca] & DOID:3744 / Cervical squamous cell carcinoma & DOID:0050940 / Endocervical adenocarcinoma	3397
DOID:5041 / Esophageal cancer [EC]	3701

As mentioned above, conversion of gene ID is required for further analyses to investigate intra-tumoral heterogeneity. For this purpose, Algorithm 18 was used for ID conversion process from UniProt to Ensembl Gene ID:

Algorithm 18

```
conversion <- read_csv(file = "conversion.csv")
conversion <- rename(conversion, gene = ensembl_gene_id)
conversion <- tbl_df(conversion)
conversion <- select(conversion, 2, 4, 9)
DEG_conversion_finish <- left_join(uniqu_deg_list, conversion) %>%
filter(!is.na(gene)) %>% group_by(Cancer.Ontology) %>% unique() %>% ungroup()
write.csv(DEG_conversion_finish, file = "DEG_conversion_finish.csv")
DEG_conversion_finish <- read_csv(file = "DEG_conversion_finish.csv")
deg_table_converted <- DEG_conversion_finish %>% group_by(Cancer.Ontology) %>%
summarise(DEG_number_converted = n())
write.csv(deg_table_converted, "deg_table_converted.csv")
deg_table_converted<-read.csv(file = "deg_table_converted.csv",stringsAsFactors =
FALSE)
kable(deg_table_converted)
```

After ID conversion, converted DEG list was stored as csv file and number of DEG in each cancer type is shown in Table 6.3:

Table 6.3 Number of DEG for each cancer after ID conversion process

Cancer Ontology	Number of DEG
DOID:10283 / Prostate cancer [PCa]	1617
DOID:10534 / Stomach cancer [Stoca]	947
DOID:11054 / Urinary bladder cancer [UBC]	4102
DOID:1324 / Lung cancer [Lunca] & DOID:3907 / Lung squamous cell carcinoma	5656
DOID:1324 / Lung cancer [Lunca] & DOID:3910 / Lung adenocarcinoma	20
DOID:1612 / Breast cancer [BRCA]	1836
DOID:1781 / Thyroid cancer [Thyca]	814
DOID:1793 / Pancreatic cancer [PACA] & DOID:4074 / Pancreas adenocarcinoma	177

Table 6.3 Number of DEG for each cancer after ID conversion process (cont'd)

DOID:1993 / Rectum cancer [Recca] & DOID:1996 / Rectum adenocarcinoma	8
DOID:263 / Kidney cancer [Kidca] & DOID:4465 / Papillary renal cell carcinoma	680
DOID:263 / Kidney cancer [Kidca] & DOID:4471 / Chromophobe adenocarcinoma	10
DOID:263 / Kidney cancer [Kidca] & Kidney renal clear cell carcinoma	4645
DOID:3571 / Liver cancer [Livca]	610
DOID:363 / Uterine cancer [Uteca]	4431
DOID:4362 / Cervical cancer [Cerca] & DOID:3744 / Cervical squamous cell carcinoma & DOID:0050940 / Endocervical adenocarcinoma	3397
DOID:5041 / Esophageal cancer [EC]	3701

Although ID conversion might seem trivial it is prone to unexpected hurdles. Due to disagreements between different databases, one gene ID from one database can map to multiple gene IDs in another database. In our case, one Uniprot ID in DEG list can map to several different Ensembl Gene IDs. In this step, DEG list was joined with all_geneTissuePair list, which was obtained our results of analyses and including Ensembl Gene ID of genes, to increase reliability of ID conversion using Algorithm 19 and the final DEG list was saved as a csv file as follows:

Algorithm 19

```
DEG_conversion_tissue_specific_all <-DEG_conversion_finish %>% select(3, 4) %>%
  distinct() %>% inner_join(all_geneTissuePair) %>% select(uniprotswissprot, gene,
  tissue, expressionLevel) %>% distinct() %>% group_by(uniprotswissprot) %>%
top_n(1,
  expressionLevel) %>% inner_join(DEG_conversion_finish, by =
"uniprotswissprot")%>%  arrange(Cancer.Ontology) %>% select(1, 2, 6, 8) %>%
distinct()
table_ final <- DEG_conversion_tissue_specific_all %>% group_by(Cancer.Ontology)
%>% summarise(new_number  = (n()))
write.csv(table_ final, "table_final.csv")

table_final <- read.csv(file = "table_final.csv", stringsAsFactors = FALSE)
```

New clear and reliable DEG list was generated and number of genes is shown below as a new table finally:

Table 6.4 Number of DEG after ID conversion using a reliable Ensembl Gene ID list (generating in the thesis study including specific genes)

Cancer Ontology	Number of DEG
DOID:10283 / Prostate cancer [PCa]	918
DOID:10534 / Stomach cancer [Stoca]	537
DOID:11054 / Urinary bladder cancer [UBC]	1640
DOID:1324 / Lung cancer [Lunca] & DOID:3907 / Lung squamous cell carcinoma	2568
DOID:1324 / Lung cancer [Lunca] & DOID:3910 / Lung adenocarcinoma	17
DOID:1612 / Breast cancer [BRCA]	989
DOID:1781 / Thyroid cancer [Thyca]	502
DOID:1793 / Pancreatic cancer [PACA] & DOID:4074 / Pancreas adenocarcinoma	108
DOID:1993 / Rectum cancer [Recca] & DOID:1996 / Rectum adenocarcinoma	6
DOID:263 / Kidney cancer [Kidca] & DOID:4465 / Papillary renal cell carcinoma	425
DOID:263 / Kidney cancer [Kidca] & DOID:4471 / Chromophobe adenocarcinoma	6
DOID:263 / Kidney cancer [Kidca] & Kidney renal clear cell carcinoma	2421
DOID:3571 / Liver cancer [Livca]	385
DOID:363 / Uterine cancer [Uteca]	1607
DOID:4362 / Cervical cancer [Cerca] & DOID:3744 / Cervical squamous cell carcinoma & DOID:0050940 / Endocervical adenocarcinoma	1298
DOID:5041 / Esophageal cancer [EC]	1699

6.3 Intersection of Tissue-Specific Genes and Differentially Expressed Genes in Cancer

Having gathered all specific-genes and DEG for 16 different cancer types, the next step was to intersect and find overlapping genes for each cancer type, to investigate and understand tumoral heterogeneity and cancer cell behavior through a new reliable robust and rigorous computational approach in the context of this thesis study. Number of calculated tissue-specific genes for parent tissue and results from this process will be given in Results and Discussion. Specific genes for all tissues were listed in the following code, Algorithm 20:

Algorithm 20

```
library(tidyverse)
all_geneTissuePair <- read.csv(file = "all_geneTissuePair.csv", stringsAsFactors = FALSE)
only_parent_geneTissuePair <- read.csv(file = "only_parent_geneTissuePair.csv",
stringsAsFactors = FALSE)
all_geneTissuePair <- tbl_df(all_geneTissuePair)
only_parent_geneTissuePair <- tbl_df(only_parent_geneTissuePair)
parent <- only_parent_geneTissuePair %>% select(parent_tissue) %>% unique()
parent_spe <- only_parent_geneTissuePair %>% select(gene, parent_tissue) %>%
group_by(parent_tissue,  gene) %>% unique() %>% summarise(spe = n())

write.csv(parent_spe,"specific_genes_final.csv")
stomach <- filter(parent_spe, parent_tissue == parent[1, ])
brain <- filter(parent_spe, parent_tissue == parent[2, ])
spleen <- filter(parent_spe, parent_tissue == parent[3, ])
skin <- filter(parent_spe, parent_tissue == parent[4, ])
bone.marrow  <- filter(parent_spe, parent_tissue == parent[5, ])
testis <- filter(parent_spe, parent_tissue == parent[6, ])
appendix <- filter(parent_spe, parent_tissue == parent[7, ])
liver <- filter(parent_spe, parent_tissue == parent[8, ])
lung <- filter(parent_spe, parent_tissue == parent[9, ])
prostate <- filter(parent_spe, parent_tissue == parent[10, ])
thyroid  <- filter(parent_spe, parent_tissue == parent[11, ])
salivary.gland <- filter(parent_spe, parent_tissue == parent[12, ])
small.intestine <- filter(parent_spe, parent_tissue == parent[13, ])
placenta <- filter(parent_spe, parent_tissue == parent[14, ])
lymph.node  <- filter(parent_spe, parent_tissue == parent[15, ])
heart <- filter(parent_spe, parent_tissue == parent[16, ])
adrenal.gland <- filter(parent_spe, parent_tissue == parent[17, ])
esophagus <- filter(parent_spe, parent_tissue == parent[18, ])
kidney <- filter(parent_spe, parent_tissue == parent[19, ])
```

```
ovary <- filter(parent_spe, parent_tissue == parent[20, ])
pancreas <- filter(parent_spe, parent_tissue == parent[21, ])
colon <- filter(parent_spe, parent_tissue == parent[22, ])
adipose.tissue <- filter(parent_spe, parent_tissue == parent[23, ])
rectum <- filter(parent_spe, parent_tissue == parent[24, ])
oviduct <- filter(parent_spe, parent_tissue == parent[25, ])
seminal.vesicle <- filter(parent_spe, parent_tissue == parent[26, ])
vagina <- filter(parent_spe, parent_tissue == parent[27, ])
tongue <- filter(parent_spe, parent_tissue == parent[28, ])
uterus <- filter(parent_spe, parent_tissue == parent[29, ])
spinal.cord <- filter(parent_spe, parent_tissue == parent[30, ])
epididymis <- filter(parent_spe, parent_tissue == parent[31, ])
penis <- filter(parent_spe, parent_tissue == parent[32, ])
breast <- filter(parent_spe, parent_tissue == parent[33, ])
vas.deferens <- filter(parent_spe, parent_tissue == parent[34, ])
whole.blood <- filter(parent_spe, parent_tissue == parent[35, ])
tibial.nerve <- filter(parent_spe, parent_tissue == parent[36, ])
```

The purpose of using DEG is illustrated in Figure 6.1 below:

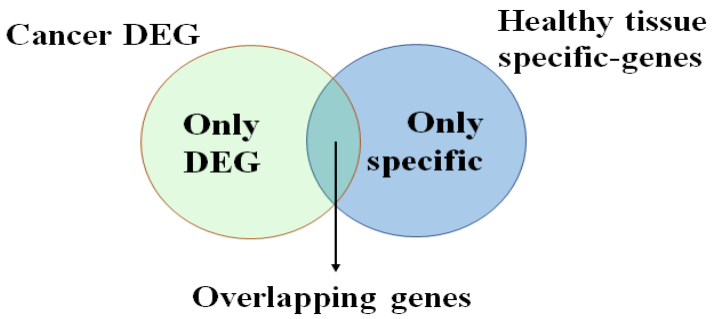


Figure 6.1 Intersection process between DEG and specific genes

We calculated overlapping genes between cancer DEG for each type of cancer and tissue-specific genes for each healthy tissue matching the cancer type. We generate results that contain number of overlapping genes and their lists, number of only DEG for each cancer and number of only tissue-specific genes for corresponding tissue. Furthermore, to analyze intra-tumoral heterogeneity we tried to show whether DEG for a single cancer type are specific to other tissues instead of related tissue or not. The lists

of genes were computed with the Algorithm 21. This process was applied to all cancer DEG data mentioned in this Section. The sample code below is for only a few cancers (breast cancer, liver cancer, cervix cancer):

Algorithm 21

```
# 1. Breast cancer (breast tissue)
breastcancer<-
filter(DEG_conversion_tissue_specific_all,Cancer.Ontology=="DOID:1612 / Breast
cancer [BRCA]")
intersection_breast<-data.frame(intersect(breast$gene,breastcancer$gene.x))
only_breast_deg<- data.frame(setdiff(breastcancer$gene.x,breast$gene))
only_breast_spe<- data.frame(setdiff(breast$gene,breastcancer$gene.x))
od_breast<-
rename(only_breast_deg,gene=setdiff.breastcancer.gene.x..breast.gene.)%>%
left_join(parent_spe, by=c("gene"))%>%
 filter(!is.na(spe))%>%
 filter(parent_tissue!="breast")%>%
 group_by(parent_tissue) %>%
 summarise(number=n())
# 2. Liver cancer (liver tissue)
livercancer<- filter(DEG_conversion_tissue_specific_all,Cancer.Ontology=="DOID:3571
/ Liver cancer [Livca]")
intersection_liver<-data.frame(intersect(liver$gene,livercancer$gene.x))
only_liver_deg<- data.frame(setdiff(livercancer$gene.x,liver$gene))
only_liver_spe<- data.frame(setdiff(liver$gene,livercancer$gene.x))
od_liver<-rename(only_liver_deg,gene=setdiff.livercancer.gene.x..liver.gene.)%>%
left_join(parent_spe, by=c("gene"))%>%
 filter(!is.na(spe))%>%
 filter(parent_tissue!="liver")%>%
 group_by(parent_tissue)%>%
 summarise(number=n())
# 3. Cervix cancer (cervix,ectocervix,endocervix tissues)
```

```
cervix<-all_geneTissuePair%>%
 select(gene,tissue)%>%
 group_by(tissue,gene)%>%
 unique()%>%
 summarise(spe=n())%>%
 filter(tissue=="cervix")
cervixcancer<-
filter(DEG_conversion_tissue_specific_all,Cancer.Ontology=="DOID:4362 / Cervical cancer [Cerca] & DOID:3744 / Cervical squamous cell carcinoma & DOID:0050940 / Endocervical adenocarcinoma")
intersection_cervix<-data.frame(intersect(cervix$gene,cervixcancer$gene.x))
only_cervix_deg<- data.frame(setdiff(cervixcancer$gene.x,cervix$gene))
only_cervix_spe<- data.frame(setdiff(cervix$gene,cervixcancer$gene.x))
od_cervix<rename(only_cervix_deg,gene=setdiff.cervixcancer.gene.x..cervix.gene.)%>
%left_join(parent_spe, by=c("gene"))%>%
 filter(!is.na(spe))%>%
 filter(parent_tissue!="cervix")%>%
 group_by(parent_tissue)%>%
 summarise(number=n())
```

Differentially expressed genes were calculated for 16 different cancer types. After that, DEG of cancer type and genes specifically expressed in the matching tissues were overlapped. In addition to all, intersected genes and only DEG were investigated in the tissue specificity perspective.

6.4 Functional Annotation of All Gene Groups

Functional annotation is very important for understanding the role of genes in biological processes. For this reason, The Database for Annotation, Visualization and Integrated Discovery (DAVID) [278] was used for functional annotation in this study. DAVID can be used to understand biological meaning of genes.

Intersected genes might be crucial for targeted treatment of solid tumors. They can be used as biomarker for related cancer because they are specifically expressed in that

tissue or organ. In this context, exploring these genes in context of molecular pathways or biological processes is essential. Hence, both all tissue-specific genes and only intersected genes were analyzed by using DAVID annotation tool to detect related diseases, pathways and as an important parameter, GO terms (depends on biological processes, molecular functions, cellular components) for each gene group compiled in this study. While genes are annotated with their functional properties in organism, **Benjamini-Hochberg adjusted *P* value** is the criterion of being significant. It is a powerful procedure that decreases the false discovery rate. Benjamini-Hochberg is an adjustment method including the Bonferroni correction in which the p-values are multiplied by the number of comparisons and it is used when there are more than one p value [279]. If this score is $\leq 10^{-2}$, listed functions, diseases, pathways are significant and associated with examined genes. Usage of DAVID and annotation results as an example was indicated in Figure 6.2:

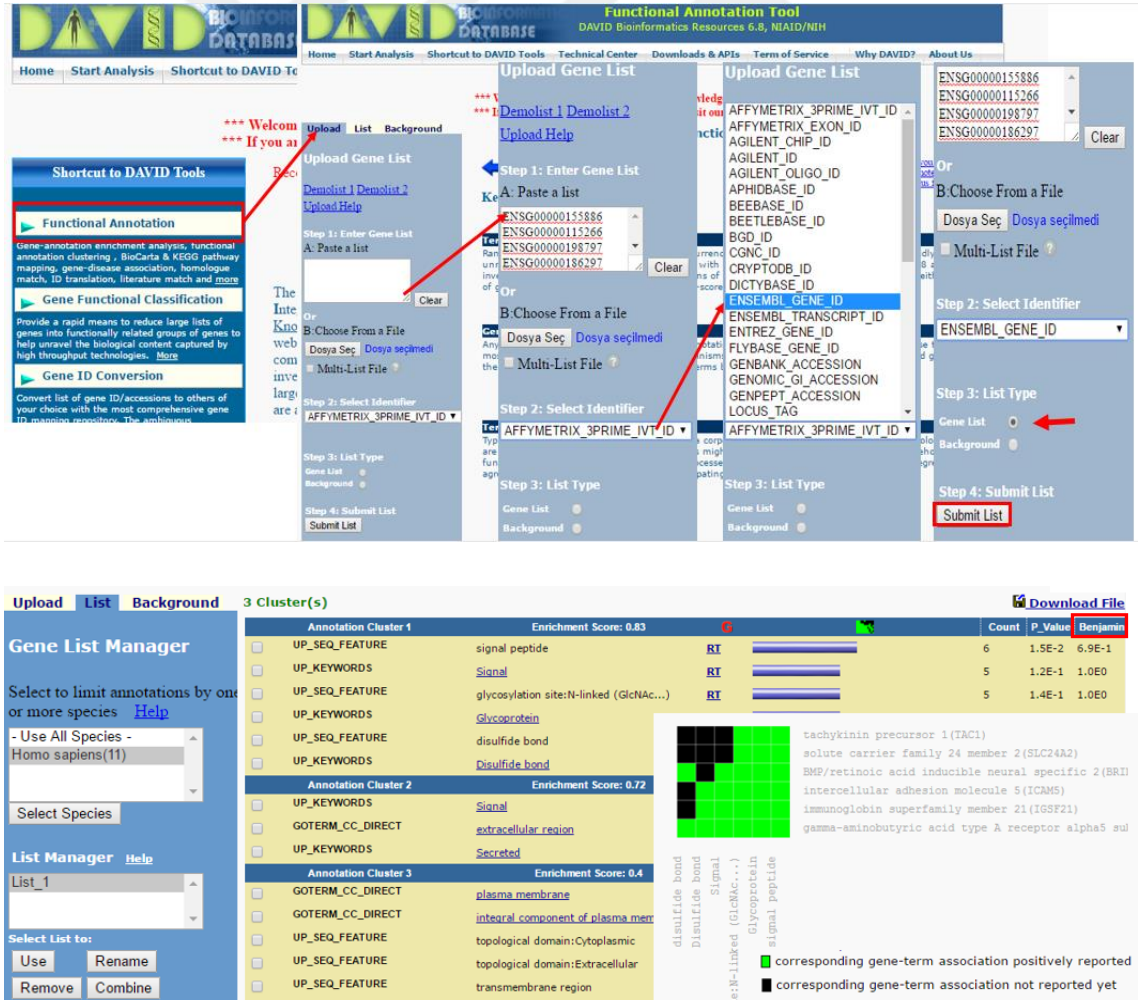


Figure 6.2 Usage of DAVID for functional annotation

In this thesis study, first aim is to identify protein-coding gene profiles and determine specifically expressed genes in 96 child tissues that are parts of 39 parent tissues. Next aim is to reveal intra-tumoral heterogeneity and tumor cell behavior of different cancer types. Therefore, two important gene groups which are specifically expressed in healthy tissues and differentially expressed genes in cancers were intersected, in brief.

The applications performed in this section can be summarized by the following flow chart, in Figure 6.3:

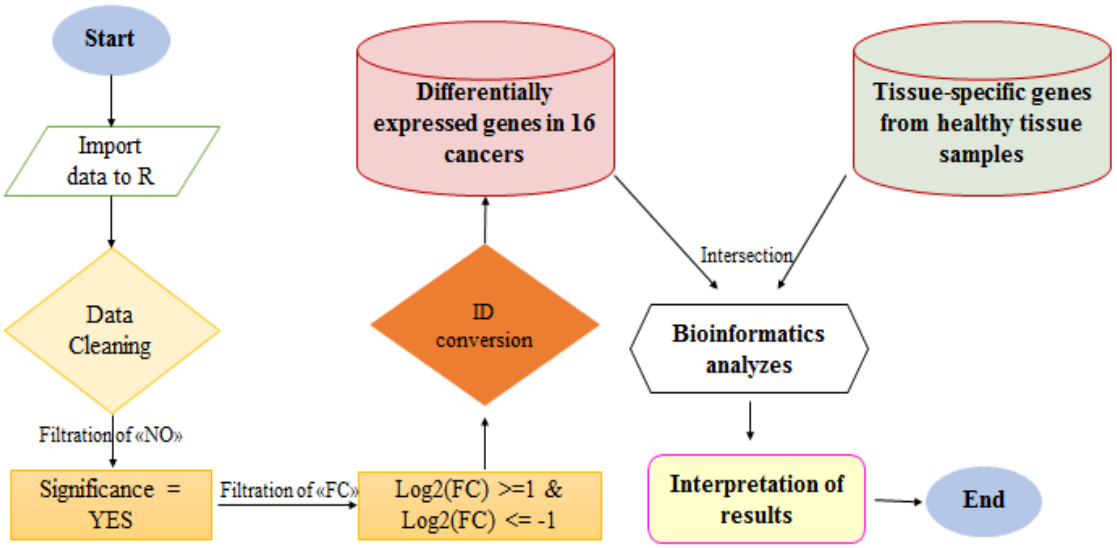


Figure 6.3 Flow chart of obtaining DEG and other processes

In addition to DEG analysis it was decided that only cancer patient data should be examined in terms of tissue specificity.

6.5 TCGA Data for Various Cancer Types

Gene expression may show alteration in cancer patients depending on cancer type, stage and environmental conditions. Determination of gene expression of tumor tissue and normal tissue in cancer patients are very significant perspective to understand mechanism of cancer cell and this process. In this part of study, gene expression profile was examined to investigate cancer cell mechanism, tumoral heterogeneity and to find out new specific targets for cancer in the context of tissue-specific genes for various cancer types and patients in different conditions.

TCGA (The Cancer Genome Atlas) [280] is a cancer data portal that collects and analyzes tumor samples from many patients who have many different cancer types. There are a lot of data types in TCGA such as clinical data, DNA-sequencing, miRNA-sequencing, RNA-sequencing, DNA methylation etc. In this study, FPKM normalized RNA-sequencing data was retrieved and used for downstream analysis. Cancer types which had high number of samples and were solid tumors, were selected among TCGA data. Data were downloaded by TCGA portal specific R packages. Selected cancer types and number of cancerous tissue and normal tissue samples in each cancer are presented in Table 6.5:

Table 6.5 Number of cancerous and healthy tissue samples obtained from TCGA

Cancer type	**Number of tumor tissue samples**	**Number of normal tissue samples**
Bladder Urothelial Carcinoma [BLCA]	412	19
Breast invasive carcinoma [BRCA]	1097	50
Colon adenocarcinoma [COAD]	461	41
Esophageal carcinoma [ESCA]	185	11
Kidney renal clear cell carcinoma [KIRC]	536	72
Kidney renal papillary cell carcinoma [KIRP]	291	32
Liver hepatocellular carcinoma [LIHC]	377	50
Lung adenocarcinoma [LUAD]	521	59
Pancreatic adenocarcinoma [PAAD]	185	4
Thyroid carcinoma [THCA]	507	59
Uterine Corpus Endometrial Carcinoma [UCEC]	548	24

Cancer data were downloaded from TCGA, organized as SummarizedExperiment object and archived as RDS file in R. TCGAbiolinks, GenomicDataCommons and

SummarizedExperiment packages were used for mentioned tasks. The code below, as an example, downloads Breast cancer (BRCA) data, organizes into SummarizedExperiment object and then saves as rds file using Algorithm 22:

Algorithm 22

```
# source('https://bioconductor.org/biocLite.R')
# biocLite('TCGAbiolinks')
# biocLite('SummarizedExperiment')
# biocLite("GenomicDataCommons")
library(TCGAbiolinks)
library(SummarizedExperiment)
library(GenomicDataCommons)
suppressPackageStartupMessages(library(tidyverse))
query.exp.hg19 <- GDCquery(project = "TCGA-BRCA",
                           data.category = "Gene expression",
                           data.type = "Gene expression quantification",
                           platform = "Illumina HiSeq",
                           file.type  = "normalized_results",
                           experimental.strategy = "RNA-Seq",
                           legacy = TRUE)
GDCdownload(query.exp.hg19, method = "client", files.per.chunk = 5)
downloaded_data <- GDCprepare(query.exp.hg19)
saveRDS(downloaded_data,"tcga_brca_rnaseq_data.rds")
```

11 different cancer data were used in this study, data were obtained as rds file as follows:

tcga_brca_rnaseq_data.rds	tcga_kirc_rnaseq_data.rds	tcga_paad_rnaseq_data.rds
tcga_blca_rnaseq_data.rds	tcga_kirp_rnaseq_data.rds	tcga_thca_rnaseq_data.rds
tcga_coad_rnaseq_data.rds	tcga_lihc_rnaseq_data.rds	tcga_ucec_rnaseq_data.rds
tcga_esca_rnaseq_data.rds	tcga_luad_rnaseq_data.rds	

The code below demonstrates that TCGA data were read from rds file, sample information (patient data such as age, gender, race, tumor stage, etc) has been extracted and joined with expression data in tidy format. Liver cancer (lihc) and pancreas cancer (paad) examples were only shown in Algorithm 23:

Algorithm 23

```
library(SummarizedExperiment)
library(tidyverse)
library(stringr)
library(biobroom)

#Liver cancer data (lihc_data)
lihc <- readRDS("tcga_lihc_rnaseq_data.rds")
exp_metadata_lihc <- as.tibble(colData(lihc))
names(rowRanges(lihc)) <- paste0(rowRanges(lihc)$ensembl_gene_id,"-
",rowRanges(lihc)$entrezgene)
lihc_data<-exp_metadata_lihc %>%
 select(1:3,4,5,10,11,22,23,28:29,31:33,41:44,46,55)  %>%
 mutate(age_at_diagnosis=as.integer(ceiling(age_at_diagnosis/ 365.2424))) %>%
 inner_join(tidy(lihc), by=c("barcode"="sample")) %>%
 select(gene,name,value,everything()) %>%
 select(gene,value,sample,shortLetterCode)  %>%
 spread(shortLetterCode,value)  %>%
 arrange(gene,sample)

#Pancreas cancer data (paad_data)
paad <- readRDS("tcga_paad_rnaseq_data.rds")
exp_metadata_paad <- as.tibble(colData(paad))
names(rowRanges(paad)) <- paste0(rowRanges(paad)$ensembl_gene_id,"-
",rowRanges(paad)$entrezgene)
paad_data<-exp_metadata_paad %>%
 select(1:3,4,5,10,11,22,23,28:29,31:33,41:44,46,55)  %>%
 mutate(age_at_diagnosis=as.integer(ceiling(age_at_diagnosis/ 365.2424))) %>%
```

```
inner_join(tidy(paad), by=c("barcode"="sample")) %>%
select(gene,name,value,everything()) %>%
select(gene,value,sample,shortLetterCode) %>%
spread(shortLetterCode,value) %>%
arrange(gene,sample)
```

TCGA data have some information about patients such as, tumor type (primary tumor, metastatic tumor etc.), stage (stage I, II, III etc.), alcohol consumption, smoking, sex, age and nationality. We used tumor type information. Tumor types had been coded in TCGA data like that:

Primary Solid Tumor → TP

Recurrent Solid Tumor → TR

Metastatic Tumor → MT

Solid Tissue Normal → NT

There are some steps in here to compare cancer data which were obtained from TCGA and normal tissue data which were used for tissue specificity exported from ArrayExpress and ExpressionAtlas. In Algorithm 24,

- Raw expression data from EMTAB datasets were merged as a single file for comparison with the TCGA data.
- Percentages of expression of each specifically expressed gene in tissues were calculated, added as a new column and saved as rds file that is called "raw_expression_percentage_all". Moreover, there is a file defined as "joined_genepair" which was prepared before for plotting graphics. It contains specific genes for all datasets, related tissues and tau scores of genes.

Algorithm 24

```
data1733<-read_csv("data1733.csv")
data2836<-read_csv("data2836.csv")
data5214<-read_csv("data5214.csv")
data4344<-read_csv("data4344.csv")
data3358<-read_csv("data3358.csv")
```

```
raw_expression_all<-data1733%>%
 rename(Ensembl.Gene.ID=EnsemblGeneID)%>%
 gather(tissue,expression,-X1,-Ensembl.Gene.ID)%>%
 mutate(sample="data1733")
raw_expression_all<-data2836%>%
 select(-Gene.Name)%>%
 gather(tissue,expression,-X1,-Ensembl.Gene.ID)%>%
 mutate(sample="data2836")%>%
 bind_rows(raw_expression_all)
raw_expression_all<-data5214%>%
 select(-Gene.Name)%>%
 gather(tissue,expression,-X1,-Ensembl.Gene.ID)%>%
 mutate(sample="data5214") %>%
 bind_rows(raw_expression_all)
raw_expression_all<-data4344%>%
 select(-Gene.Name)%>%
 gather(tissue,expression,-X1,-Ensembl.Gene.ID)%>%
 mutate(sample="data4344") %>%
 bind_rows(raw_expression_all)
raw_expression_all<-data3358%>%
 select(-Gene.Name)%>%
 gather(tissue,expression,-X1,-Ensembl.Gene.ID)%>%
 mutate(sample="data3358") %>%
 bind_rows(raw_expression_all)
raw_expression_all%>%
 select(sample)%>%
 count(sample)
saveRDS(raw_expression_all,"raw_expression_all.rds")
raw_expression_percentage_all<-raw_expression_all %>%
 group_by(Ensembl.Gene.ID,sample)%>%
 mutate(percentage=expression / sum(expression)*100)
saveRDS(raw_expression_percentage_all,"raw_expression_percentage_all.rds")
```

In this section, we try to understand and interpret intra-tumoral heterogeneity, cancer cell migration and behavior using TCGA data instead of DEG. For this purpose, we used some filtrations when we processed together three different data as mentioned above: raw expression percentage data, tissue-specific genes data and TCGA expression data. Some criteria must be defined for the combining of TCGA cancer data with specific gene data.

- First criterion was used for TCGA cancer data. The averages of normal tissue samples were calculated for each gene from cancer data. Tumor samples were evaluated separately because of different expression levels among patients. After that this criterion is as follows: **TP > (average_NT + (2*stddev_NT))**

  TP: expression levels of primary tumor samples

  Average_NT: average of expression levels of normal tissue samples

  Stddev_NT: standard deviation of expression levels of normal tissue samples

  If expression of a gene in primary tumor sample is above the criterion, this gene will be filtered in related cancer type to be used for further analysis.

- Second criterion is about primary tumor samples. TP value of a gene for patients in single cancer must be greater than 50 as expression value to minimize experimental errors and to achieve high accuracy results **(TP>50)**.

- Third criterion is about raw expression data of tissue-specific genes. If a gene is specific to a tissue, it may be specific to other tissues thanks to our new approach. Furthermore, it may be expressed in other tissues in very small quantities. Percentage of raw expression of specific-genes calculated from EMTAB datasets must be **higher than 95%** because we desire to select restrictedly expressed specific-genes in a particular single tissue. By this way, strictly expressed specific-genes were preferred for further analysis. Meaning of this criterion is that a gene is not only specifically but also restrictedly expressed at a very high proportion in a particular tissue.

- Fourthly, the more important criterion concerns both data: raw expression of a tissue specific-gene and normal tissue expression retrieved from TCGA for the

corresponding tissue. This criterion can be better explained with an example. Assume that a gene is specific to pancreas tissue strictly **(raw expression > 95% of total)**. This gene was shown to be expressed based on TCGA data in primary tumor sample of liver cancer. Normally, healthy liver sample must have low expression level for this gene because it is specific and restrictedly expressed in pancreas. Therefore, there is a significant criterion relationship between expression of gene in both raw expression of gene and normal sample value of patients. Criterion was adjusted as follows: **average_NT< (0.1*raw expression)**

Note: Raw expression values of genes come from EMTAB data.

- The last criterion is about number of patients. Genes were filtered according at least 20 patients with all the criteria mentioned above were accepted **(number of patients > 20)**.

All criteria and combination of three different gene information tables are explained in brief as a simple figure.

Specific gene list		Raw expression values from EMTAB			TCGA (example: ucec data)				
gene	tissue	gene	raw expression	percentage	gene	patients	TP	NT	criteria
geneA	esophagus	geneA	493.21	97.83%	geneA	X	9207.67	2.12	8.13
.	.	.	.	.	.	.	.	.	.
geneD	heart	geneD	1126	99%	geneD	T	202.63	23.19	79.01
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.

Percentage > 95% : **restrictedly expression of gene**
TP > 50 : **expression of primary tumor**
criteria = average_NT + (2*stddev_NT) : **an important criteria**
TP > criteria : **TP expression is higher than criteria**
average_NT< (0.1*raw expression) : **very low expression in normal tissue**
Number of patients > 20 : **at least 20 patients**

Figure 6.4 Data were joined, and all criteria were applied to reveal significant genes

The above-mentioned procedures an example with uterine carcinoma (ucec) in Figure 6.4 were performed for each 11 different cancer types from TCGA, and the procedure above was repeated for all cancer data using following Algorithm 25 (only liver cancer and pancreas cancer data code samples were shown):

Algorithm 25

```
#Liver cancer data (lihc_data)

only_calculated<-lihc_data %>%
 group_by(gene) %>%
 mutate_all(funs(ifelse(is.na(.),0,.))) %>%
 mutate(count=n_distinct(NT)-1)%>%
 arrange(gene) %>%
 mutate(total_NT=Reduce("+",NT)) %>%
 mutate(average_NT=total_NT / count) %>
 mutate(NT_NA=ifelse(NT==0,NA,NT)) %>%
 mutate(stddev_NT=sd(NT_NA,na.rm=TRUE)) %>%
 mutate(criteria=average_NT+(2*stddev_NT))

 only_calculated_criteria<-only_calculated %>%
 filter(TP>criteria) %>%
 separate(gene, c("ensembl", "entrez"), "-")
join_lihc_specific_rawexp<-joined_genepair %>%
 mutate(sample=str_replace(sample,"E-MTAB-","data")) %>%


left_join(raw_expression_percentage_all,by=c("Ensembl.Gene.ID","tissue","sample"))
%>%
 left_join(only_calculated_criteria,by=c("Ensembl.Gene.ID"="ensembl")) %>%
 arrange(-percentage) %>%
 filter(tissue!="liver",percentage>95,TP>50,average_NT< (0.1*expression)) %>%
 count(Ensembl.Gene.ID,tissue)

join_lihc_specific_rawexp20<-join_lihc_specific_rawexp %>%
 group_by(Ensembl.Gene.ID,tissue) %>%
 select(2,3,13) %>%
 unique() %>%
 count() %>%
```

```
 filter(n>20)
saveRDS(join_lihc_specific_rawexp20,"join_lihc_specific_rawexp20.rds")

#Pancreas cancer data (paad_data)

 paad_only_calculated<-paad_data %>%
 group_by(gene) %>%
 mutate_all(funs(ifelse(is.na(.),0,.))) %>%
 mutate(count=n_distinct(NT)-1)%>%
 arrange(gene) %>%
 mutate(total_NT=Reduce("+",NT)) %>%
 mutate(average_NT=total_NT / count) %>%
 mutate(NT_NA=ifelse(NT==0,NA,NT)) %>%
 mutate(stddev_NT=sd(NT_NA,na.rm=TRUE)) %>%
 mutate(criteria=average_NT+(2*stddev_NT))


 paad_only_calculated_criteria<-paad_only_calculated %>%
 filter(TP>criteria) %>%
 separate(gene, c("ensembl", "entrez"), "-")
join_paad_specific_rawexp<-joined_genepair %>%
 mutate(sample=str_replace(sample,"E-MTAB-","data")) %>%
left_join(raw_expression_percentage_all,by=c("Ensembl.Gene.ID","tissue","sample"))
%>%
 left_join(paad_only_calculated_criteria,by=c("Ensembl.Gene.ID"="ensembl")) %>%
 arrange(-percentage) %>%
 filter(tissue!="pancreas",percentage>95,TP>50,average_NT< (0.1*expression))

join_paad_specific_rawexp20<-join_paad_specific_rawexp %>%
 group_by(Ensembl.Gene.ID,tissue) %>%
 select(2,3,14) %>%
 unique() %>%
 count() %>%
 filter(n>20)

saveRDS(join_paad_specific_rawexp20,"join_paad_specific_rawexp20.rds")
```

After generating new lists, some analyses were performed, and detailed graphs were plotted. The genes which are highly expressed in a single cancerous tissue and passing all criteria were detected although they were specific to another particular tissue. These genes were described as "selected genes" in the thesis study. Moreover, selected genes were grouped according to the number of related cancer types and they were expressed according to stringent criterion described above. Analyses were performed for these genes in each cancer type. They were selected according to new thresholds:

- A single gene is specific according to all EMTAB datasets (consensus),
- Genes passing the above criteria and common all cancer types,
- Genes passing the above criteria appearing in 5 cancer types
- Genes passing the above criteria appearing in only one cancer type

All steps are summarized in Figure 6.5:

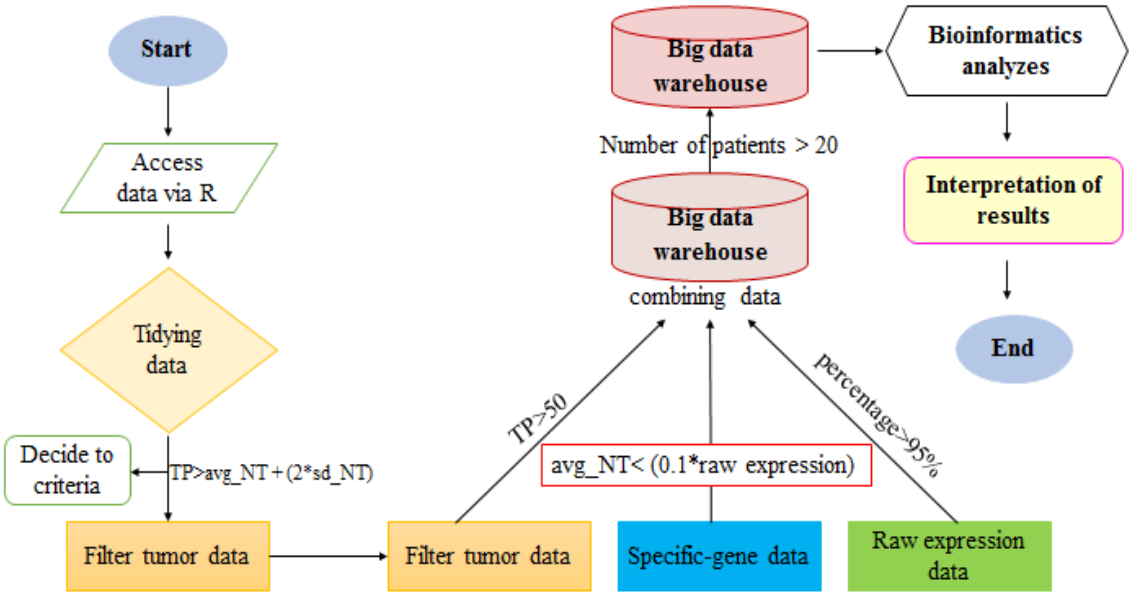


Figure 6.5 Flow chart of TCGA cancer data analysis

6.6 Understanding the Molecular Functions of Selected Genes

Understanding the functions of the identified genes is very important for the purpose of this study. The gene names and basic functions of selected genes which have Ensembl Gene ID are listed using BioMart.

GeneCards [281] is a searchable, integrative and user-friendly database that provides comprehensive and actual information on all annotated and predicted human genes. GeneCards was used to identify gene function, expression, paralogs, orthologs, disorders and pathways and more.

WebGestalt [282] (WEB-based Gene SeT AnaLysis Toolket) is a functional enrichment analysis web tool for gene lists which are interested in research in bioinformatics. One of the complementary methods for enrichment analysis is Over-Representation Analysis (ORA) in WebGestalt. We used ORA to reveal enrich functions of genes.

6.7 Network Analysis of Selected Genes

Genes are regulated by various transcription factors or microRNAs. The analysis above has supported estimation and interpolation of heterogeneity and interesting mechanisms of cancer cells. Besides, network analysis of these selected genes has potential to provide more insight.

NetworkAnalyst (network-based visual analytics for gene expression profiling, meta-analysis and interpretation) [283] is an interactive, simple and easy to use tool to support integrative analysis of gene expression data through statistical, visual and network-based approaches. It may be very useful to find related miRNAs, TFs, drugs etc. It helps identifying and interpreting genes via generating networks. We carried out network analysis for all cancer types and selected genes to elicit the related miRNAs.

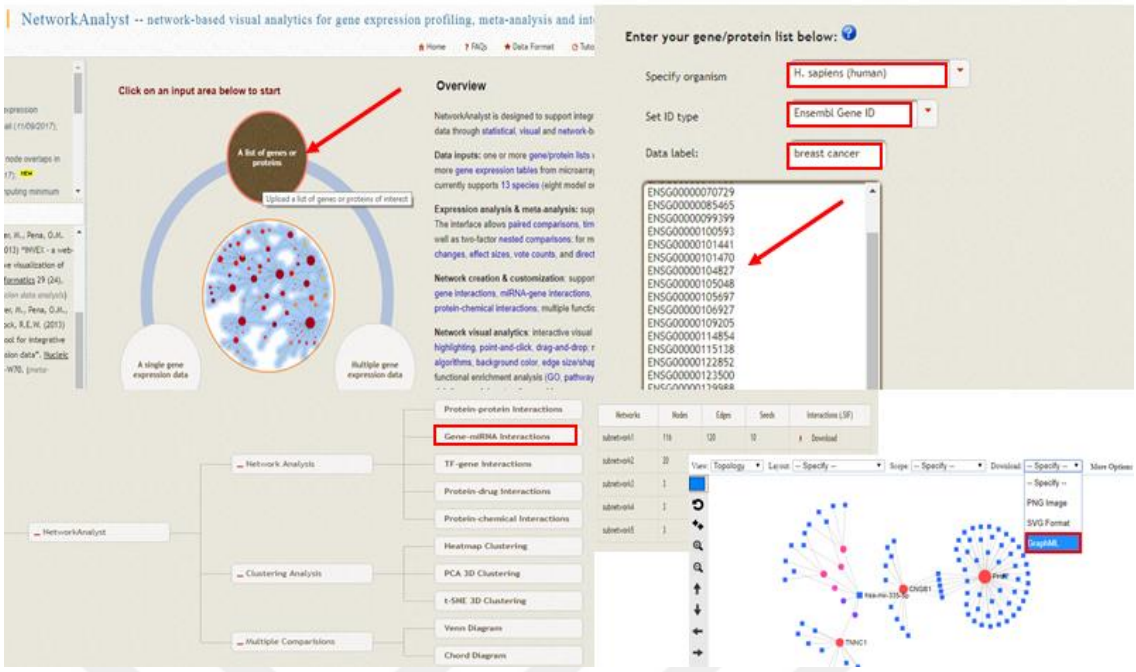


Figure 6.6 Obtaining related miRNAs and their networks via NetworkAnalyst

After network analysis of genes, detected miRNAs were examined in terms of related diseases. For this purpose, **miR2Disease Base** [284] was used. miR2Disease is a manually curated database, aims at providing a comprehensive knowledge of miRNA deregulation in various human diseases. Each result of miR2Disease contains detailed information on a miRNA-disease relationship, detection method for miRNA expression, and literature reference.

NetworkAnalyst allows exporting results into different file formats such as siff, png, svg and graphml. We exported our analysis as graphml file and visualized the networks via CytoScape.

Cytoscape [285] is an open source software platform for visualizing molecular interaction networks and biological pathways. It can be downloaded and used in personal computers easily. Even though CytoScape is capable of not only network visualization but also network analysis, we used this software for visualizing simple networks using selected genes and related miRNAs. All analyses stated in Section 6.6 and 6.7 were performed for DEG data in Section 6.3 and TCGA data in Section 6.5.

In this chapter, DEG cancer patients' gene data for various cancer types were examined. Furthermore, the association of them with tissue-specific genes were found out.

Significant genes which may be related to tumoral heterogeneity and cancer cell mechanisms were detected using some stringent criteria and thresholds. After that, bioinformatic analysis were carried out for understanding functions of genes and related miRNAs.

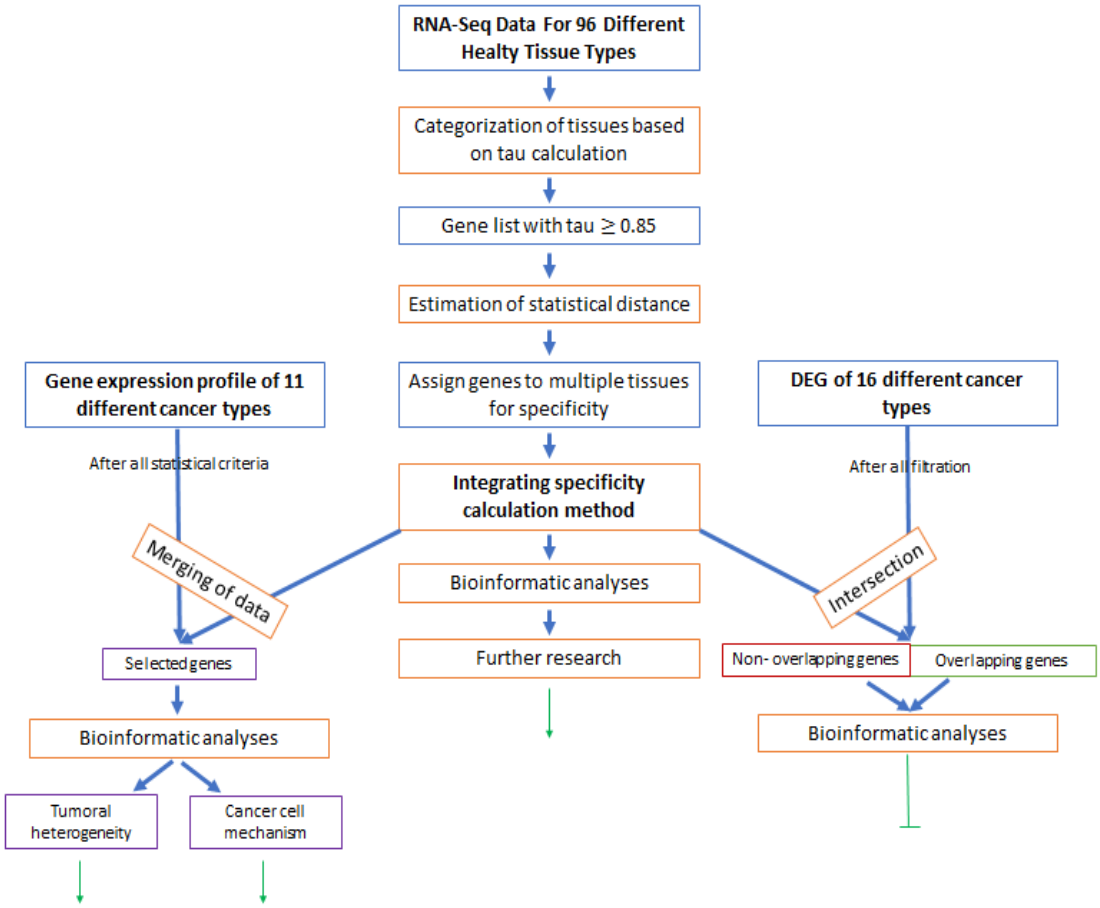


Figure 6.7 All analyses were performed during the thesis study

In this study, tissue-specific genes were identified using a robust and rigorous method. After that DEG lists for 16 different cancer types were examined in the context of tissue specificity. Some disadvantage of DEG are limited obtaining the detailed information and reliability of results are lower due to down regulated genes in DEG. For this reason, gene expression profiles of cancer patients were examined to obtain comprehensive and reliable information about tumoral heterogeneity and cancer cell invasive behaviors. All analyses were summarized as a flow chart in Figure 6.7 in order to discuss tumoral heterogeneity and complex mechanisms of cancer cell

Discovering new biomarkers for cancer detection, effective treatment and to solve heterogeneity mechanism will have huge impact for patients with cancer all over the world. This thesis study serves similar purpose. RNA-seq data was used because of its advantages. Besides, all computational analyses were performed using R programming language. The code of some comprehensive graphs is presented in Appendix-A Section.

CHAPTER 7

RESULTS AND DISCUSSION

Biological applications of computer science algorithms have led to the emergence of field of bioinformatics which facilitates significant developments in various medical problems. Upon improvements in NGS technology, bioinformatic analyses have become more widespread and crucial. NGS technologies comprehensively and systematically confirm nucleotide sequence, copy number of genomic loci, copy number of mRNA molecules, also known as “level of gene expression”. Innovative bioinformatics and computational techniques have been well applied in various medical research fields, and oncology is one such field.

Objective of this thesis study can be summarized as,

- Understanding of gene expression profiles,
- Generating robust and rigorous classification of tissue-specific genes
- Determining differentially expressed genes for various tumors, intersection of two important gene groups,
- Functional annotation of selected gene groups,
- Examination of gene expressions in cancer patients with primary tumors,
- Assessment of expression of genes in terms of tissue-specific genes,
- Discussing intra-tumoral heterogeneity,
- Interpreting cancer cell mechanisms and their behaviors
- Suggesting tissue-specific targets for each cancer and generating effective diagnostic biomarkers.

To achieve all the objectives, a comprehensive literature search and detailed analyzes were conducted. Five big datasets that are freely accessible were used in this study for

healthy gene expression. 96 different tissue types were classified via tissue ontology and a variety of cancer types were examined in the context of tumoral heterogeneity and cancer mechanisms. Our findings were presented in tables, figures and networks. This study was designed and conducted so that its output would contribute to the already growing body of information about cancer research.

7.1 Evaluation of Tissue Specificity Results

Gene expression is a crucial phenomenon for all living organisms to understand all steps of life. Therefore, in order to discover gene expression profiles, new devices, technologies and bioinformatic tools have been developed all over the world. If a gene is expressed in many cells, tissues or organs, it is classified as housekeeping gene and since it has a function needed in many organs, tissues and cell types such as general metabolic activity like glycolysis, cell cycle, cell development and cell death. On the other hand, if a gene is expressed in a single or several cells/tissues specifically, it is classified as tissue-specific gene. Tissue-specific genes have more specific roles and functions in their tissues. Recent studies indicate that to find out tissue-specificity is functionally informative for diagnosis, and treatment learning developmental stages [38].

There are tissue-specific genes which are associated with cancer morphology, pathology and moreover, intra-tumoral heterogeneity. Thus, identification of these genes is crucial for the discovery of tissue-specific drug targets and identifying tumoral heterogeneity. Intra-tumoral heterogeneity is an important obstacle which complicates the development of molecular targeted agents and can cause poor diagnosis. In this study, we aimed to determine tissue-specific genes, find intersection between differentially expressed genes in cancer and tissue-specific genes, for the purpose of understanding tumor heterogeneity using a new dynamic, powerful computational method. After the analyses of association between DEG and tissue-specific genes, gene expression profiles of cancer patients who have primary tumors were investigated in terms of tissue-specific patterns to understand tumoral heterogeneity and cancer specific mechanisms. Specifically expressed genes in one or several tissues can ensure important clues about gene functions and molecular mechanisms of disorders. Tau is a specificity index of genes that has been used in various studies [38], [118], [120], [121], [122], [123], [124].

Besides, tissue specificity can be computed in a comparable manner in different datasets or different species, as shown by Kryuchkova-Mostacci and Robinson-Rechavi's another study [286].

There are many studies from the beginning of 2000 about tissue specificity because it is a valuable information that has relationship with various disorders, discovering developmental stages of multicellular organisms. Su et al. in 2004 [287] and Liang et al. in 2006 [288], independently, generated tissue-specific mRNA expression profiles using thousands of genes across a large number of tissue types (130 tissue types combined) from normal human samples through microarray data. VeryGene as a tool was enhanced by Yang et al. [102] which link tissue-specific genes to diseases and drugs using microarray data. VeryGene and other tools or databases as mentioned before such as TisGeD, TiGER, PAGenBase used microarray or ESTs data via various specificity scores mentioned in Section 3.6.1. We used RNA-seq data and a new effective method integrated with the best specificity score, tau. Therefore, our study will make a significant contribution to the literature.

There are a lot of different specificity scores or algorithms in literature. If we investigate their advantages and disadvantages, we understand that tau score is the best although it has some deficiencies. Assis and Bachtrog studied tissue-specific genes via tau score in their research about young duplicate genes in *Drosophila* [289]. Bush et al. used tau score to explain relationship between lineage-specific sequence evolution and expression level in *A. thaliana* [290]. Additionally, tissue-specific genes were calculated to understand tissue and organ evolution in another study [120].

We can see from the examples that; one-dimensional (i.e. one gene mapped to one tissue) tissue specificity indices are generally used. However, it is limited in its capacity to identify and categorize specific genes. To overcome this, we tried to develop a new extensive procedure that find the tissue-specific genes in a rigorous manner. In this thesis study, specific expression of a gene in one or more tissues were calculated using tau as a robust specificity index and statistical distance as a rigorous method via RNA-seq data of 96 human tissues for protein coding genes. Statistically significant interval estimation is a statistical procedure which provides rigorous results. As a result, we used a novel approach that is the integration of tau score and statistical distance procedure

to calculate protein coding tissue-specific genes from RNA-seq data. We also have investigated the fact that how many genes specific to a tissue plays a role in cancer development in that tissue. After all statistical analysis and calculation of gene groups which are called intersected genes, only DEG and only specific-genes, we performed functional annotation to understand their roles in cancer and improve new effective targeted treatment. In addition, we integrated our findings with gene expression data in cancer patients' samples and normal samples obtained from TCGA. In order to recognize the intra-tumoral heterogeneity and cancer cell mechanism, cancer data was evaluated by integrating it with tissue-specific gene information. Some criteria were applied to the combined data for generating reliable results.

In this study, there are two main parts. First part, it contains calculation of tissue-specific genes for 96 different tissue types. Second part, it includes discussion of intra-tumoral heterogeneity and cancer mechanism for various types of cancer.

Before beginning to analyze the data, the distribution of datasets was examined. Kolmogorov- Smirnov test was applied for each dataset. Both K-S test and Q-Q plots showed that distribution of datasets fit normal distribution and expression values are close to each other.

When we tried to investigate gene expression profiles, we obtained null expression, weak expression, wide-spread expression and specific expression gene lists for each dataset. The criteria for each category is described in Section 5.2. The table below summarizes number of genes in each category for each dataset.

Table 7.1 Number of genes in each expression class for all dataset

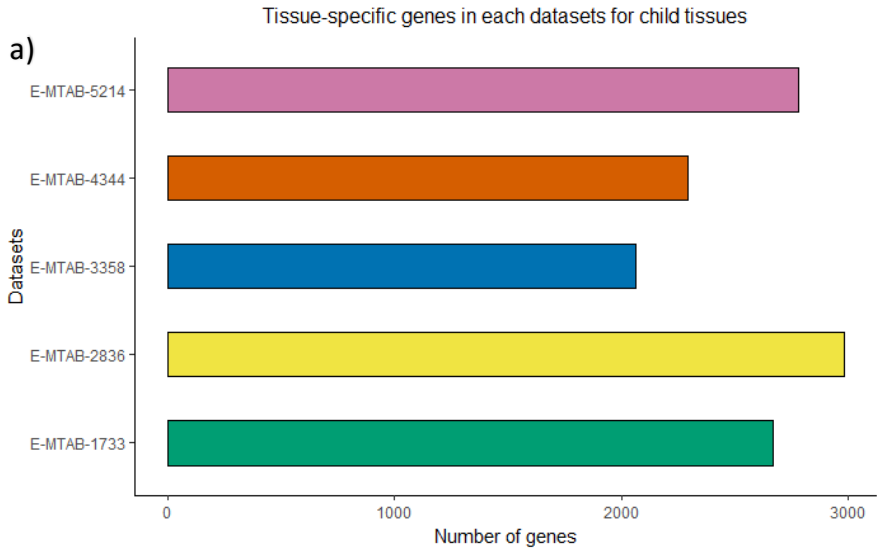
Gene profile	**E-MTAB-1733**	**E-MTAB-2836**	**E-MTAB-5214**	**E-MTAB-3358**	**E-MTAB-4344**
Null expression	1260	2427	2672	1808	3394
Weak expression	1808	2533	2788	3869	2976
Wide-spread expression	13126	11733	11434	8698	11013
Specific expression	2669	2983	2782	2063	2293

According to Table 7.1, the number of tissue-specific genes is comparable between all samples. This result shows us that the method which was used for the determination of tissue specificity and to separate null expressed genes, weakly expressed genes and wide-spread expressed gene is very effective and accurate. If we compare tissue specificity using only tau score with extended tau calculation as our new approach which includes integration of tau score and statistically significantly interval estimation from maximum expression of a gene, our new approach gives more accurate and comprehensive results. We can see comparison of only tau score results and extended tau approach results for each dataset in Table 7.2:

Table 7.2 Comparison of results between only tau and extended tau

Datasets	**Only tau (gen-tissue pairs)**	**Extended tau (gen-tissue pairs)**
EMTAB_1733	2669	3370
EMTAB_2836	2983	4257
EMTAB_5214	2782	4680
EMTAB_3358	2063	3982
EMTAB_4344	2293	3097

According to Table 7.2, number of gene-tissue pairs in new approach is higher than tau score results. Because we assigned genes to multiple tissues, genes might be specific to one or several tissues or cell types. It is quite natural result that the number of genes in the newly developed method is higher.



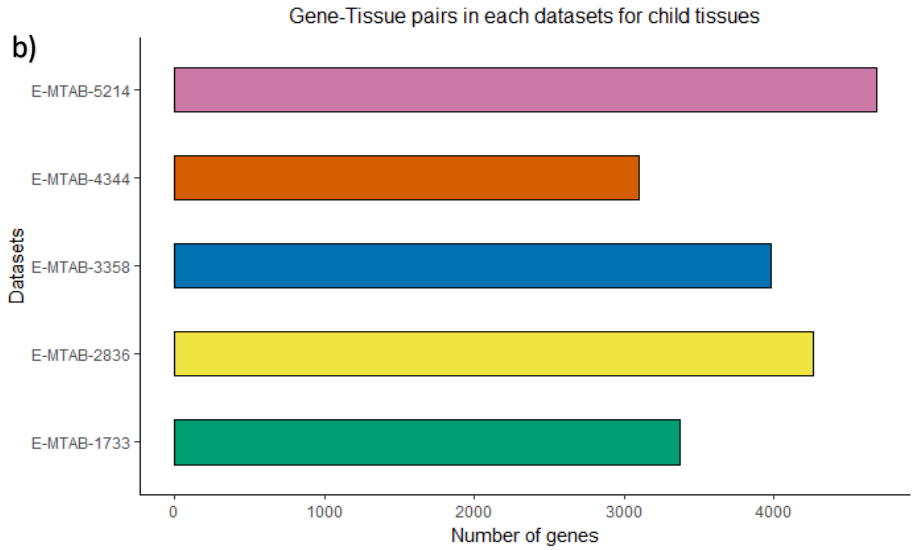


Figure 7.6 Number of tissue-specific genes(a) and genes-tissue pairs (b) (in each dataset for 96 child tissues)

In this study, there are 96 different tissue types, however some tissues are part of other tissues. For instance, cerebellum and brain are different tissue types examined in all datasets we have used. However, cerebellum is a part of brain. Thus, we categorized the all tissues and named cerebellum as “child tissue” and brain is named “parent tissue”. We did parent-child mappings according to Brenda Tissue Ontology. Generally, the results were evaluated separately in both of the two groups. According to Figure 7.1 as a demonstration of Table 7.2, gene-tissue mapping counts are comparable among all datasets for all child tissue counts. The figure shows comparable proportions of each dataset for the number of tissue specific genes, considering all tissues.

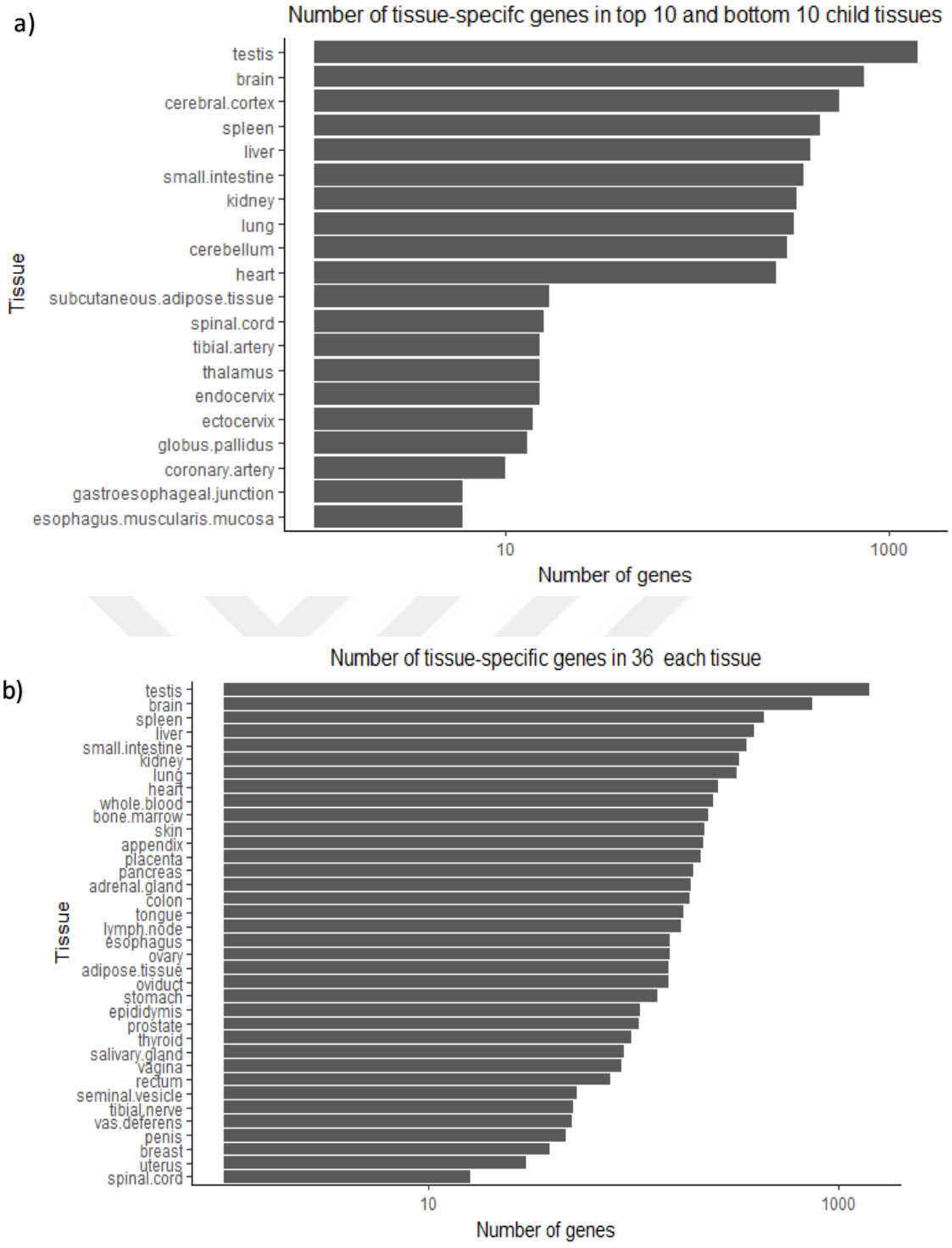


Figure 7.2 Distribution of tissue-specific genes (for top 10 and bottom 10 child (a) and all parent (b) tissues)

Figure 7.2 demonstrates the number of tissue-specific genes per tissues, considering either parent only or child tissues. Testis and brain have highest number of tissue-specific protein coding genes for both parent and child distributions. Table 7.3 lists the

number of tissue-specific genes per tissue. Tissue-specific genes were compared among all datasets and number of them was obtained uniquely.

Table 7.3 Number of tissue-specific genes for parent tissues

Tissues	Number of specific genes	Tissues	Number of specific genes
adipose tissue	150	pancreas	196
adrenal gland	191	penis	47
appendix	220	placenta	213
bone marrow	233	prostate	107
brain	748	rectum	77
breast	39	salivary gland	90
colon	190	seminal vesicle	53
epididymis	108	skin	224
esophagus	152	small intestine	360
heart	261	spinal cord	16
kidney	330	spleen	438
liver	390	stomach	131
lung	320	testis	1427
lymph node	171	thyroid	98
ovary	151	tibial nerve	51
oviduct	149	tongue	177
uterus	30	vas deferens	50

According to our results, testis and brain have highest number of specific genes and liver has the third highest count. If we compare our results with all protein data results of The Human Protein Atlas database, they are in agreement with each other, when we evaluate results as number of tissue-specific genes independent from expression level. Tissue Atlas classified, using proteomics data, approximately 20 000 human genes according to their protein expression across a large number of tissues representing all major organs and tissue types. The database defines “enriched or specific genes” as a gene which is expressed at least five-fold higher in terms of in a particular tissue as compared to all other tissues. Moreover, testis shows the largest number of tissue enriched and specific genes, followed by brain and liver according to Tissue Atlas results, respectively. In this case, the database supports our results in terms of the rank of tissues. There is an important difference between our results and The Human Protein Atlas [291]. On the other side, number of specific genes in each tissue is higher than The Human Protein Atlas because of interval estimation and assigning genes to multiple tissues based on newly generated extended tau method.

Brain tissue is characterized by a high level of gene expression; at least 30–50% of approximately 20,000 known protein coding genes are expressed across all parts of the brain. Moreover, the human brain has the highest level of gene expression compared to other mammals such as mouse, chimpanzee [142], [292], [293], [294].

Testis is the special male tissue responsible for spermatogenesis and steroidogenesis with complex gene expressions and <10% of testis-enriched genes are male-biased in expression. Spermatogenesis is an excellent model for studying the regulation of gene expression during differentiation and testis is important tissue for body building processes [295], [296]. In literature, testis-specific expression is most abundant among the genes exhibiting tissue-specific expression [297]. Yamashita et al. used microarray data to calculate specific genes and they found 1012 testis-specific genes [295]. Our results are correlated with the literature.

In brain, there are genes specifically expressed based on cell types and specific layers or regions of brain. Significant differences in cell composition of the various anatomical brain regions result in cell-specific differences in gene expression. For this reason, there may be many specific-genes all over the brain [298].

Some characteristic functions of liver cells including the capacity to synthesize albumin, urea, and the storage of glycogen are specifically expressed in liver and they were reported in literature. If we compare these with our results, there are high correlation between them. Although cretainin18 and creatinine 19 were defined liver specific genes in previous literature [299], our results suggest that they are not specific to neither the liver nor other tissues. When we examined the RNA-seq data which were used in this study we found that they had shown a wide-spread expression, they have higher expression in liver, though.

Tissue-specific gene perspective is significant for biological mechanism and evolution and organogenesis as well. A research about the evolution of tissues and organs in terms of tissue specificity showed that genes usually in the direction of stronger selection with higher expression in brain. There are also significant partial correlations for esophagus, prostate, adrenal, colon, and endometrium according to their study [122]. Their results have a good correlation with our study, although aim of studies are different from each other.

Figure 7.3 shows another evaluation about tissue specificity results. We used five large datasets and some datasets excluded tissues studied in other datasets. Thus, not every tissue is covered 5 times by the tissue-specific genes. Therefore, a gene may not be specific in a particular tissue according to five datasets.

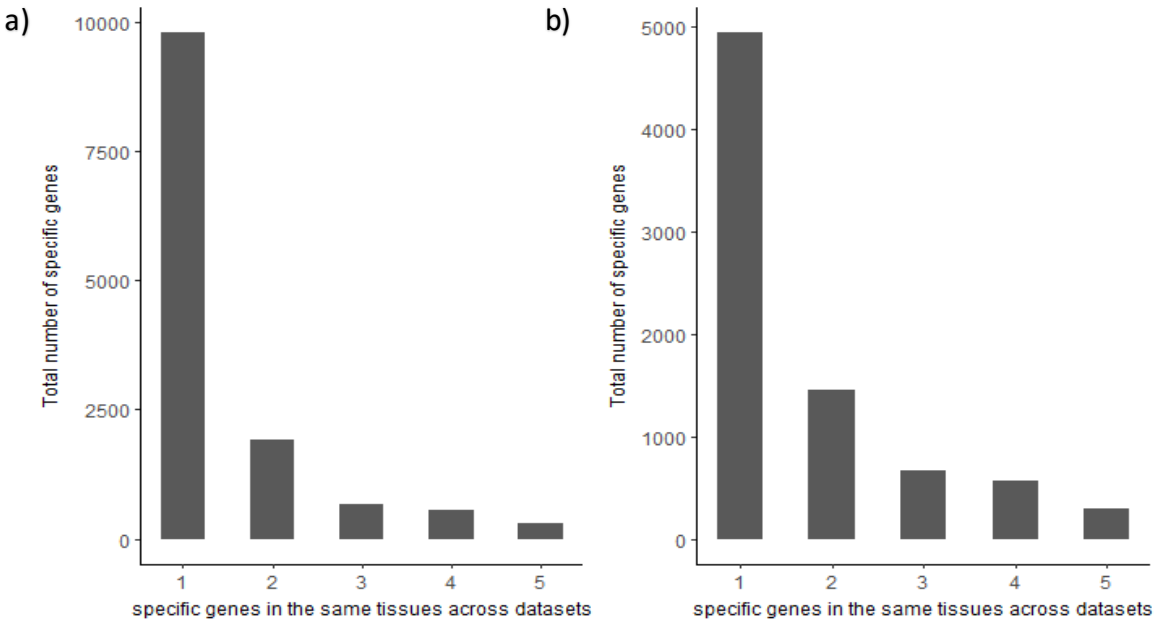


Figure 7.3 Number of same specific genes in the same tissue from different datasets, (a) for child tissues, (b) for parent tissues

If a gene is calculated to be specific according to five different datasets, it absolutely is specific to related tissue. On the other hand, if a gene calculated specifically according to only one dataset, it is specific, but it may need to be re-tested. Number of specific genes according to only one dataset in Figure 7.3 is highest. Besides, genes that are specific for five datasets, is lower in number than other groups. However, we are sure that these genes are specifically expressed in related tissues precisely and five datasets have certain consensus for these genes. The situation described in Figure 7.3 is shown separately according to the datasets in Figure 7.4.

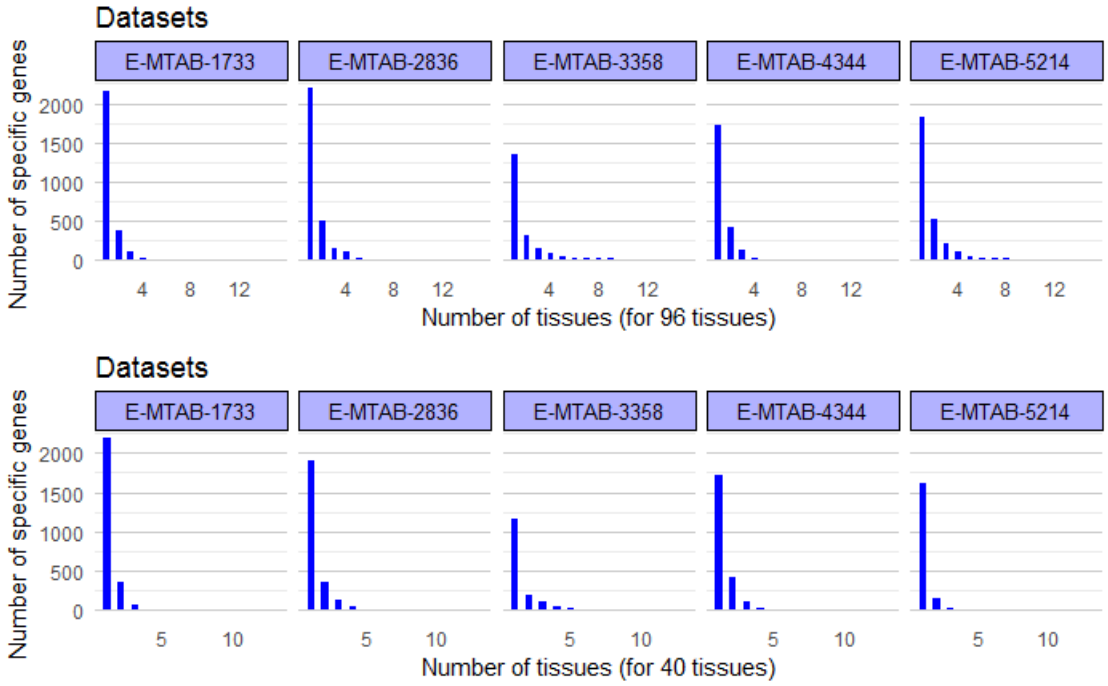


Figure 7.4 How many genes are specific in how many tissues per dataset, (a) for child tissue, (b) for parent tissues

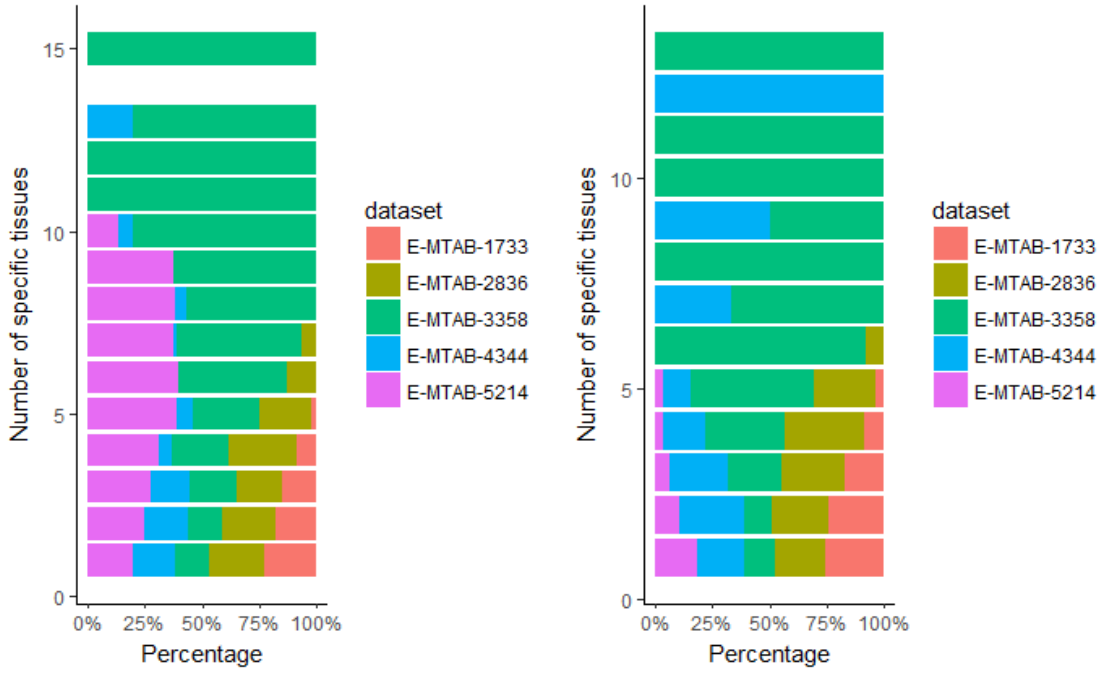


Figure 7.5 How many genes are specific in each dataset (a) for child tissue, b) for parent tissues)

Figure 7.5 shows distribution of tissue-specific genes in each dataset according to how many genes are specific to how many tissues as a ratio. We obtained this figure because we assigned genes to multiple tissues. EMTAB-3358 have a bigger area than other datasets. EMTAB-3358 generally is a little bit different from other data due to its

normalization method. For instance, our results show that EMTAB-1733 dataset does not have a gene which is specific to more than 5 tissues. Actually, we may comment that normalization of RNA-seq data is important for the correctness and consistency of the results. For this reason, data type, normalization methods, usage of extensive data, comparison of data, minimization of errors are special necessities.

Genes were assigned to multiple tissues in this thesis study through cooperation of tau and statistically distance procedure as a new computational approach defines as extended tau calculation method. Genes can be expressed specifically one or several cells or tissues. This phenomenon is the difference between specific-genes and marker genes. One of the main objectives of this study is that genes are assigned to multiple tissues via the improved new effective method. According to graphs, genes generally are specific to only one single tissue. In addition to this some genes can be specifically expressed in two and more different tissues. Number of tissues usually do not exceed five. Moreover, results of both child and parent tissues have a good correlation. Specific genes determined during the thesis are consistent with the specificity definition in the literature.

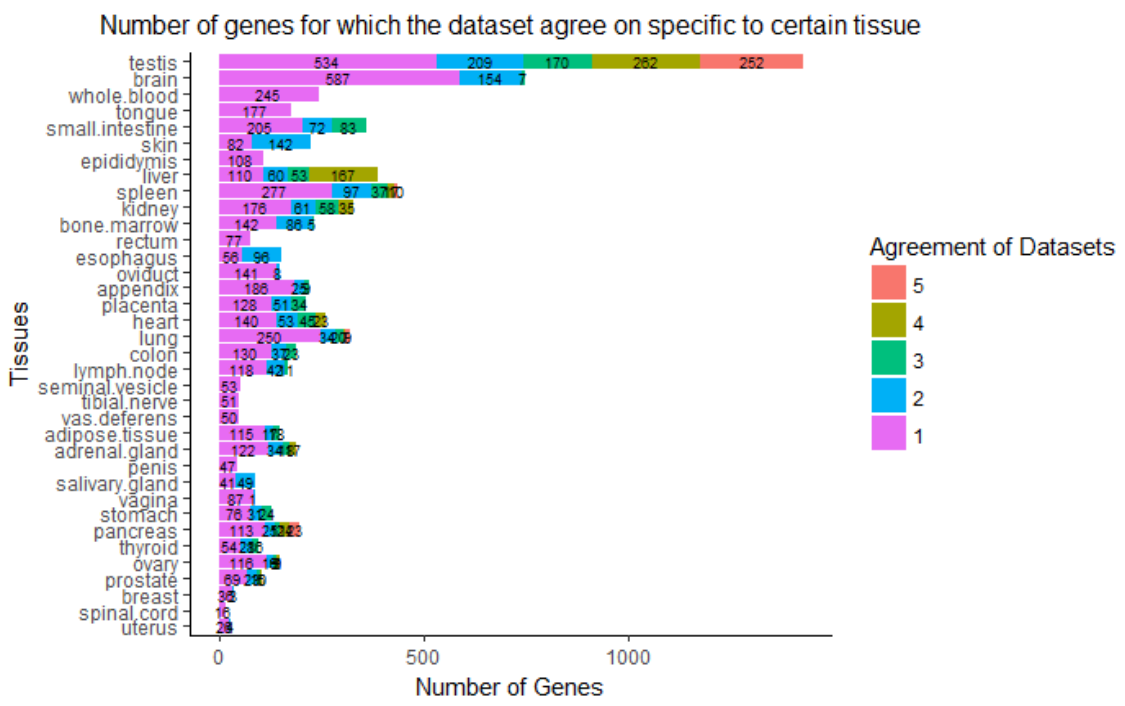


Figure 7.6 Number of tissue-specific genes based on supporting datasets

Figure 7.6 illustrates the distribution of tissue specificity consensus for each tissue. For instance, 252 genes are specific to testis according to all five datasets and 262 genes are specific to testis which is supported by 4 datasets. This result is important for accuracy. On the other hand, if tissues are included only in a single dataset, then all genes specific to that tissue will be supported by only one dataset, naturally. For instance, gene expression in whole blood, tongue, epididymis, seminal vesicle, tibial nerve, vas deferens and spinal cord are examined in only one dataset, therefore genes are specific to these tissues are supported by one dataset only.

Figure 7.7 summarizes the comparison of tissue-specific gene lists across different datasets via Venn Diagram:

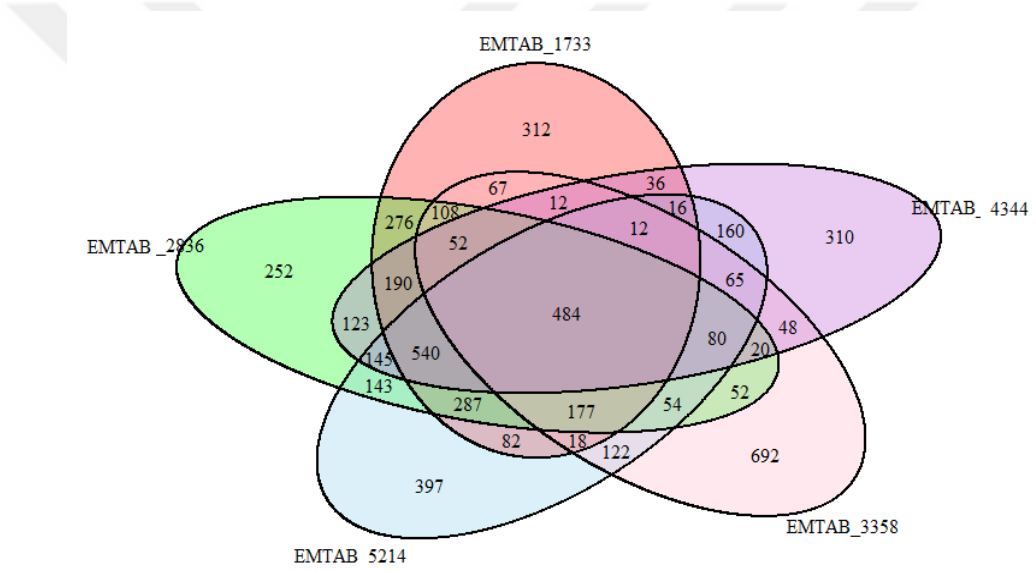


Figure 7.7 Comparison of datasets depends on number of tissue-specific genes

In this study, 484 genes were found to be tissue-specific for all datasets. In other words, all datasets agree on tau score of 484 genes. Although some genes are found to be specifically expressed only in one dataset, this result suggested that our approach is suitable and effective for the determination of tissue-specific genes in RNA-seq data. This comparison is very important for the correctness and reliability of the results. It is natural that there are differences in the results as they are created with different experimental conditions, laboratory conditions, different people's tissues. If we examine Figure 7.7, 692 genes are determined to be tissue-specific by only EMTAB-3358 dataset. This number is the biggest compared to others. Remember that EMTAB-3358 expression

values were normalized using a different method, despite the same experimental procedure. The situation illustrated by the Venn Diagram has been drawn as a network of genes and datasets in Figure 7.8. The dataset nodes are colored as green and labeled with dataset ID. The small red nodes represent genes. There's edge between a gene and a dataset if that dataset reports gene's tau score above 0.85. So, 484 genes mentioned in Figure 7.7 will have 5 edges, i.e., to all datasets, in Figure 7.8. The code to generate this network graph is provided in Appendix-A.

Next, we examined the similarity between datasets according to gene expression values. Figure 7.9 is a correlation plot between datasets in which gene expression values are used to calculate the distance between datasets.

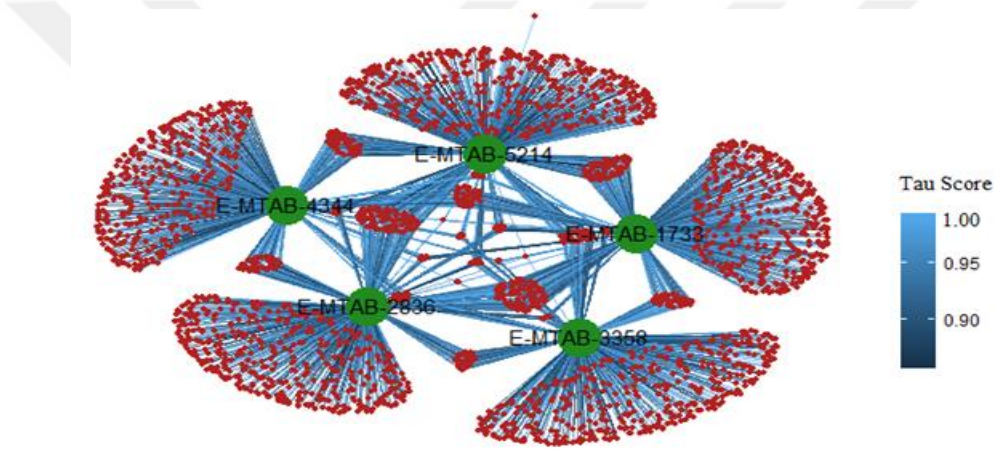


Figure 7.8 Network of tau scores in five datasets

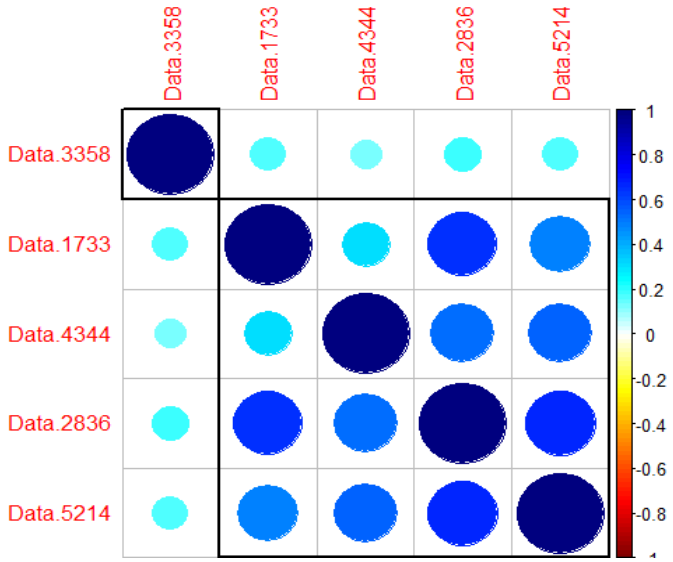


Figure 7.9 Compatibility of datasets with each other based on expression profiles (null, widespread, weak and specific)

Similarity is defined as dark color in these graphs and size of nodes shows number of specific genes which are common between two datasets. If color is dark blue between two datasets, they have high similarity, i.e., similar gene expression profile. Otherwise, if color is light blue, two related datasets are far away from each other as expression values. When we examine this figure, we can show that the similarity of EMTAB-5214 and EMTAB-2836 is higher than any other dataset pair. EMTAB-3358 is quite different from all other datasets as it was observed in the previous results. Figure 7.10 shows the similarity between EMTAB-2836 and EMTAB-5214, dissimilarity between EMTAB-3358 and EMTAB-4344 via scatterplots.

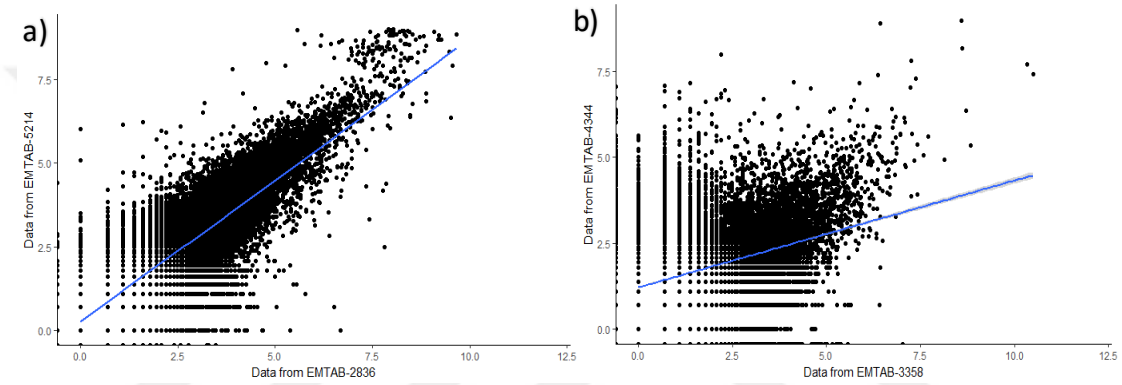


Figure 7.10 Scatter plots for data correlation, a) relationship between two closer datasets (EMTAB-5214 and EMTAB-2836), b) relationship between two unrelated data sets (EMTAB-3358 and EMTAB-4344 (data were log transformed.)

After examining the datasets for similarity for gene expression, in the context of raw data so to speak, the datasets were compared according to tau score results. The datasets were compared for their tau score correlation, for the genes which had tau score greater than 0.85 and accompanying correlation plot has been drawn in Figure 7.11. It was shown that EMTAB-5214 and EMTAB-2836 have high correlation, EMTAB-3358 and EMTAB-4344 have not.

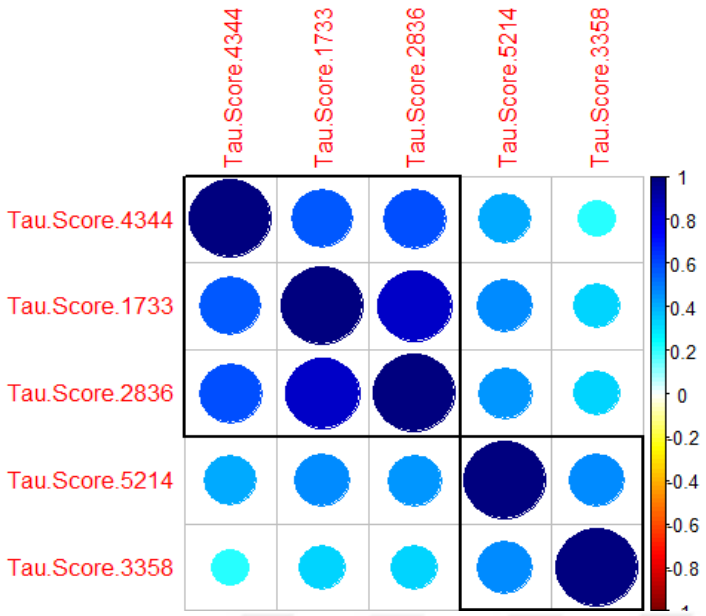


Figure 7.11 Compatibility of datasets with each other based on tau (≥0.85)

Here, EMTAB-4344, EMTAB-1733 and EMTAB-2836 gives similar results after the determination of tissue specificity. After that, EMTAB-5214 appears close to them. As before, EMTAB-3358 is also far from the other datasets. Two data that give the closer results are EMTAB-1733 and EMTAB-2836. On the other hand, two data that give the most distant results are EMTAB-3358 and EMTAB-4344 according to Figure 7.11.

As mentioned before, tissue-specific expression profiles can be used for many different purposes such as the understanding of molecular mechanism of health organism, examining gene profiles for various disorders, enhancing efficiency of therapies, discovering new biomarkers for diagnosis and also targeted treatment of disease like cancer. Except those, there is another thing that is very important and related to all of them, organogenesis. Progression of tissues, organs and systems in living organisms can be understood identifying tissue-specific genes and their roles in organism. In addition to this, interpretation of relationship between tissues in the context of specific-genes is a very crucial approach to decipher not only tissue progress but also some diseases like metastatic cancer. Shared tissue-specific genes between different tissues might give clue to unravel relationships between various tissues. With this purpose in mind, networks showing that relationship were generated for tissues in each dataset. The nodes of networks are tissues and edges are tissue-specific genes in these networks. The size of the nodes indicates the number of tissue-specific genes for that tissue and edge

width is proportional to the number of shared tissue-specific genes. Figure 7.12 depicts the tissue networks based on tissue-specific genes, which discloses the relationships between different tissues.

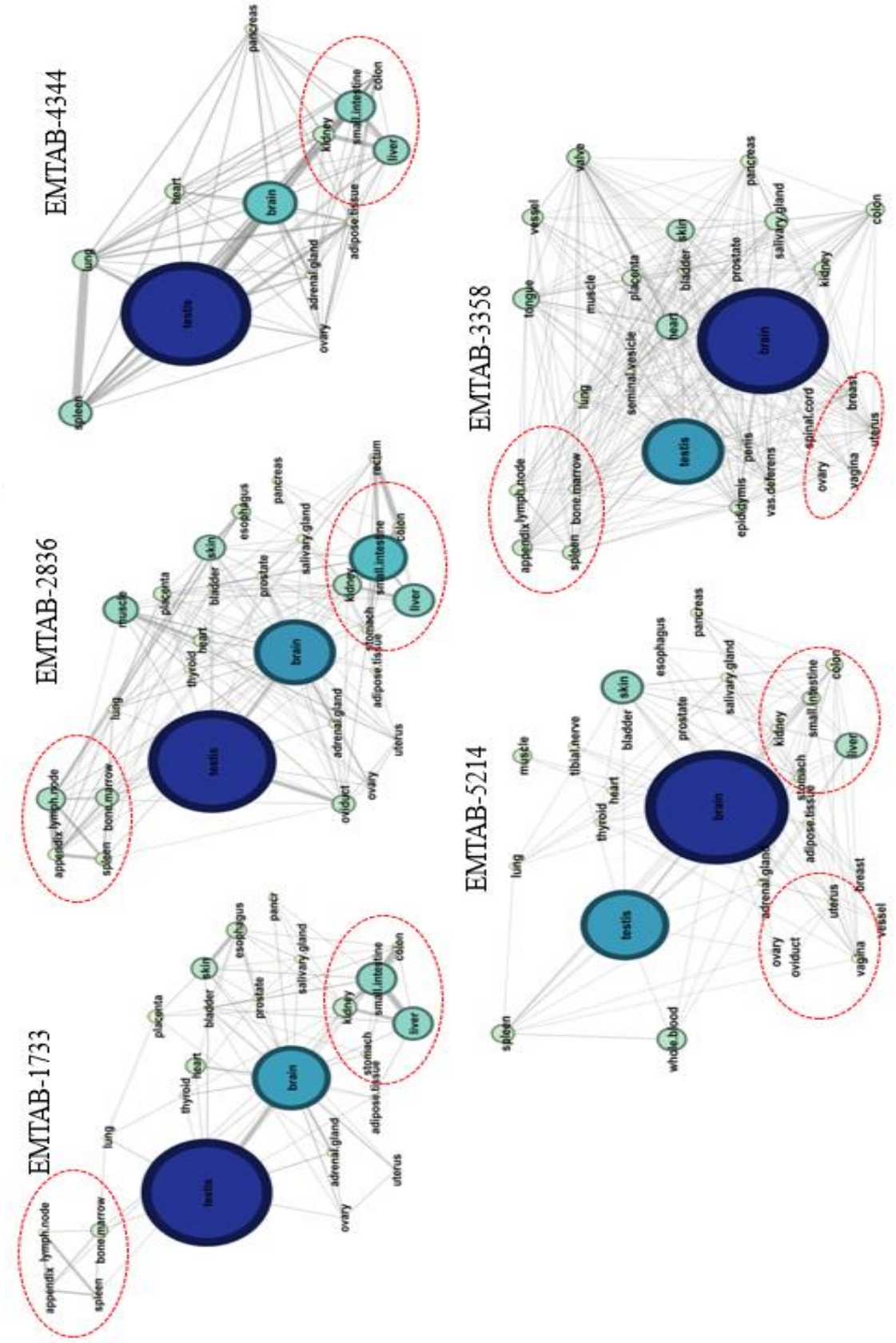


Figure 7.12 Relationship among tissues as a network (in the context of tissue-specific genes for EMTAB-1733, EMTAB-2836, EMTAB-4344, EMTAB-5214 and EMTAB-3358, respectively.)

In these networks, testis and brain have highest number of tissue-specific genes. According to The Human Protein Atlas, transcriptome analysis of human shows that 74% of all proteins are expressed in the brain and 1460 of these genes have higher expression in the brain compared to other tissue types. Interestingly, brain has connections to all of the tissues because brain expression may reflect developmental ontogeny, in other words, stages of developmental processes of human body [300]. In all the networks, bone marrow, spleen and lymph node are tightly connected. We know that they are member of lymph system [301]. Another group of tissues is related to female reproductive system including vagina, uterus, ovary and oviduct suggesting that these tissues express a list of genes which are specifically expressed in these tissues [302]. Another good example is human digestive system organs that are small intestine, colon, rectum, liver and also stomach which share tissue-specific genes. However, it has also been observed that some specific genes related to digestive system are associated with kidney. It is possible because different organs can act in similar subsequent processes. For example, ammonia is toxic and is not readily removed from the circulation by the kidneys. Therefore, hepatocytes convert it to urea, which is less toxic than ammonia and is secreted into the circulation and then eliminated by the kidneys in the urine [303]. As we can see in this example, different organs or tissues can have functions which are related processes and the single gene can be specific the both of two organs. In addition, epididymis, vas deferens, penis and testis are tightly connected to each other as parts of male reproductive system. They have shared tissue-specific genes according to the results and our findings are consistent with both literature and Brenda Tissue Ontology.

7.1.1 Comparison of Tissue Specificity Results with Protein Data

Tissue specificity is an important phenomenon for all developmental stage and diseases. In this study, we aimed to generate a large-tissue specific gene list for a large number of tissues so that it can be used as a robust, rigorous and effective resource providing deeper insight during biological research. In order to validate our findings, expression data from The Human Protein Atlas and its subgroup Tissue Atlas were compared with our list of tissue-specific genes.

After filtration of protein level which is described as "not detected" in data, we can discover matching genes with our specificity results. When we examine these not detected genes in protein level, we noticed that there are several genes which are specifically expressed, and their tau scores are equal or higher than 0.85. Another result of comparison RNA-seq data and protein data, we indicate that expression levels based on RNA-seq data of some specific genes are lower than 10 and number of them is listed in Table 7.4. We had categorized data in Section 5.2 according to expression level. If a gene has lower than 10 as mRNA expression in all tissue samples in each dataset, this expression of related gene was called as "weak expression". Some specific genes might have expression level that is lower than 10 due to assignment of genes to multiple tissues. These genes assigning to several tissues in the context of specificity are not called weak expression because they can have higher than 10 as expression level. For instance, a gene is specifically expressed in pancreas and stomach and its expression level is 14 and 9, respectively. This is not called weak expression because of expression level value in pancreas.

Table 7.4 Number of specific genes that are expressed lower than 10 in RNA-seq

Parent tissue	**Number of specific genes expressed lower than 10**	**Parent tissue**	**Number of specific genes expressed lower than 10**
testis	77	skin	11
spleen	69	lymph node	10
lung	60	prostate	10
brain	48	stomach	10
appendix	38	oviduct	9
pancreas	32	small intestine	9
kidney	26	uterus	9
ovary	22	vagina	9
liver	21	thyroid	8
colon	17	esophagus	6

Table 7.4 Number of specific genes that are expressed lower than 10 in RNA-seq data (cont'd)

placenta	14	rectum	6
epididymis	11	breast	5

When genes with RNA expression less than 10 were counted, the first four tissue are testis, spleen, lung and brain, as shown in Table 7.4. More detailed categorization of genes in this case can be developed in further analysis.

After that, protein data were classified and illustrated as a box plot according to expression level: Not detected, low, medium and high, respectively. Raw protein data obtained from immunochemistry and raw RNA-Seq data from EMTAB datasets were tested for suitability of expression levels shown in Figure 7.13.

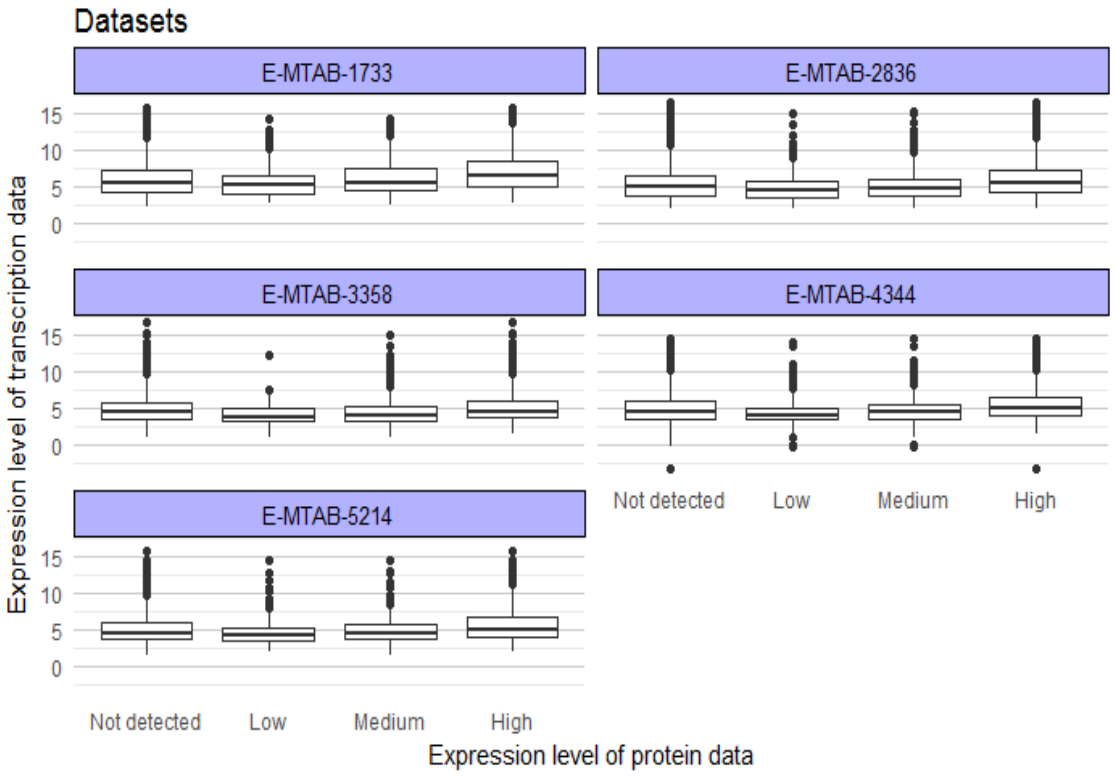


Figure 7.13 Comparison of protein expression and RNA expression of genes (transcription data were log transformed)

Our expectation is that if the protein level is "low", RNA expression level is also low, and this rule is valid for both "medium" and "high" definitions. Not detected protein have generally low expression level in RNA level. They cannot be detected during

experimental processes. According to box plot, two different types of data are compatible with each other, in general. In EMTAB-1733 and EMTAB-5214, RNA expression levels appear to be in concordance with protein expression categories. In contrast, the protein level may be low in database while the RNA level is high in fact like colipase in pancreas. It is specific because of restricted expression toward pancreas [304]. There is a dilemma and it may be caused by experimental errors. In this case, we concluded that RNA-seq data and protein data are not in concordance as raw expression values. Further studies should be performed in general to investigate probable causes of such discrepancy.

Except of those, all protein knowledge from The Human Tissue Atlas was combined with all tissue-specific genes in each dataset. It is supported that the gene at any protein expression level may be specific to the corresponding tissue. In other words, while expression in protein level is low, gene can be specifically expressed. Insulin is specifically expressed in pancreas in high level according to both of two data. Lysyl oxidase like 4 is specific gene to pancreas in low level according to both of two data types. Additionally, if we compare the number of tissue-specific gene lists in our results and The Human Protein Atlas there is a good correlation between them. Brain, testis and liver are the first three tissues in the context of tissue specificity according to both of the two data types as mentioned before.

The final section of this thesis study can be summarized as;

After filtering out transcripts with low level of expression, protein coding genes were assessed for specific expression in tissues. We successfully assigned genes to multiple tissues for specificity in oppose to the assignment to single tissue by original tau study. Results were examined using tables and figures.

All analysis and statistical calculations were performed in R programming language using new algorithms. While improving of a new algorithm or tool, suitability for every data is absolutely important. If developed algorithm is appropriate for any datasets, it is a successful approach. In this context we must use our new approach for the calculation of tissue-specific genes with a lot of datasets and also different experiments like microarray. If we do this, we can advance our computational approach because it is necessary to increase the sensitivity of the algorithm and the accuracy of the results.

7.2 Evaluation of Intra-Tumoral Heterogeneity Results in The Context of DEG

NGS technology and bioinformatic applications contribute to the understanding of human genome in detail, the roles of genes in development, progression, invasion and metastasis of tumors, generating effective and useful therapies using marker molecules.

Tumor heterogeneity is an important obstacle in the better diagnosis and influent treatment of cancer, and it decreases life quality and survival of cancer patients. Aggressive behaviors of tumors originate from heterogeneity [305]. Besides, tumoral heterogeneity can occur not only between patients but also in the same single cancerous tissue or organ in genetic level. This is a major challenge that needs to be elucidated. Furthermore, heterogeneity results in functionally distinct, reduced sensitivity to targeted therapy and poor diagnosis. Thereby, understanding heterogeneity mechanism and role of genes is crucial for medical fields focusing on cancer [179].

Developments in genome sequencing techniques and innovative bioinformatic analyses have provided valuable insight into the remarkable genetic complexity and also heterogeneity due to various cells and tissue layers of malignant tumors. There are many evidences that solid tumors may comprise of subpopulations of cells or tissues that is called intra-tumoral heterogeneity. In recent years, the number of studies to identify heterogeneity increased. Intra-tumoral heterogeneity was shown by different research groups [306], [307], [308]. Genomic heterogeneity of renal cell carcinoma has been recently documented as biomarker interpretation by Gerlinger et al. [180]. Additionally, Kreso et al. investigated the perspective of intra-tumoral heterogeneity in colorectal cancer [309].

Cancer heterogeneity may have profound implications for therapy. On the other hand, diagnosis of cancer is affected badly due to heterogeneity. We know that early diagnosis is vital for cancer patients. The identification of tumor heterogeneity can be useful to overcome clinical challenges partly [310]. For all reasons, a systematic approach to the intra-tumoral heterogeneity or genetic heterogeneity in cancer is required. We can achieve that via using bioinformatic applications. In this study, we tried to understand intra-tumoral heterogeneity through a computational approach because of an easy, dynamic, comprehensive and certain system.

Second main part of the thesis study focused on trying to identify the heterogeneity of various solid tumors. For this purpose, we used our approach to assign tissue specificity published DEG lists for different cancers, firstly. After initial exploration and analyses of DEG, we deflected the study into the cancer expression data for a lot of patients with various tumors.

DEG for various cancers and tissue-specific genes were intersected to unveil potential biomarkers for tumors specifically and describing intra-tumoral heterogeneity for each cancer type. As a result, the number of overlapping genes were found to be quite small, suggesting that DEG in a particular cancer are not specific to its tissue. Differentially expressed genes can be up regulated or down regulated in a particular cancer type. For this reason, number of overlapping genes can be small because tissue specific genes have restrictedly higher expression in related healthy tissue. These intersected gene groups are important for the targeted treatment of solid tumors, especially. In addition, this small number of genes can be used as biomarkers for early diagnosis because these intersected genes are specific to related tissue and their expressions are up regulated or down regulated in the cancer which is originated from the same tissue. This situation can be appropriate to enhance specific diagnostic method for cancer patients. When we analyzed results, we noticed that many genes differentially expressed in a cancer type might contain genes specific to other tissues. These results can suggest that cancerous tissue samples might be heterogeneous and cancer network might have complex interplay with tissue-specific gene expression. On the other hand, tissue-specific genes may provide insight about tumor heterogeneity, malignancy and metastasis to other tissues.

After all analyses with tissue-specific genes and DEG for each cancer type, some gene groups were compiled. These overlapping genes are only differentially expressed for related cancer and only specific genes for related tissue. Number of genes in each group was presented in Table 7.5.

Table 7.5 Number of genes in each group after analyses

Cancer type	Number of DEG	Number of tissue-specific genes	Number of overlapping genes	Number of non-overlapping DEG	Number of non-overlapping specific genes
DOID:1612 / Breast cancer [BRCA]	989	39	23	964	16
DOID:3571 / Liver cancer [Livca]	384	390	31	353	359
DOID:5041 / Esophageal cancer [EC]	1699	152	29	1661	123
DOID:263 / Kidney cancer [Kidca] & DOID:4471 / Chromophobe adenocarcinoma	6	330	0	6	330
DOID:263 / Kidney cancer [Kidca] & DOID:4467 / Renal clear cell carcinoma	2421	330	217	2195	113
DOID:263 / Kidney cancer [Kidca] & DOID:4465 / Papillary renal cell carcinoma	425	330	38	385	292
DOID:1324 / Lung cancer [Lunca] & DOID:3907 / Lung squamous cell carcinoma	2568	320	257	2292	63
DOID:1324 / Lung cancer [Lunca] & DOID:3910 / Lung adenocarcinoma	17	320	0	17	320
DOID:1793 / Pancreatic cancer [PACA] & DOID:4074 / Pancreas adenocarcinoma	108	196	9	99	187
DOID:10283 / Prostate cancer [PCa]	918	107	30	886	77
DOID:1993 / Rectum cancer [Recca] & DOID:1996 / Rectum adenocarcinoma	6	77	0	6	77
DOID:10534 / Stomach cancer [Stoca]	537	131	13	523	118
DOID:1781 / Thyroid cancer [Thyca]	502	98	15	486	83
DOID:363 / Uterine cancer [Uteca]	1607	30	21	1579	9
DOID:4362 / Cervical cancer [Cerca] & DOID:3744 / Cervical squamous cell carcinoma & Endocervical adenocarcinoma	1289	39	27	1263	12
DOID:11054 / Urinary bladder cancer [UBC]	1640	46	16	1615	30

According to Table 7.5, the intersection of two main groups is low, it is a good result because these genes can be used for both targeted therapeutics or biomarkers for diagnosis. Their functions and roles in molecular mechanisms are very significant. Thereby, their functional annotations were carried out in the context of this thesis study. Some cancers with smaller number of intersected genes do not give certain and meaningful results. They should be investigated more extensively in further studies. Rectum cancer, lung adenocarcinoma and chromophobe adenocarcinoma could not be considered comprehensively because of their smaller number of overlapping genes. If we want to examine intersected genes for functional annotation we can use DAVID functional annotation tool. We investigate related disease, GO terms and pathways for intersected genes in order to understand their roles and functions in the molecular mechanisms of cancer.

As a result,

- Obtaining uterus cancer DEG after intersection, some GO terms about biological processes was examined and determined according to Benjamini-Hochberg score. Terms seem to be related to reproductive system such as uterus development, embryonic skeletal system development, adrenal gland development. Moreover, genes have been confirmed to be uterine specific using Uniprot data.
- Obtaining pancreas cancer DEG after intersection, similarity of result was confirmed after comparison with Uniprot depends on specific expression in pancreas. Some related pancreatic diseases were observed.
- Obtaining lung cancer DEG after intersection, diseases such as cancer, infections and immune system related diseases were identified. GO terms as biological processes: immune response, chemotaxis, respiratory gaseous change, chemokine-mediated signaled pathways, inflammatory response and leukocyte migration, as cell component: cell surface, lamellar body, integral component of plasma membrane, clathrin-coated endocytic vesicle, collagen trimer, as molecular pathway: carbohydrate binding was shown. Genes are related to tuberculosis pathway. Moreover, genes are specifically expressed in lung according to Uniprot.

- Obtaining kidney cancer DEG after intersection, genes are concentrated in renal diseases, and also related to certainly excretion, sodium ion transport, ion transmembrane transport, calcium ion homeostasis, phosphate ion transport, urate metabolic process, sodium-dependent phosphate transport, anion transmembrane transport and other ion transportations as biological processes; apical plasma membrane, extracellular exosome, vacuolar proton-transporting V-type ATPase complex as cellular component; anion:anion antiporter activity, inorganic anion exchanger activity, sodium:phosphate symporter activity, hydrogen ion transmembrane transporter activity as molecular functions. They are associated with metabolic pathways according to KEGG database for pathways. The genes are completely concentrated in the kidney according to Uniprot.
- Obtaining esophagus cancer DEG after intersection, they are associated certainly with sarcoplasmic reticulum membrane and keratin filament as cellular components. The genes are related to tongue and esophagus according to Uniprot.
- Obtaining prostate cancer DEG after intersection, genes were found to be related to skeletal muscle and prostate. No certain targets could be observed for prostate cancer.
- Obtaining liver cancer DEG after intersection, cancer, HIV infection were shown. High-density lipoprotein particle, organelle membrane and collagen trimer as molecular components, metabolic pathways were found to be in relation with corresponding intersected genes. The genes are completely concentrated in the liver according to Uniprot.
- Obtaining breast cancer DEG after intersection, lipid metabolic process, uterus development, regulation of apoptotic process as biological processes was examined. Intersected genes could not be shown as targets.

Results are meaningful in some cancers but not significant in others in the context of functional annotation of intersected genes. On the other hand, all intersected genes are not strictly target, they are worth examining of their functions in the body.

In this thesis study, genes and related miRNAs were used to generate networks. Intersected genes can be used as markers in cancer diagnosis if they have certain function. The miRNAs related to these genes are also very important in this respect. Unfortunately, a specific miRNA which regulate the related genes has not been observed. Obtained miRNAs are associated with various cancer types and other diseases.

After examination and identification of functional properties of intersected genes, only DEGs not overlapping with tissue specific genes were analyzed to see whether it is specific to other tissues or not. This approach can give us insight about intra-tumoral heterogeneity. After examination of only DEG excluding intersected genes, we notice that genes are specific to other tissues. Figure 7.14 illustrates our findings using liver cancer as example.

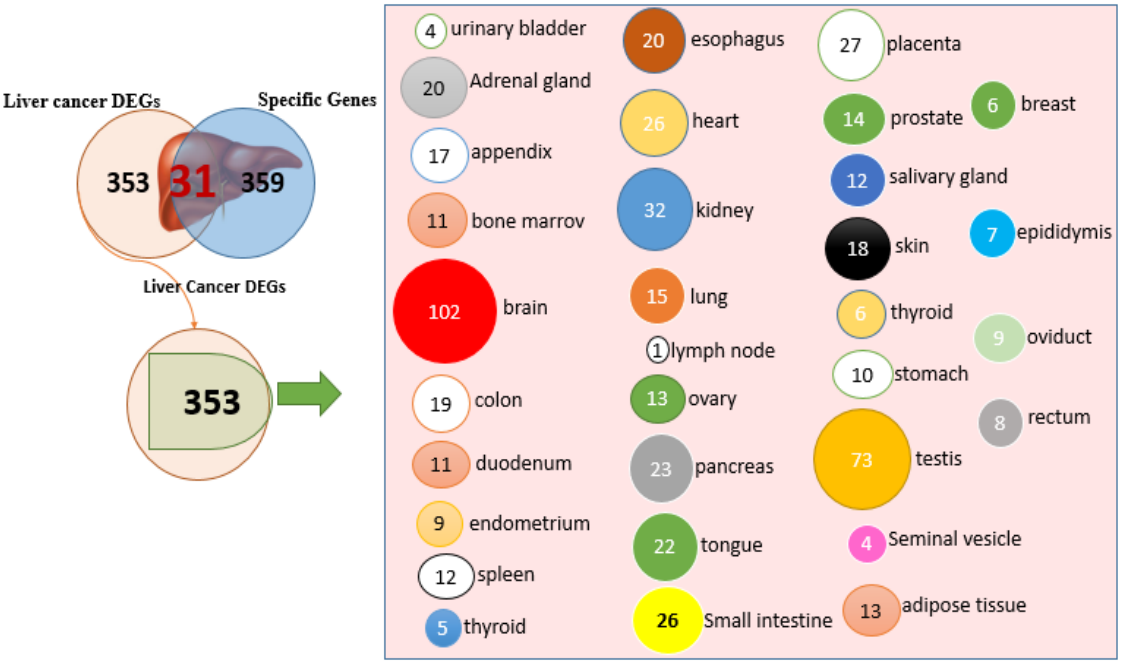


Figure 7.14 Illustration of number of tissue specific genes excluding healthy liver tissue in a single liver tumor

In liver cancer example illustrated in Figure 7.14, 384 differentially expressed genes are up or down regulated in cancerous liver tissue obtaining from liver cancer patients via RNA-sequencing experiments. When two groups of genes are intersected, it is shown that 353 DEG are not specific to liver, however they have altered expression in liver cancer patients. If we examine these genes in the context of specificity to other tissues,

we showed that these genes can be specific to other irrelevant tissues. Number of specific genes in brain and testis is higher than others. This heterogeneous nature of liver cancer is also observed in other cancer types. So non-overlapping differentially expressed genes in other 16 different cancer types which were examined in this thesis study have similar behaviors with liver cancer in the context of specifically expressed genes in unrelated healthy tissues. It is quite puzzling that all cancers have the same distribution among other tissues which might indicate calculation or conceptual shortcomings in our methods Furthermore, brain and testis have large number of tissue-specific genes in almost all cancer DEG. Figure 7.14 clearly demonstrates this phenomenon in liver cancer as an example. However, we know that differentially expressed genes can be up regulated or down regulated in a single cancer. A down regulated DEG in liver cancer, for example, may have specifically higher expression in another tissue/organ type. It is natural condition and also it cannot demonstrate tumoral heterogeneity. For this reason, DEG analyses associated with tissue-specific genes are not enough to discuss and explain intra-tumoral heterogeneity and cancer cell mechanisms in a high accuracy. In order to understand the tumoral heterogeneity in the context of tissue-specific genes up regulated or restrictedly higher expression of genes in various cancer types must be examined. Hence, gene expression data for cancer patients was downloaded from TCGA and used for interpretation of intra-tumoral heterogeneity.

7.3 Evaluation of Intra-Tumoral Heterogeneity Results in Context of Gene Expression of Cancer Patients

Gene expression profiling has facilitated various aspects in cancer such as defining prognosis, stage, biomarkers and targeted treatment, etc. Although almost all the genetic variability and mutations of cancers has been identified for decades, gene expression profiling has a huge importance in recent years to demonstrate the heterogeneity on the genomic level in various cancer [311]. This is why describing intra-tumoral heterogeneity is essential for an effective therapy and increasing cancer patients' life quality. Computerized algorithms can be designed in order to identify features of tumor microenvironments in terms of intra-tumoral heterogeneity using gene expression in cancer and specifically expressed genes in tissues effectively.

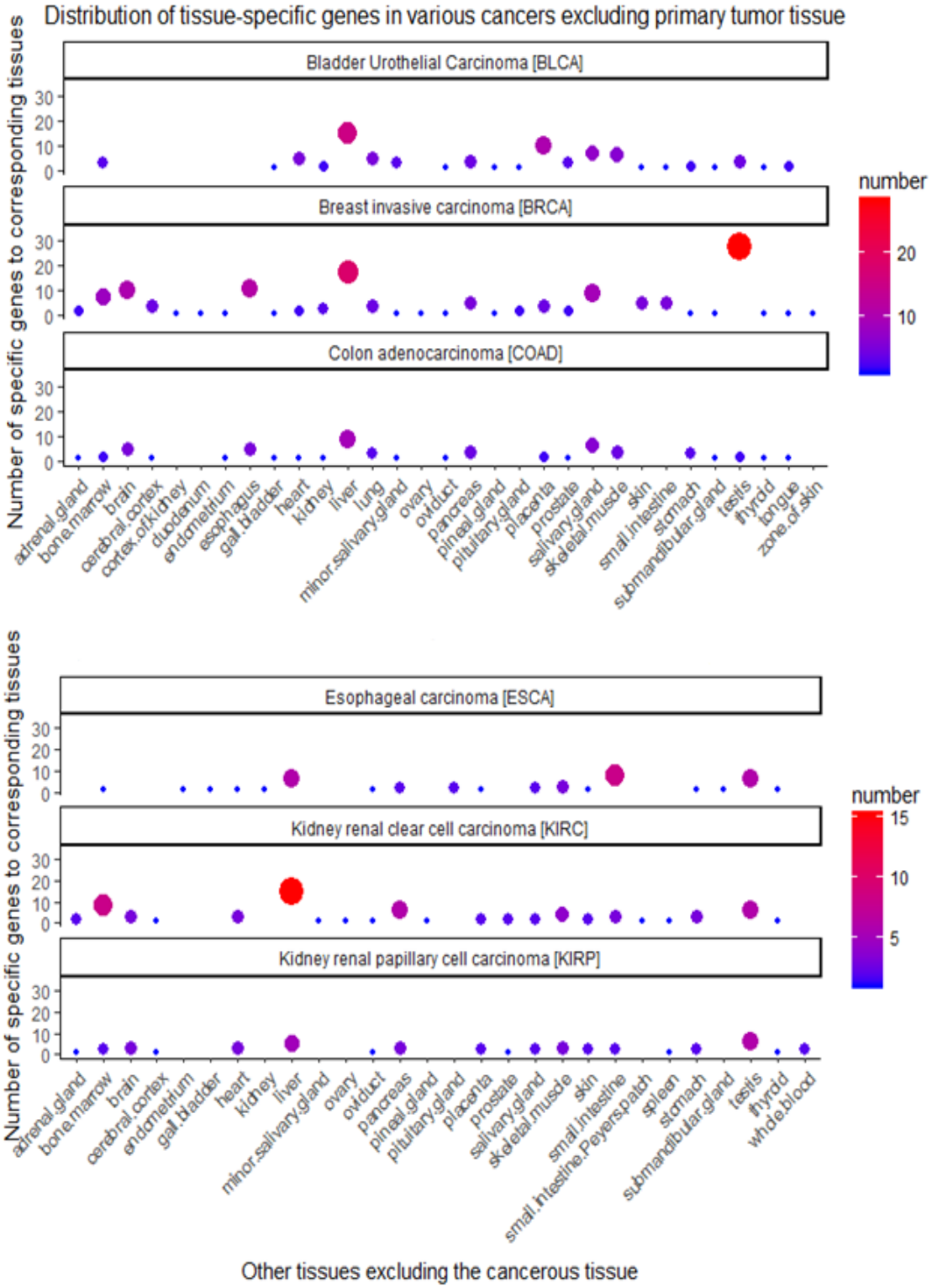
Gene expression in a single cancer for a lot of patients can give important signatures for tumoral heterogeneity among patients and in the body of a patient. This approach also provides a good perspective on tumor microenvironment and cancer cell behaviors. Validated specificity of gene expression can be meaningful for heterogeneity and cancer cell behaviors in terms of diagnosis, prognosis or effective therapy. In this Section, the results of analysis done with the expression data of various cancers downloaded from TCGA portal were integrated with our tissue-specificity results.

Heterogeneity in tumors was explained in Section 4.6 and discussed in Section 7.2. Tumor microenvironment is a term which is closely associated with heterogeneity because different cells which have various molecular functions in the body can migrate to the cancerous tissue to contact or support tumor stroma. Tumor microenvironment is composed of non-cancer cells and their stroma. Although, these cells are not cancer cells, they form the tumor stroma in the corresponding cancer tissue. This structure has been implicated in the regulation of cancer cell growth with the secretion of some molecules. Tumor stroma consists of various cell types including angiogenic vascular cells (endothelial cells and pericytes), fibroblasts, smooth muscle cells (a-smooth muscle actin (a-SMA) myofibroblasts), glial, epithelial, adipocytes and immune cells like neutrophils, macrophages, T-cells, B-cells, dendritic cells, lymphocytes and mesenchymal stem cells [301], [312], [313], [314]. Extracellular matrix (ECM) and secreted molecules support to the tumor stroma. These cells migrate to the cancer tissue and cause formation a heterogeneous structure at site. While none of these cells are themselves malignant, they gain an abnormal phenotype and can lose their functions related with their origin after a while due to their interactions with each other, and directly or indirectly with the cancer cells [312]. Migrated cells from other parts of the body support tumor stroma by the secretion of some molecules. Their secretions are relevant to tissues or organs of their origin. Hence, healthy tissue-specific gene expression can be used to explain role of cells in tumor microenvironment and tumoral heterogeneity. For example, T cells originated from bone marrow with important roles as increasing tumor prognosis are included in tumor microenvironment [313]. Therefore, bone marrow specific gene expression can be seen in various cancer gene expression since a sample from tumor for gene profiling is a bulk tissue consisting of

tumor cells and their microenvironment. Other cell types are likely to be found which are in tumor stroma due to intra-tumoral heterogeneity. As a result, this interesting situation can easily be explained with an example like colorectal cancer. In colorectal cancer, there can be seen bone marrow specific-genes during the examination of colorectal cancer gene expression profile because we know that T-cells are located in colorectal cancer microenvironment and they cause also intra-tumoral heterogeneity [315]. In this thesis study, tissue-specific genes and cancer gene expression were associated to reveal this situation in different types of cancer.

There are 11 different cancer types in this study downloaded from TCGA portal. They are primary solid tumors including liver, breast, lung, colon, esophagus, kidney, pancreas, thyroid cancer and more. Primary tumor samples were preferred in this study. If we use secondary tumor samples, they can contain other cell types originated from primary tumor tissue. When primary tumor cells migrate the other parts of body and colonize there, secondary tumor occurs as mentioned Section 4.7. Metastatic cancer cells have features similar to primary tumor cells because of the expression of proteins characteristic to the cell type from which it arose. So, after migration to the secondary side of the body, these cells can show their specific functions in the secondary tumor. In this way, medical doctor can determine origin of a secondary tumor. On the other hand, for instance, liver cancer expresses some of but not all the proteins characteristic of normal healthy liver cells and after a while, they may ultimately evolve to a state in which they lack most liver-specific functions [316]. Behaviors of metastatic cancer cells are different compared to cancer cell in a primary tumor because they have colonization potential in other tissues or organs in the body. This is possible with genetic and epigenetic alterations in these metastatic cells [317]. They can cause genetic heterogeneity in secondary tumor except other supportive cells in tumor microenvironment. Moreover, association of tissues or organs in specific gene levels can give important clues about metastasis. For instance, colorectal cancer and stomach cancer can easily metastasize to liver tissue in advanced stage [318]. If we know connection of tissues using specific-genes we can predict the tissues to which cancer will metastasize. This is another invaluable phenomenon of heterogeneity. However, the main purpose of our study is to investigate intra-tumoral heterogeneity in a single

cancerous tissue caused by stromal microenvironment. To achieve this goal, tissue-specific genes for various healthy tissues/organs and cancer gene expression data were merged to examine the heterogeneity. Some criteria have been set for this objective. When we analyzed a cancer type in itself, even if a gene is specifically and restrictedly expressed in a tissue (expression percentage is higher than 95% of all heathy tissue sample), it was shown that gene has high expression level in another unrelated cancerous tissue passing the specified criteria in Section 6.5. Moreover, the expression of corresponding gene in normal sample of cancer patients is very low (lower than 1/10 of tissue that is specifically expressed). After analyses and calculations of significant genes within the context of the above criteria, specifically expressed genes in unrelated cancerous tissue and their numbers were shown in order to recognize them easily. Figure 7.15 summarizes all findings in a comprehensive manner to demonstrate that tissues from which specifically expressed genes are detected in selected cancer data. These important genes specifically expressed in various tissue types but also has significantly high expression in other irrelevant cancerous tissues. Although, this is an unusual suspect, it can help to reveal new information about intra-tumoral heterogeneity and cancer cell genetic mechanisms.


Distribution of tissue-specific genes in various cancers excluding primary tumor tissue
Bladder Urothelial Carcinoma [BLCA]
Breast invasive carcinoma [BRCA]
Colon adenocarcinoma [COAD]
Esophageal carcinoma [ESCA]
Kidney renal clear cell carcinoma [KIRC]
Kidney renal papillary cell carcinoma [KIRP]
Number of specific genes to corresponding tissues
number
adrenal.gland
bone.marrow
brain
cerebral.cortex
cortex.of.kidney
duodenum
endometrium
esophagus
gall.bladder
heart
kidney
liver
lung
minor.salivary.gland
ovary
oviduct
pancreas
pineal.gland
pituitary.gland
placenta
prostate
salivary.gland
skeletal.muscle
skin
small.intestine
small.intestine.Peyers.patch
spleen
stomach
submandibular.gland
testis
thyroid
tongue
whole.blood
zone.of.skin
Other tissues excluding the cancerous tissue

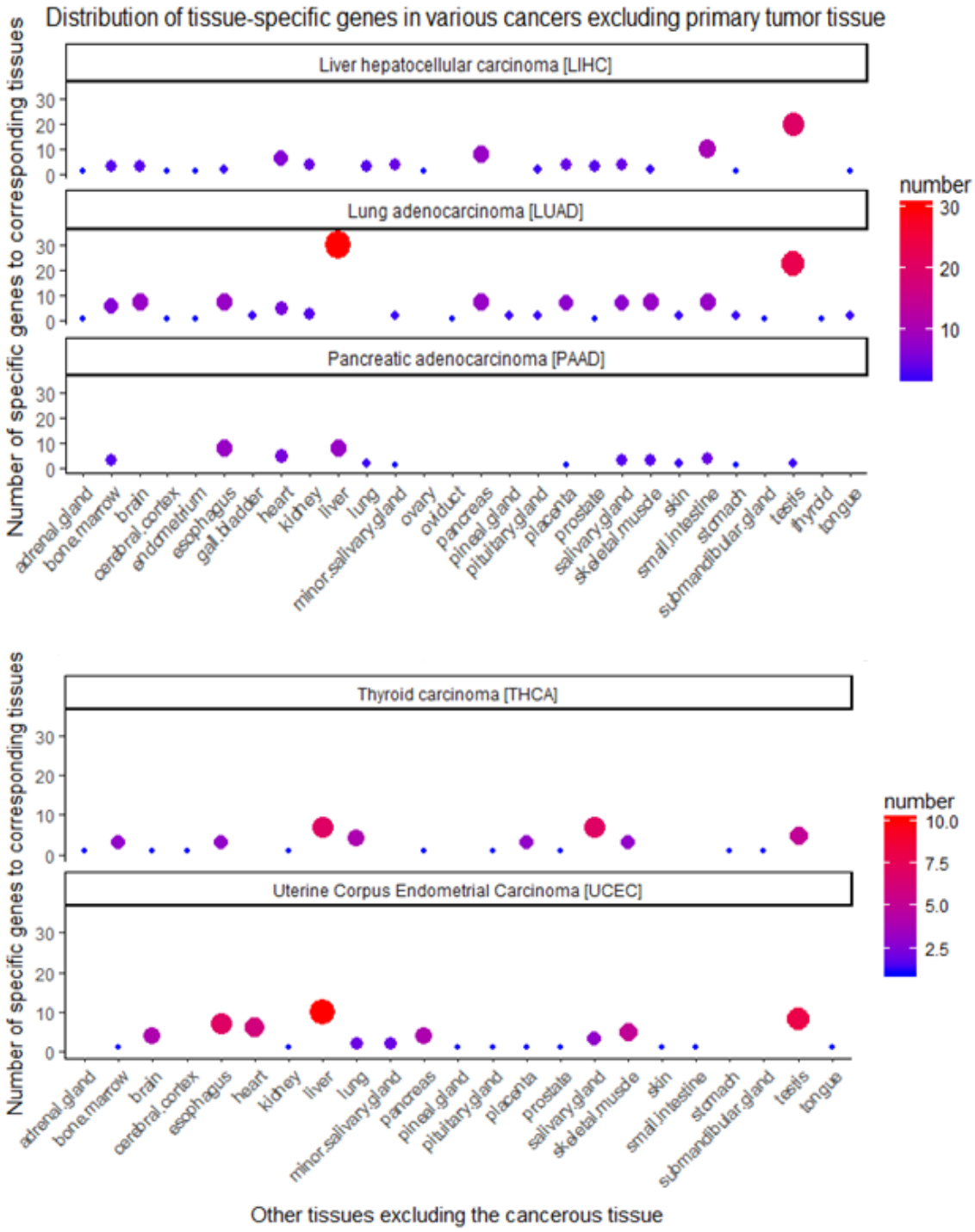


Figure 7.15 Distribution of tissue-specific genes in various cancer types excluding primary tumor tissue

Figure 7.15 shows an interesting table of intra-tumoral heterogeneity in the context of tissue-specific genes. Genes can specifically have expressed in a variety of tissue types excluding related tumorous tissue according to Figure 7.15. Color gradient indicates number of tissue-specific genes.

This illustration presents a very interesting landscape about gene expression in cancer tissues. To give an example, Bladder Urothelial Carcinoma (BLCA), the first cancer type in Figure 7.15, have gene expression which are specifically expressed in other types of tissue instead of urinary bladder. Number of liver specific, placenta specific, salivary gland and skeletal muscle specific genes have higher number in bladder urothelial carcinoma. If we examine all the cancer types in Figure 7.15, specific genes of liver, testis, bone marrow, salivary gland, brain and muscle have higher number of gene expression than other tissues in unrelated cancer tissue. There may be two main reasons which are likely to produce this unexpected outcome. First one is the intra-tumoral heterogeneity in primary tumors due to different cell types in its microenvironment. The second one is complicated mechanisms of cancer cells and their genetic instability leading to activation of genes which are normally expressed in that tissue prior to cancer. Another possible reason can be tissue contamination during the biopsy process from patients for tissues close to each other such as liver and pancreas or lung and esophagus, but it is less effective than other two main explanations. We wondered why specific genes to above tissues are observed in various cancers. Bone marrow and muscle specific genes may be due to cells in the microenvironment of tumor. Cancer cells are antigenic and they can be recognized by host T cells. CD8+ T cells within both tumor and pretumoral stroma have significant positive prognostic effects. Colorectal cancer [315], breast cancer [319], renal cell carcinoma [320], ovarian cancer [321] and lung tumors [322] have been shown to harbor T-cell infiltration. Immunoregulatory properties of dendritic cells in the tumor microenvironment have been suggested in the inhibitory direction of tumor. Lymphocytes in some cases contribute to the support of tumor growth. Macrophages and other myeloid cells are universally found in the solid tumor microenvironment [313]. It is important to consider that the accumulation of bone marrow and muscle originated cells in the tumor microenvironment could be a secondary phenomenon, reflecting a distinct underlying tumor biology. Bone marrow originated cells can have tissue specific gene expression in tumors according to our results. Literature supports this perceptive in terms of tumoral-heterogeneity due to complex microenvironment of tumors. The identification of such genes might be significant improvement for effective immunotherapy.

Otherwise, testis specific genes are shown in a large number and in more cancer types like breast, liver, lung, kidney and uterine cancer, interestingly. The literature shows that there is a group of genes expressed in normal testis restrictedly and have capacity to elicit immune response when expressed in tumor cell that are called cancer/testis genes. Some testis specific genes are mainly expressed in a variety by tumor types such as in lung, ovarian, bladder, breast tumors [323], [324] . This can be verified the our results in Figure 7.15 in the context of testis specific genes. However, testis/cancer genes are not associated with tumor microenvironment features because testis cell do not migrate to the tumor stroma. This is a genetic alteration related with cancer cell genetic mechanisms. Few authors have suggested that these testis specific genes should be considered as cancer-germline genes instead of cancer/testis genes, as they can also be presented in ovary and placenta [323], [324], [325]. According to da Silva et al. [324], FAM290A, MARCH10, GSTF1, TEX19, HORMAD1, C5ORF58 are cancer/testis genes. The same genes are demonstrated in our calculations performed in cancer expression data and testis specific genes. This is an important result of our this study because these genes can be targeted for breast cancer, liver cancer and kidney cancer according to our study. In Figure 7.15 testis gene counts high in almost all cancer types excluding pancreatic cancer. In the research of da Silva et al., they showed that bladder, breast, colon, kidney renal clear cell, lung adenocarcinoma and uterine corpus endometrial carcinoma have cancer/testis genes similar to results shown in Figure 7.15. Our results are in such good agreement with da Silva findings that they did not show pancreatic cancer has cancer/testis genes and so. Figure 7.15 shows no testis specific genes in pancreatic cancer. Besides, other studies suggested that PIWIL2, MAGEB4, BRDT and MAEL [326] are cancer testis antigens for bladder cancer, they also are detected in our study. LEM Domain 1 (LEMD1) was characterized as cancer/testis gene for colorectal cancer [327] , similarly we find out this gene is testis specific according to four RNA-seq data used in the thesis study, and it is restrictedly expressed in colon cancer based on TCGA data. It was seen in colon cancer in this study, thus it can be used as an effective diagnostic marker. Placenta specific genes can also be evaluated and investigated in detail for breast, bladder, lung and thyroid cancers in detail, like cancer/testis genes. Findings in literature and our study suggests that specifically expressed genes in placenta which are expressed in tumor cells are excellent candidates for biomarkers and

targeted therapy. As a conclusion, each gene detected in Figure 7.15 must be comprehensively investigated in future studies because they can be specific targets for a single solid tumor, associated with cancer cell mechanisms which renders them valuable candidates in cancer research.

Above, one of the tissues, testis, which had high number of tissue-specific genes was subject to a detailed analysis. Now, the other tissue, brain, will be analyzed in detail. A microarray-based data to calculate tissue selective genes and overlapping genes with various cancer types was conducted [328], and it showed that brain specific genes were shown in endometrium, kidney, liver, lung, prostate thyroid and colon cancer. Their findings is in agreement with Figure 7.15, which shows a number of brain specific genes in various cancers. Besides, several evidences point out to the possibility of the formation of new nerve endings inside the tumors as microenvironment feature. Nerve endings within the tumors have been expressed in bladder, prostate, breast, pancreas and colon cancer in a review [329]. The role of the nerve fibers in the tumors can be possible by the migration of the perineural invading cells. However, new research has suggested that the nervous system is functionally relevant with tumor progression, modulating a complex network of mediators because of an interaction between nervous system and cancer cells. By the way, cancer cells are also shown to secrete the neurogenic factors [329], [330]. As a result, the appearance of brain specific genes in various tumors in Figure 7.15 can be originated from two fundamental reasons. These are; migration of neuronal cells to tumor microenvironment causing tumoral heterogeneity or secretion of neurologic moleculs by tumor cells due to complicated tumor cell behaviors stimulated by nervous system.

Liver specific genes were also found to have overlapped expression in all cancers which were examined in another study [328] gave comparable results with our study. All cancer types have higher gene expression which are specific to normal liver tissue in this study. Liver plays significant roles in metabolic pathways related to all human body including lipid metabolism, oxidation-reduction process, complement activation [331]. Cancer cell metabolic activity is faster than normal; thus, liver specific genes can be seen in tumor sample due to the complex mechanism of cancer cells and diverse and extensive microenvironment of tumors. As additional information, primary liver cancer

has intrahepatic metastases, lymph node, bone, lung and distant organ metastasis, moreover, the intra-tumor heterogeneity is higher in hepatocellular carcinomas [206], [207], [208], [209].

After some bioinformatic analyses and statistical calculations we determined genes which are highly expressed high level in various cancers and specific to another tissue, which can be explained by tumoral heterogeneity or cancer cell mechanisms. Cellular and molecular components of tumor microenvironment may vary depending on tumor type and location and thus renders each tumor unique [314]. For this reason, each tumor and calculated genes for tumor must be considered and deal with individually. These genes, as a table in Appendix-B, are very important for that cancer research and they must be investigated by bioinformatic analysis and validated in laboratory in further studies.

We can suggest that there are two main phenomena. The first one is tumoral heterogeneity and tumor microenvironment and the second one is cancer cell mechanisms. To approach the problem from a different angle, we integrated miRNA data since miRNAs control the gene expression. For this purpose, we examine related miRNAs and their networks with the calculated genes. For breast cancer, selected tissue specific genes excluding breast tissue after all calculation and their associated miRNAs were demonstrated as networks. Like breast cancer, some other networks are shown as follows:

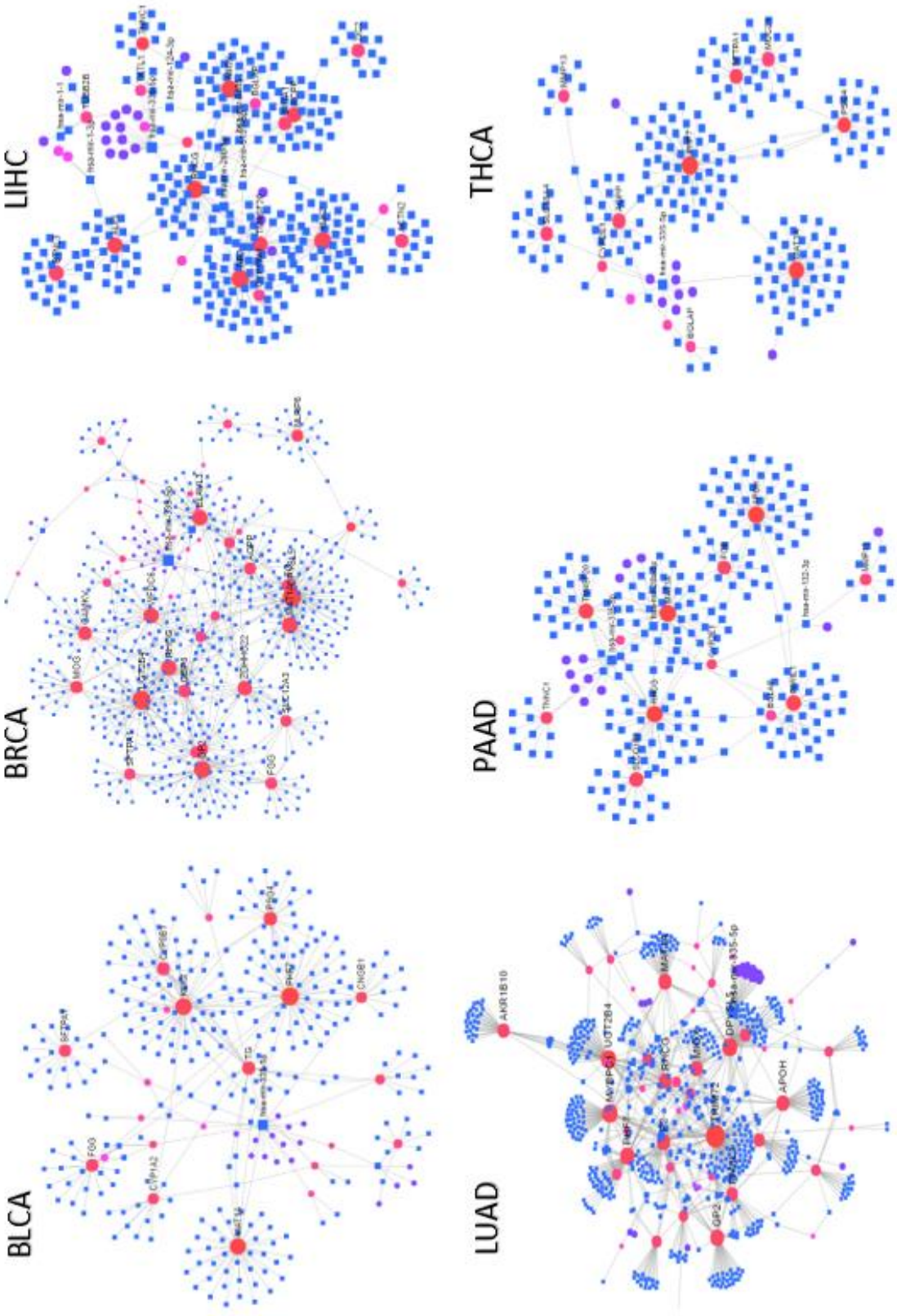


Figure 7.16 miRNAs and targeted genes as a network for selected cancer types

According to miRNA networks in Figure 7.16, **mir335-5p, mir-1-3p, and miR-181** are related to almost all cancer types. These miRNAs are present in the networks of pancreas, colorectal, chronic lymphocytic leukemia, papillary thyroid carcinoma, lung cancer, breast cancer, multiple myeloma, ovarian cancer and uterine cancer [330], [332], [333], [334]. We understood that observed miRNAs are general for cancer mechanisms. Hence, they did not show crucial and specific signatures about tumor heterogeneity or cancer cell mechanisms.

After all, selected genes were examined in detail with their functions, related GO terms especially, biological processes, and enrichment analysis were performed for these genes.

Three genes expressed highly in nine different cancers are specific to muscle, bone marrow and salivary gland. These are troponin T1 slow skeletal type (TNNT1), bone gamma-carboxyglutamate protein (BGLAP) and cystatin SN (CST1) according to tissue specificity, respectively. Extracellular matrix can be supported with them according to GeneCards data, and they can be important for tumor microenvironment. Eight genes expressed strictly in seven or eight cancer types are specific to salivary gland, liver, heart and pancreas in general. Matrix metalloproteinase (MMP13) is one of them and it is very important for tumor stroma because matrix metalloproteinase family is produced by fibroblast or tumor associated fibroblast in tumor microenvironment and MMP-13 was identified as a common up-regulated gene by cancer invasion-related factors [312], [335]. Other genes are generally associated with catabolic metabolism, energy metabolism (negative regulation of ATPase activity) and tissue homeostasis according to GeneCards and enrichment analysis results. Cancer cells can change gene expression to protect themselves and increase migration and invasion capacity. We know that cancer cells have genetic instability. Additionally, there is an effective consensus on the EMTAB data on whether the genes mentioned are specific to tissue or not. 50 genes were shown to be specific to three to five different cancers are specific to muscle, lung, brain, testis, liver and more are enriched different functions like muscle filament sliding, skeletal muscle contraction, multicellular organismal movement, actin-filament based movement and respiratory gaseous exchange. These genes can be related to invasive mechanisms of cancer cells. In addition to all, some genes are specific to a tissue but

expressed in another irrelevant one or two cancerous tissues. As stated before LEMD1, a cancer/testis antigen, is expressed significantly in colon cancer but it is specific to testis.There are more genes showing similar pattern in our study. Dickkopf like acrosomal protein 1(DKKL1) and membrane associated ring-CH-type finger 10 (MARCH10) are testis specific genes,but they are expressed in breast cancer. Moreover, fetal and adult testis expressed 1 (FATE1) and schlafen like 1 (SLFNL1) are specific to testis but they are expressed in kidney cancer subtypes significantly. Such genes have potential to be biomarkers for effective diagnosis and thus require detailed examination and validation. This further analysis can be performed for all selected genes in the context of this thesis study. They can provide a novel perspective to cancer research.

In the context of thesis study, we select six genes from all selected genes to illustrate their expression in 11 different cancers.

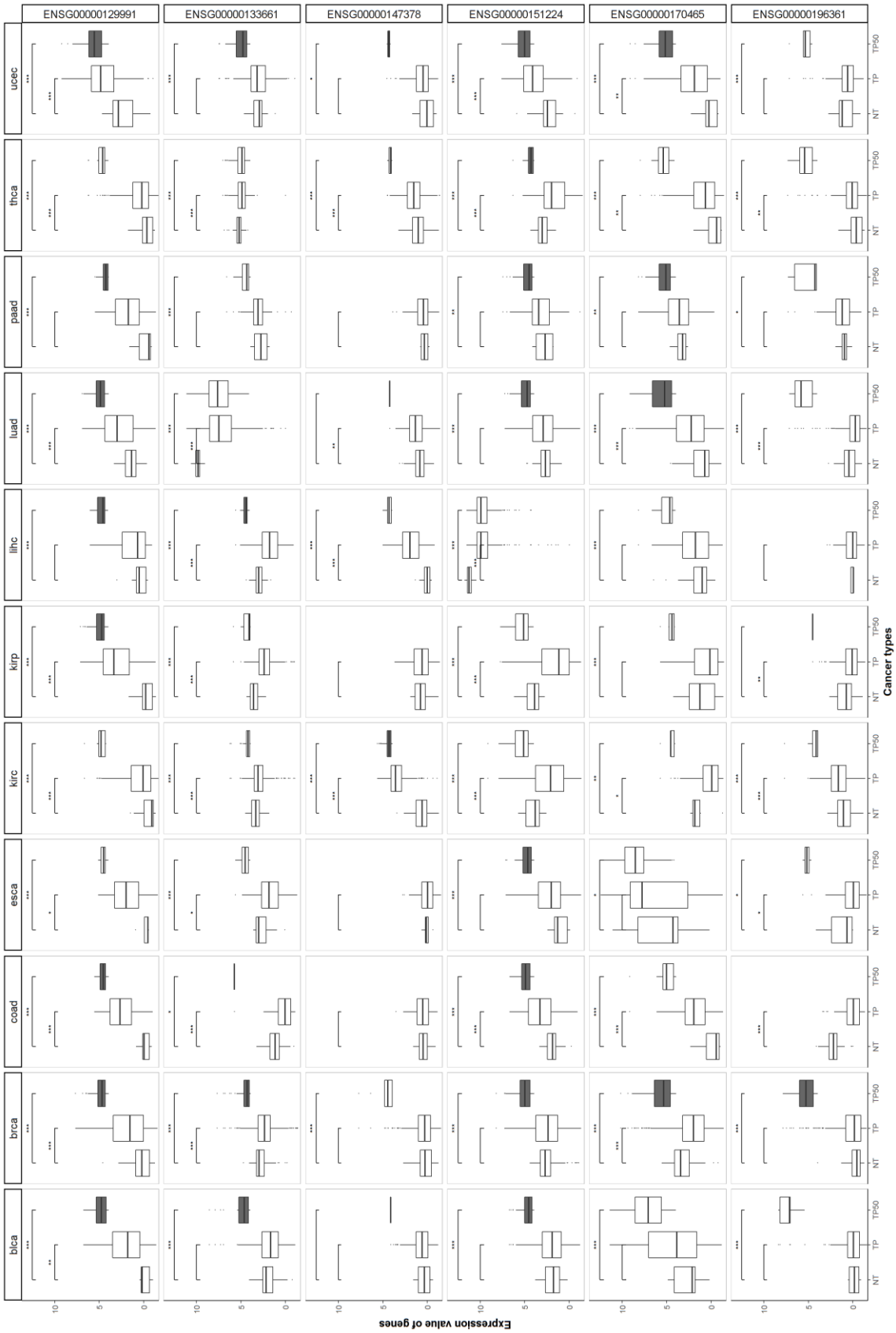


Figure 7.17 Significantly expressed genes in a particular cancer tissue despite being significant to irrelevant tissue

In the Figure 7.17, there are six genes selected from all results are plotted for clarity. NT is normal tissue sample, TP is primary tumor sample and TP50 is expression of primary tumor for genes with expression level is higher than 50 in least 20 patients. T-test was performed with each pair of groups, NT-TP and NT-TP50 to find out significant difference from each other for all cancer types. Box plots filled with dark grey indicate significant TP50 expression level for that gene is given cancer. For instance, in Figure 7.17 troponin I3, cardiac type (ENSG00000129991, TNNI3) is found to be specifically expressed in heart with high consensus of expression data. It has important functions that are vasculogenesis and skeletal and smooth muscle contraction and more. Vasculogenesis is very important formation for migration and angiogenesis of malignant cancer cells from primary tumor tissue to seconder part of the body. Since, angiogenesis is a common feature in solid tumor, it's not surprising to find heart specific gene expression responsible for angiogenesis to be present in eight cancers because of genetic transformation of cancer cells towards angiogenesis process. On the other hand, tumor microenvironment has endothelial cells to support angiogenesis. Heart specific genes can be seen in various cancers due to endothelial cells towards the tumor microenvironment. Thus, our findings suggest that this gene may be an interesting signature of intra-tumoral heterogeneity.

Surfactant protein D (ENSG00000133661, SFTPD) is a specific gene to lung with perfect consensus, i.e., all RNA-Seq data confirm its specificity. SFTPD has meaningful expression in four cancer types according to Figure 7.17. It has significant functions in terms of intra-tumoral heterogeneity like regulation of cytokine production and negative regulation of T cell proliferation. Cytokines are released from stromal cells in tumor microenvironment. Stromal cells migrate other side of the body to support cancer cells to secrete important molecules like cytokines for a strong stroma.

FATE1 (ENSG00000147378) is a testis specific gene with perfect consensus according to EMTAB data. It has important expression in only renal clear cell carcinoma, it may be an essential biomarker for diagnosis and also targeted treatments and hence needs further analysis and validation.

Methionine adenosyl transferase 1A (ENSG00000151224, MAT1A) is a liver specific gene, with high consensus of EMTAB data, related to metabolic processes is expressed

restrictedly in eight different cancer types. Because of the relation with metabolic processes, this gene can be expressed cancer cells in corresponding cancer types. In short, MAT1A may be associated with complicated cancer cell mechanism.

Keratin 6C (ENSG00000170465, KRT6C), specifically expressed in esophagus, is shown to be expressed in four cancers. Its fundamental function is intermediate filament cytoskeleton organization which is related to structure of tumor microenvironment.

The last gene is ELAV like RNA binding protein 3 (ENSG00000196361, ELAVL3) is a brain specific gene. It is expressed only in breast cancer significantly. Its functions are associated with nervous system development and cell differentiation. Neural cells can migrate the tumor microenvironment to support and generate nerve endings in tumor microenvironment according to literature as mentioned before. This phenomenon can demonstrate intra-tumoral heterogeneity due to cell migration. Otherwise, cancer cell can interact with the neural cells to produce nerve endings alone according to other researches as stated before. It is proved of comprehensive and complex behaviors of cancer cells.

All selected genes which passing important criteria after combination of two different data (gene expression in cancer and healthy tissue samples) to investigate tumoral heterogeneity are discussed separately. Some of the selected genes are associated with intra-tumoral heterogeneity and aslo important properties of tumor stroma. Others are associated with complicated and instable genetic mechanisms of cancer cells.

7.4 Conclusion

In this thesis study, tissue-specific genes were determined comprehensively in order to understand tumoral heterogeneity and complex mechanisms of cancer cells in various solid tumors. Newly improved calculational method was described as extended tau for tissue specificity. Results demonstrate that there are some significant genes to understand tumoral heterogeneity and some of them are associated with cancer cells genetic instability and gene alteration of different cancer cells. These genes should be investigated in computationally using bioinformatic analysis and also validated experimentally to give a new perspective to cancer research.

Intra-tumoral heterogeneity can easily be associated with tissue-specific genes. There is general lack of research about intra-tumoral heterogeneity because researches have executed, recently. For this reason, there should be studies focusing on cancer heterogeneity to contribute both medical researches and human life. We generate a new dynamic, effective, rigorous and powerful computational approaches to be easily integrated with the experimental processes, so that less time is consumed and by consequently more validations are performed. However, these analyzes should be repeated with more datasets and also with other data types like microarray. This study is important to provide new knowledge and perspective to literature. Because, tissue specific genes were identified with great care based on extended tau method. These results can be used in various objectives and diseases. After that intra-tumoral heterogeneity was tried to be explained for various tumors and discussed, functional properties of selected genes were investigated to provide new insight about better early diagnosis and effective targeted therapy for various cancers. As a new bioinformatic application, this thesis study has been put forward to increase the life quality of cancer patients and survival, to decrease medical problems. Additionally, determined tissue-specific genes and heterogeneity approaches can be used other fatal diseases in human life.

REFERENCES

[1] Wu, D. Rice, C.M. and Wang, X., (2012). "Cancer bioinformatics: A new approach to systems clinical medicine", BMC bioinformatics, 13: 71.

[2] Dodd, L.G. Kerns, B.J. Dodge, R.K. and Layfield, L.J., (1997). "Intratumoral heterogeneity in primary breast carcinoma: study of concurrent parameters", Journal of surgical oncology, 64: 280-288.

[3] O'Connor, J.P. Rose, C.J. Waterton, J.C. Carano, R.A. Parker, G.J. and Jackson, A., (2015). "Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome", Clinical Cancer Research, 21: 249-257.

[4] Janku, F., (2014). "Tumor heterogeneity in the clinic: is it a real problem?", Therapeutic advances in medical oncology, 6: 43-51.

[5] Oh, P. Li, Y. Yu, J. Durr, E. Krasinska, K.M. Carver, L.A. Testa, J.E. and Schnitzer, J.E., (2004). "Subtractive proteomic mapping of the endothelial surface in lung and solid tumours for tissue-specific therapy", Nature, 429: 629-635.

[6] Hogeweg, P., (2011). "The roots of bioinformatics in theoretical biology", PLoS computational biology, 7: e1002021.

[7] Luscombe, N.M. Greenbaum, D. and Gerstein, M., (2001). "What is bioinformatics? A proposed definition and overview of the field", Methods of information in medicine, 40: 346-358.

[8] Blazer, D.G. and Hernandez, L.M., (2006). Genes, behavior, and the social environment: Moving beyond the nature/nurture debate: National Academies Press.

[9] Eisen, M.B. Spellman, P.T. Brown, P.O. and Botstein, D., (1998). "Cluster analysis and display of genome-wide expression patterns", Proceedings of the National Academy of Sciences, 95: 14863-14868.

[10] Livesey, F. Furukawa, T. Steffen, M. Church, G. and Cepko, C., (2000). "Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene Crx", Current Biology, 10: 301-310.

[11] Šali, A. and Blundell, T.L., (1993). "Comparative protein modelling by satisfaction of spatial restraints", Journal of molecular biology, 234: 779-815.

[12] Venter, J.C. Adams, M.D. Myers, E.W. Li, P.W. Mural, R.J. Sutton, G.G. Smith, H.O. Yandell, M. Evans, C.A. and Holt, R.A., (2001). "The sequence of the human genome", science, 291: 1304-1351.

[13] Collins, F.S. Patrinos, A. Jordan, E. Chakravarti, A. Gesteland, R. and Walters, L., (1998). "New goals for the US human genome project: 1998-2003", science, 282: 682-689.

[14] Ng, S.-K. and Wong, L., (2004). "Accomplishments and challenges in bioinformatics", IT professional, 6: 44-50.

[15] National Cancer Institute, NCI, https://www.cancer.gov/, 10 November 2017.

[16] Al-Rajab, M. and Lu, J., (2012). "Bioinformatics: an overview for cancer research": The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[17] Pubmed, Bioinformatic and Cancer Search, https://www.ncbi.nlm.nih.gov/pubmed/?term=bioinformatics+cancer, 9 November 2017.

[18] Kihara, D. Yang, Y.D. and Hawkins, T., (2006). "Bioinformatics resources for cancer research with an emphasis on gene function and structure prediction tools", Cancer Informatics, 2: 25.

[19] Bensmail, H. and Haoudi, A., (2003). "Postgenomics: proteomics and bioinformatics in cancer research", BioMed research international, 2003: 217-230.

[20] Meric-Bernstam, F. Johnson, A. Holla, V. Bailey, A.M. Brusco, L. Chen, K. Routbort, M. Patel, K.P. Zeng, J. and Kopetz, S., (2015). "A decision support framework for genomically informed investigational cancer therapy", Journal of the National Cancer Institute, 107: djv098.

[21] Margulies, M. Egholm, M. Altman, W.E. Attiya, S. Bader, J.S. Bemben, L.A. Berka, J. Braverman, M.S. Chen, Y.-J. and Chen, Z., (2005). "Genome sequencing in microfabricated high-density picolitre reactors", Nature, 437: 376-380.

[22] Whiteford, N. Skelly, T. Curtis, C. Ritchie, M.E. Löhr, A. Zaranek, A.W. Abnizova, I. and Brown, C., (2009). "Swift: primary data analysis for the Illumina Solexa sequencing platform", Bioinformatics, 25: 2194-2199.

[23] Valouev, A. Ichikawa, J. Tonthat, T. Stuart, J. Ranade, S. Peckham, H. Zeng, K. Malek, J.A. Costa, G. ve McKernan, K., (2008). "A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning", Genome research, 18: 1051-1063.

[24] Parson, W. Strobl, C. Huber, G. Zimmermann, B. Gomes, S.M. Souto, L. Fendt, L. Delport, R. Langit, R. and Wootton, S., (2013). "Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM)", Forensic Science International: Genetics, 7: 543-549.

[25] Ju, J. Kim, D.H. Bi, L. Meng, Q. Bai, X. Li, Z. Li, X. Marma, M.S. Shi, S. and Wu, J., (2006). "Four-color DNA sequencing by synthesis using cleavable fluorescent

nucleotide reversible terminators", Proceedings of the National Academy of Sciences, 103: 19635-19640.

[26] Shendure, J. Porreca, G.J. Reppas, N.B. Lin, X. McCutcheon, J.P. Rosenbaum, A.M. Wang, M.D. Zhang, K. Mitra, R.D. and Church, G.M., (2005). "Accurate multiplex polony sequencing of an evolved bacterial genome", science, 309: 1728-1732.

[27] Pushkarev, D. Neff, N.F. and Quake, S.R., (2009). "Single-molecule sequencing of an individual human genome", Nature biotechnology, 27: 847-850.

[28] Eid, J. Fehr, A. Gray, J. Luong, K. Lyle, J. Otto, G. Peluso, P. Rank, D. Baybayan, P. and Bettman, B., (2009). "Real-time DNA sequencing from single polymerase molecules", science, 323: 133-138.

[29] Sanger, F. Nicklen, S. and Coulson, A.R., (1977). "DNA sequencing with chain-terminating inhibitors", Proceedings of the National Academy of Sciences, 74: 5463-5467.

[30] Maxam, A.M. and Gilbert, W., (1977). "A new method for sequencing DNA", Proceedings of the National Academy of Sciences, 74: 560-564.

[31] Consortium, I.H.G.S., (2004). "Finishing the euchromatic sequence of the human genome", Nature, 431: 931-945.

[32] Van Dijk, E.L. Auger, H. Jaszczyszyn, Y. and Thermes, C., (2014). "Ten years of next-generation sequencing technology", Trends in genetics, 30: 418-426.

[33] McCarthy, D.J. Chen, Y. and Smyth, G.K., (2012). "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation", Nucleic Acids Research, 40: 4288-4297.

[34] Knief, C., (2014). "Analysis of plant microbe interactions in the era of next generation sequencing technologies", Frontiers in plant science, 5.

[35] Rahnenführer, J. and Bozinov, D., (2004). "Hybrid clustering for microarray image analysis combining intensity and shape features", BMC bioinformatics, 5: 47.

[36] Su, Z. Li, Z. Chen, T. Li, Q.-Z. Fang, H. Ding, D. Ge, W. Ning, B. Hong, H. and Perkins, R.G., (2011). "Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys", Chemical research in toxicology, 24: 1486-1493.

[37] Zhao, S. Fung-Leung, W.-P. Bittner, A. Ngo, K. and Liu, X., (2014). "Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells", PloS one, 9: e78644.

[38] Kryuchkova-Mostacci, N. and Robinson-Rechavi, M., (2016). "A benchmark of gene expression tissue-specificity metrics", Briefings in bioinformatics, 18: 205-214.

[39] Vigliar, E. Malapelle, U. Luca, C. Bellevicine, C. and Troncone, G., (2015). "Challenges and opportunities of next-generation sequencing: a cytopathologist's perspective", Cytopathology, 26: 271-283.

[40] Ioannidis, J.P. Allison, D.B. Ball, C.A. Coulibaly, I. Cui, X. Culhane, A.C. Falchi, M. Furlanello, C. Game, L. and Jurman, G., (2009). "Repeatability of published microarray gene expression analyses", Nature genetics, 41: 149-155.

[41] Consortium, G.P., (2010). "A map of human genome variation from population-scale sequencing", Nature, 467: 1061-1073.

[42] Davey, J.W. Hohenlohe, P.A. Etter, P.D. Boone, J.Q. Catchen, J.M. and Blaxter, M.L., (2011). "Genome-wide genetic marker discovery and genotyping using next-generation sequencing", Nature Reviews Genetics, 12: 499-510.

[43] Trapnell, C. Roberts, A. Goff, L. Pertea, G. Kim, D. Kelley, D.R. Pimentel, H. Salzberg, S.L. Rinn, J.L. and Pachter, L., (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks", Nature protocols, 7: 562-578.

[44] Robinson, M.D. and Oshlack, A., (2010). "A scaling normalization method for differential expression analysis of RNA-seq data", Genome biology, 11: R25.

[45] Wang, Z. Gerstein, M. and Snyder, M., (2009). "RNA-Seq: a revolutionary tool for transcriptomics", Nature Reviews Genetics, 10: 57-63.

[46] Mortazavi, A. Williams, B.A. McCue, K. Schaeffer, L. and Wold, B., (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq", Nature methods, 5: 621-628.

[47] Sachidanandam, R. Weissman, D. Schmidt, S.C. Kakol, J.M. Stein, L.D. Marth, G. Sherry, S. Mullikin, J.C. Mortimore, B.J. and Willey, D.L., (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms", Nature, 409: 928-933.

[48] Hill, J.T. Demarest, B.L. Bisgrove, B.W. Gorsi, B. Su, Y.-C. and Yost, H.J., (2013). "MMAPPR: mutation mapping analysis pipeline for pooled RNA-seq", Genome research, 23: 687-697.

[49] Grabherr, M.G. Haas, B.J. Yassour, M. Levin, J.Z. Thompson, D.A. Amit, I. Adiconis, X. Fan, L. Raychowdhury, R. and Zeng, Q., (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome", Nature biotechnology, 29: 644-652.

[50] Degner, J.F. Marioni, J.C. Pai, A.A. Pickrell, J.K. Nkadori, E. Gilad, Y. and Pritchard, J.K., (2009). "Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data", Bioinformatics, 25: 3207-3212.

[51] Trapnell, C. Williams, B.A. Pertea, G. Mortazavi, A. Kwan, G. Van Baren, M.J. Salzberg, S.L. Wold, B.J. and Pachter, L., (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation", Nature biotechnology, 28: 511-515.

[52] Solas, About RNa-Seq, http://cmb.molgen.mpg.de/2ndGenerationSequencing/Solas/RNA-seq.html, 7 November 2017.

[53] Pareek, C.S. Smoczynski, R. and Tretyn, A., (2011). "Sequencing technologies and genome sequencing", Journal of applied genetics, 52: 413-435.

[54] Xi, R. Kim, T.-M. and Park, P.J., (2010). "Detecting structural variations in the human genome using next generation sequencing", Briefings in functional genomics, 9: 405-415.

[55] Vermeer, S. Hoischen, A. Meijer, R.P. Gilissen, C. Neveling, K. Wieskamp, N. de Brouwer, A. Koenig, M. Anheim, M. and Assoum, M., (2010). "Targeted next-generation sequencing of a 12.5 Mb homozygous region reveals ANO10 mutations in patients with autosomal-recessive cerebellar ataxia", The American Journal of Human Genetics, 87: 813-819.

[56] Kuhlenbäumer, G. Hullmann, J. and Appenzeller, S., (2011). "Novel genomic techniques open new avenues in the analysis of monogenic disorders", Human mutation, 32: 144-151.

[57] Singleton, A.B. Hardy, J. Traynor, B.J. and Houlden, H., (2010). "Towards a complete resolution of the genetic architecture of disease", Trends in genetics, 26: 438-442.

[58] Kingsley, C.B., (2011). "Identification of causal sequence variants of disease in the next generation sequencing era", Disease Gene Identification: Methods and Protocols: 37-46.

[59] Meyerson, M. Gabriel, S. and Getz, G., (2010). "Advances in understanding cancer genomes through second-generation sequencing", Nature Reviews Genetics, 11: 685-696.

[60] Buermans, H.P. Ariyurek, Y. van Ommen, G. den Dunnen, J.T. and AC't Hoen, P., (2010). "New methods for next generation sequencing based microRNA expression profiling", BMC genomics, 11: 716.

[61] Morin, R.D. Zhao, Y. Prabhu, A.-L. Dhalla, N. McDonald, H. Pandoh, P. Tam, A. Zeng, T. Hirst, M. and Marra, M., (2010). "Preparation and analysis of microRNA libraries using the Illumina massively parallel sequencing technology", RNAi and microRNA-Mediated Gene Regulation in Stem Cells: Methods, Protocols, and Applications: 173-199.

[62] Costa, V. Gallo, M.A. Letizia, F. Aprile, M. Casamassimi, A. and Ciccodicola, A., (2010). "PPARG: gene expression regulation and next-generation sequencing for unsolved issues", PPAR research, 2010.

[63] Werner, T., (2010). "Next generation sequencing in functional genomics", Briefings in bioinformatics, 11: 499-511.

[64] Yang, J.-H. Shao, P. Zhou, H. Chen, Y.-Q. and Qu, L.-H., (2009). "deepBase: a database for deeply annotating and mining deep sequencing data", Nucleic Acids Research, 38: D123-D130.

[65] Hajirasouliha, I. Hormozdiari, F. Alkan, C. Kidd, J.M. Birol, I. Eichler, E.E. and Sahinalp, S.C., (2010). "Detection and characterization of novel sequence insertions using paired-end next-generation sequencing", Bioinformatics, 26: 1277-1283.

[66] Hestand, M.S. Klingenhoff, A. Scherf, M. Ariyurek, Y. Ramos, Y. van Workum, W. Suzuki, M. Werner, T. van Ommen, G.-J.B. and den Dunnen, J.T., (2010).

"Tissue-specific transcript annotation and expression profiling with complementary next-generation sequencing technologies", Nucleic Acids Research, 38: e165-e165.

[67] Gerstein, M.B. Bruce, C. Rozowsky, J.S. Zheng, D. Du, J. Korbel, J.O. Emanuelsson, O. Zhang, Z.D. Weissman, S. and Snyder, M., (2007). "What is a gene, post-ENCODE? History and updated definition", Genome research, 17: 669-681.

[68] BioNinja, DNA Structure, http://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/26-structure-of-dna-and-rna/dna-structure.html, 12 November 2017.

[69] BRIEF, A., "The Human Genome: Structure and Function of Genes and Chromosomes".

[70] Birney, E. Stamatoyannopoulos, J.A. Dutta, A. Guigó, R. Gingeras, T.R. Margulies, E.H. Weng, Z. Snyder, M. Dermitzakis, E.T. and Thurman, R.E., (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project", Nature, 447: 799-816.

[71] Ohshima, K., (2013). "RNA-mediated gene duplication and retroposons: retrogenes, LINEs, SINEs, and sequence specificity", International journal of evolutionary biology, 2013.

[72] Taft, R.J. Pheasant, M. and Mattick, J.S., (2007). "The relationship between non-protein-coding DNA and eukaryotic complexity", Bioessays, 29: 288-299.

[73] Harrow, J. Nagy, A. Reymond, A. Alioto, T. Patthy, L. Antonarakis, S.E. and Guigó, R., (2009). "Identifying protein-coding genes in genomic sequences", Genome biology, 10: 201.

[74] Dinger, M.E. Pang, K.C. Mercer, T.R. and Mattick, J.S., (2008). "Differentiating protein-coding and noncoding RNA: challenges and ambiguities", PLoS computational biology, 4: e1000176.

[75] Tomas Urban (2013), Animal Genetics, http://web2.mendelu.cz/af_291_projekty2/vseo/print.php?page=307&typ=html, 29 September 2017.

[76] Boivin, V. Deschamps-Francoeur, G. and Scott, M.S., (2017). "Protein coding genes as hosts for noncoding RNA expression": Elsevier.

[77] Kapranov, P. Cheng, J. Dike, S. Nix, D.A. Duttagupta, R. Willingham, A.T. Stadler, P.F. Hertel, J. Hackermüller, J. and Hofacker, I.L., (2007). "RNA maps reveal new RNA classes and a possible function for pervasive transcription", science, 316: 1484-1488.

[78] Hrdlickova, B. de Almeida, R.C. Borek, Z. and Withoff, S., (2014). "Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease", Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1842: 1910-1922.

[79] Fabian, M.R. Sonenberg, N. and Filipowicz, W., (2010). "Regulation of mRNA translation and stability by microRNAs", Annual review of biochemistry, 79: 351-379.

[80] miRBase: the microRNA database, http://www.mirbase.org/, 20 November 2017.

[81] Stefani, G. and Slack, F.J., (2008). "Small non-coding RNAs in animal development", Nature reviews Molecular cell biology, 9: 219-230.

[82] Mar-Aguilar, F. Rodríguez-Padilla, C. and Reséndez-Pérez, D., (2016). "Web-based tools for microRNAs involved in human cancer", Oncology letters, 11: 3563-3570.

[83] Calin, G.A. Dumitru, C.D. Shimizu, M. Bichi, R. Zupo, S. Noch, E. Aldler, H. Rattan, S. Keating, M. and Rai, K., (2002). "Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia", Proceedings of the National Academy of Sciences, 99: 15524-15529.

[84] Cohrt, K.O.H., Mysterious miRNAs: An Introduction to MicroRNAs, https://bitesizebio.com/23238/mysterious-mirnas-an-introduction-to-micrornas/, 10 Novermber 2017.

[85] Lonze, B.E. and Ginty, D.D., (2002). "Function and regulation of CREB family transcription factors in the nervous system", Neuron, 35: 605-623.

[86] Takeda, J. Yamasaki, C. Murakami, K. Sera, M. Hara, Y. Nagai, Y. Mamiya, K. Endo, T. Habara, T. and Obi, N., "H-Invitational Database, an integrated database of human genes and transcripts".

[87] Bhagwat, A.S. and Vakoc, C.R., (2015). "Targeting transcription factors in cancer", Trends in cancer, 1: 53-65.

[88] D'Onofrio, G. Ghosh, T.C. and Saccone, S., (2007). "Different functional classes of genes are characterized by different compositional properties", FEBS letters, 581: 5819-5824.

[89] Nguyen, T.T. Almon, R.R. DuBois, D.C. Sukumaran, S. Jusko, W.J. and Androulakis, I.P., (2014). "Tissue-specific gene expression and regulation in liver and muscle following chronic corticosteroid administration", Gene regulation and systems biology, 8: 75.

[90] Kawai, J. Shinagawa, A. Shibata, K. Yoshino, M. Itoh, M. Ishii, Y. Arakawa, T. Hara, A. Fukunishi, Y. and Konno, H., (2001). "Functional annotation of a full-length mouse cDNA collection", Nature, 409: 685-690.

[91] Carninci, P. Kasukawa, T. Katayama, S. Gough, J. Frith, M. Maeda, N. Oyama, R. Ravasi, T. Lenhard, B. and Wells, C., (2005). "The transcriptional landscape of the mammalian genome", science, 309: 1559-1563.

[92] El Amrani, K. Stachelscheid, H. Lekschas, F. Kurtz, A. and Andrade-Navarro, M.A., (2015). "MGFM: a novel tool for detection of tissue and cell specific marker genes from microarray gene expression data", BMC genomics, 16: 645.

[93] Yu, X. Lin, J. Zack, D.J. and Qian, J., (2006). "Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues", Nucleic Acids Research, 34: 4925-4936.

[94] Xiao, S.-J. Zhang, C. Zou, Q. and Ji, Z.-L., (2010). "TiSGeD: a database for tissue-specific genes", Bioinformatics, 26: 1273-1275.

[95] Dezső, Z. Nikolsky, Y. Sviridov, E. Shi, W. Serebriyskaya, T. Dosymbekov, D. Bugrim, A. Rakhmatulin, E. Brennan, R.J. and Guryanov, A., (2008). "A comprehensive functional analysis of tissue specificity of human gene expression", BMC biology, 6: 49.

[96] Petretto, E. Mangion, J. Dickens, N.J. Cook, S.A. Kumaran, M.K. Lu, H. Fischer, J. Maatz, H. Kren, V. and Pravenec, M., (2006). "Heritability and tissue specificity of expression quantitative trait loci", PLoS genetics, 2: e172.

[97] Nagaraj, S.H. Ingham, A. and Reverter, A., (2010). "The interplay between evolution, regulation and tissue specificity in the human hereditary diseasome", BMC genomics, 11: S23.

[98] Lage, K. Hansen, N.T. Karlberg, E.O. Eklund, A.C. Roque, F.S. Donahoe, P.K. Szallasi, Z. Jensen, T.S. and Brunak, S., (2008). "A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes", Proceedings of the National Academy of Sciences, 105: 20870-20875.

[99] Song, Y. Ahn, J. Suh, Y. Davis, M.E. and Lee, K., (2013). "Identification of novel tissue-specific genes by analysis of microarray databases: a human and mouse model", PloS one, 8: e64483.

[100] Greco, D. Somervuo, P. Di Lieto, A. Raitila, T. Nitsch, L. Castrén, E. and Auvinen, P., (2008). "Physiology, pathology and relatedness of human tissues from gene expression meta-analysis", PloS one, 3: e1880.

[101] Reverter, A. Ingham, A. and Dalrymple, B.P., (2008). "Mining tissue specificity, gene connectivity and disease association to reveal a set of genes that modify the action of disease causing genes", BioData mining, 1: 8.

[102] Yang, X. Ye, Y. Wang, G. Huang, H. Yu, D. and Liang, S., (2011). "VeryGene: linking tissue-specific genes to diseases, drugs, and beyond for knowledge discovery", Physiological genomics, 43: 457-460.

[103] Greene, C.S. Krishnan, A. Wong, A.K. Ricciotti, E. Zelaya, R.A. Himmelstein, D.S. Zhang, R. Hartmann, B.M. Zaslavsky, E. and Sealfon, S.C., (2015). "Understanding multicellular function and disease with human tissue-specific networks", Nature genetics, 47: 569-576.

[104] Yang, L. Wang, S. Zhou, M. Chen, X. Zuo, Y. Sun, D. and Lv, Y., (2016). "Comparative analysis of housekeeping and tissue-selective genes in human based on network topologies and biological properties", Molecular genetics and genomics, 291: 1227-1241.

[105] Tu, Z. Wang, L. Xu, M. Zhou, X. Chen, T. and Sun, F., (2006). "Further understanding human disease genes by comparing with housekeeping genes and other genes", BMC genomics, 7: 31.

[106] Mironov, V. Visconti, R.P. Kasyanov, V. Forgacs, G. Drake, C.J. and Markwald, R.R., (2009). "Organ printing: tissue spheroids as building blocks", Biomaterials, 30: 2164-2174.

[107] Golberg, A. Bruinsma, B.G. Uygun, B.E. and Yarmush, M.L., (2015). "Tissue heterogeneity in structure and conductivity contribute to cell survival during irreversible electroporation ablation by "electric field sinks"", Scientific reports, 5.

[108] What are the names of the tissue layers of the stomach?, https://socratic.org/questions/what-are-the-names-of-the-tissue-layers-of-the-stomach, 14 November 2017.

[109] Liu, X. Yu, X. Zack, D.J. Zhu, H. and Qian, J., (2008). "TiGER: a database for tissue-specific gene expression and regulation", BMC bioinformatics, 9: 271.

[110] Pan, J.-B. Hu, S.-C. Shi, D. Cai, M.-C. Li, Y.-B. Zou, Q. and Ji, Z.-L., (2013). "PaGenBase: a pattern gene database for the global and dynamic understanding of gene function", PloS one, 8: e80747.

[111] GENEVESTIGATOR, https://genevestigator.com/gv/, 19 September 2017.

[112] Schug, J. Schuller, W.-P. Kappen, C. Salbaum, J.M. Bucan, M. and Stoeckert, C.J., (2005). "Promoter features related to tissue specificity as measured by Shannon entropy", Genome biology, 6: R33.

[113] Martínez, O. and Reyes-Valdés, M.H., (2008). "Defining diversity, specialization, and gene specificity in transcriptomes through information theory", Proceedings of the National Academy of Sciences, 105: 9709-9714.

[114] Shannon, C.E., (2001). "A mathematical theory of communication", ACM SIGMOBILE Mobile Computing and Communications Review, 5: 3-55.

[115] Cheadle, C. Vawter, M.P. Freed, W.J. and Becker, K.G., (2003). "Analysis of microarray data using Z score transformation", The Journal of molecular diagnostics, 5: 73-81.

[116] Vandenbon, A. and Nakai, K., (2009). "Modeling tissue-specific structural patterns in human and mouse promoters", Nucleic Acids Research, 38: 17-25.

[117] Huminiecki, L. Lloyd, A.T. and Wolfe, K.H., (2003). "Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases", BMC genomics, 4: 31.

[118] Yanai, I. Benjamin, H. Shmoish, M. Chalifa-Caspi, V. Shklar, M. Ophir, R. Bar-Even, A. Horn-Saban, S. Safran, M. and Domany, E., (2004). "Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification", Bioinformatics, 21: 650-659.

[119] Su, A.I. Cooke, M.P. Ching, K.A. Hakak, Y. Walker, J.R. Wiltshire, T. Orth, A.P. Vega, R.G. Sapinoso, L.M. and Moqrich, A., (2002). "Large-scale analysis of the human and mouse transcriptomes", Proceedings of the National Academy of Sciences, 99: 4465-4470.

[120] Liao, B.-Y. and Zhang, J., (2006). "Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution", Molecular biology and evolution, 23: 1119-1128.

[121] Piasecka, B. Robinson-Rechavi, M. and Bergmann, S., (2012). "Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human", Bioinformatics, 28: 1865-1872.

[122] Kryuchkova-Mostacci, N. and Robinson-Rechavi, M., (2015). "Tissue-specific evolution of protein coding genes in human and mouse", PloS one, 10: e0131673.

[123] Smeds, L. Warmuth, V. Bolivar, P. Uebbing, S. Burri, R. Suh, A. Nater, A. Bureš, S. Garamszegi, L.Z. and Hogner, S., (2015). "Evolutionary analysis of the female-specific avian W chromosome", Nature communications, 6.

[124] Schuster, E.F. Blanc, E. Partridge, L. and Thornton, J.M., (2007). "Correcting for sequence biases in present/absent calls", Genome biology, 8: R125.

[125] Nelms, B.D. Waldron, L. Barrera, L.A. Weflen, A.W. Goettel, J.A. Guo, G. Montgomery, R.K. Neutra, M.R. Breault, D.T. and Snapper, S.B., (2016). "CellMapper: rapid and accurate inference of gene expression in difficult-to-isolate cell types", Genome biology, 17: 201.

[126] Zhong, Y. Wan, Y.-W. Pang, K. Chow, L.M. and Liu, Z., (2013). "Digital sorting of complex tissues for cell type-specific gene expression profiles", BMC bioinformatics, 14: 89.

[127] Kadota, K. Ye, J. Nakai, Y. Terada, T. and Shimizu, K., (2006). "ROKU: a novel method for identification of tissue-specific genes", BMC bioinformatics, 7: 294.

[128] Perter Boyle, B.L., Dünya kanser raporu 2008, http://kanser.gov.tr/Dosya/Kitaplar/dunyaraporu2008.pdf, 2 December 2017.

[129] T.C. Sağlık Bankanlığı, (2015). Türkiye kanser kontrol programı, http://kanser.gov.tr/Dosya/NCCP_2013-2018.pdf, 1 December 2017.

[130] What is Cancer?, http://www.csun.edu/~cmalone/pdf360/Ch22cancer.pdf, 28 October 2017.

[131] National Cancer Institute, https://www.cancer.gov/, 29 October 2017.

[132] Çobanoğlu, N. and Kiper, N., (2006). "Bina içi solunan havada tehlikeler", Çocuk sağlığı ve hastalıkları dergisi, 49: 71-75.

[133] Mayda, A.S. Tufan, N. and Baştaş, S., (2007). "Düzce Tıp Fakültesi öğrencilerinin sigara konusundaki tutumları ve içme sıklıkları", TSK Koruyucu Hekimlik Bülteni, 6: 364-370.

[134] Öztürk, B. Coşkun, U. Yaman, E. Kaya, A.O. Yildiz, R. Benekli, M. and Büyükberber, S., (2009). "Oral kavite kanserlerinde risk faktörleri, premalign lezyonlar ve kemoprevensiyon", UHOD, 19: 117-126.

[135] Canbay, E. Çelik, K. Kunt, T. Ertemur, M. ve Canbay, E., (2002). "Larinks kanserli hastalarda glutatyon peroksidaz aktivitesi ve lipit peroksidasyon düzeylerindeki değişiklikler", Cumhuriyet Üniversitesi Tıp Fakültesi Derg, 24: 175-178.

[136] Valko, M. Rhodes, C. Moncol, J. Izakovic, M. and Mazur, M., (2006). "Free radicals, metals and antioxidants in oxidative stress-induced cancer", Chemico-biological interactions, 160: 1-40.

[137] Antinuclear, Cancer incidence has risen in tandem with increased ionising radiation, https://antinuclear.net/2015/07/27/cancer-incidence-has-risen-in-tandem-with-increased-ionising-radiation/, 1 December 2017.

[138] World Health Organization, Cancers, The Problem, http://www.who.int/nmh/publications/fact_sheet_cancers_en.pdf, 3 December 2017.

[139] Köktürk, N., (2004). "Akciğer kanseri moleküler biyolojisi", Turkiye Klinikleri Journal of Pulmonary Medicine, 2: 177-182.

[140] Işıtmangil, T., (2008). "IASLC akciğer kanseri evrelendirme projesi: küçük hücreli dışı akciğer kanserinde TNM sınıflandırmasının yedinci düzenlemesi için öneriler", Türk Göğüs Kalp Damar Cerrahisi Dergisi, 16: 58-64.

[141] Travis, L.B. Gospodarowicz, M. Curtis, R.E. Aileen Clarke, E. Andersson, M. Glimelius, B. Joensuu, T. Lynch, C.F. van Leeuwen, F.E. and Holowaty, E., (2002). "Lung cancer following chemotherapy and radiotherapy for Hodgkin's disease", Journal of the National Cancer Institute, 94: 182-192.

[142] Group, I.A.L.C.T.C., (2004). "Cisplatin-based adjuvant chemotherapy in patients with completely resected non–small-cell lung cancer", N Engl J Med, 2004: 351-360.

[143] Delaney, G. Jacob, S. Featherstone, C. and Barton, M., (2005). "The role of radiotherapy in cancer treatment", Cancer, 104: 1129-1137.

[144] Du, G. Liu, Y. Li, J. Liu, W. Wang, Y. and Li, H., (2013). "Hypothermic microenvironment plays a key role in tumor immune subversion", International immunopharmacology, 17: 245-253.

[145] Group, E.B.C.T.C., (2005). "Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials", The Lancet, 366: 2087-2106.

[146] Kürkçü, E., (2008). "Deneysel olarak oluşturulmuş meme tümörlerinde curcumin'in arginaz enzim aktivitesi, ornitin ve nitrik oksit düzeylerine etkisi".

[147] Siegel, R. Ma, J. Zou, Z. and Jemal, A., (2014). "Cancer statistics, 2014", CA: a cancer journal for clinicians, 64: 9-29.

[148] Herrero, R., (1996). "Epidemiology of cervical cancer", Journal of the National Cancer Institute. Monographs: 1-6.

[149] Novikova, T., (2017). "Optical techniques for cervical neoplasia detection", Beilstein Journal of Nanotechnology, 8: 1844.

[150] Enzinger, P.C. and Mayer, R.J., (2003). "Esophageal cancer", New England Journal of Medicine, 349: 2241-2252.

[151] Tahara, M. Ohtsu, A. Hironaka, S. Boku, N. Ishikura, S. Miyata, Y. Ogino, T. and Yoshida, S., (2005). "Clinical impact of criteria for complete response (CR) of primary site to treatment of esophageal cancer", Japanese journal of clinical oncology, 35: 316-323.

[152] Coppin, C. Porzsolt, F. Awa, A. Kumpf, J. Coldman, A. and Wilt, T., (2004). "Immunotherapy for advanced renal cell cancer", Cochrane Database Syst Rev, 1.

[153] Sánchez-Gastaldo, A. Kempf, E. del Alba, A.G. and Duran, I., (2017). "Systemic treatment of renal cell cancer: a comprehensive review", Cancer Treatment Reviews, 60: 77-89.

[154] Huo, X. Han, S. Wu, G. Latchoumanin, O. Zhou, G. Hebbard, L. George, J. and Qiao, L., (2017). "Dysregulated long noncoding RNAs (lncRNAs) in hepatocellular carcinoma: implications for tumorigenesis, disease progression, and liver cancer stem cells", Molecular cancer, 16: 165.

[155] Yang, H. Zhang, X. Cai, X.-y. Wen, D.-y. Ye, Z.-h. Liang, L. Zhang, L. Wang, H.-l. Chen, G. and Feng, Z.-b., (2017). "From big data to diagnosis and prognosis: gene expression signatures in liver hepatocellular carcinoma", PeerJ, 5: e3089.

[156] Huang, J.-Y. Jian, Z.-H. Nfor, O.N. Ku, W.-Y. Ko, P.-C. Lung, C.-C. Ho, C.-C. Pan, H.-H. Huang, C.-Y. and Liang, Y.-C., (2015). "The effects of pulmonary diseases on histologic types of lung cancer in both sexes: a population-based study in Taiwan", BMC cancer, 15: 834.

[157] Xu, H. Ma, J. Wu, J. Chen, L. Sun, F. Qu, C. Zheng, D. and Xu, S., (2016). "Gene expression profiling analysis of lung adenocarcinoma", Brazilian Journal of Medical and Biological Research, 49.

[158] Pancreatic cancer incidence statistics, http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/pancreatic-cancer/incidence#heading-Two, 11 December 2017.

[159] Conroy, T. Desseigne, F. Ychou, M. Bouché, O. Guimbaud, R. Bécouarn, Y. Adenis, A. Raoul, J.-L. Gourgou-Bourgade, S. and de la Fouchardière, C., (2011). "FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer", New England Journal of Medicine, 364: 1817-1825.

[160] Altıntaş, E.K. Atalay, F.Ö. Aytaç, B. Vuruşkan, H. ve Nermin, Ü., "Prostatın Kansere Yolculuğunda Atipik Küçük Asiner Proliferasyonların Önemi Nedir?".

[161] Xu, J. Stolk, J.A. Zhang, X. Silva, S.J. Houghton, R.L. Matsumura, M. Vedvick, T.S. Leslie, K.B. Badaro, R. and Reed, S.G., (2000). "Identification of differentially expressed genes in human prostate cancer using subtraction and microarray", Cancer research, 60: 1677-1682.

[162] Myers, J.S. von Lersner, A.K. Robbins, C.J. and Sang, Q.-X.A., (2015). "Differentially expressed genes and signature pathways of human prostate cancer", PloS one, 10: e0145322.

[163] Radice, E. and Dozois, R., (2001). "Locally recurrent rectal cancer", Digestive surgery, 18: 355-362.

[164] Pasechnikov, V. Chukov, S. Fedorov, E. Kikuste, I. and Leja, M., (2014). "Gastric cancer: prevention, screening and early diagnosis", World journal of gastroenterology: WJG, 20: 13842.

[165] Davies, L. and Welch, H.G., (2006). "Increasing incidence of thyroid cancer in the United States, 1973-2002", Jama, 295: 2164-2167.

[166] Cooper, D.S. Doherty, G.M. Haugen, B.R. Kloos, R.T. Lee, S.L. Mandel, S.J. Mazzaferri, E.L. McIver, B. Pacini, F. and Schlumberger, M., (2009). "Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association (ATA) guidelines taskforce on thyroid nodules and differentiated thyroid cancer", Thyroid, 19: 1167-1214.

[167] Blaveri, E. Simko, J.P. Korkola, J.E. Brewer, J.L. Baehner, F. Mehta, K. DeVries, S. Koppie, T. Pejavar, S. and Carroll, P., (2005). "Bladder cancer outcome and subtype classification by gene expression", Clinical Cancer Research, 11: 4044-4055.

[168] Ploeg, M. Aben, K.K. and Kiemeney, L.A., (2009). "The present and future burden of urinary bladder cancer in the world", World journal of urology, 27: 289-293.

[169] DeSantis, C.E. Lin, C.C. Mariotto, A.B. Siegel, R.L. Stein, K.D. Kramer, J.L. Alteri, R. Robbins, A.S. and Jemal, A., (2014). "Cancer treatment and survivorship statistics, 2014", CA: a cancer journal for clinicians, 64: 252-271.

[170] Mauro, M.J. and Druker, B.J., (2001). "STI571: a gene product-targeted therapy for leukemia", Current oncology reports, 3: 223-227.

[171] Trajkovski, I. Zelezny, F. Lavrac, N. and Tolar, J., (2008). "Learning relational descriptions of differentially expressed gene groups", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38: 16-25.

[172] Yang, M. Zhang, Y. Wu, X. and Chen, W., (2016). "Critical genes of hepatocellular carcinoma revealed by network and module analysis of RNA-seq data", European review for medical and pharmacological sciences, 20: 4248-4256.

[173] Sheng, Z. Sun, Y. Zhu, R. Jiao, N. Tang, K. Cao, Z. and Ma, C., (2015). "Functional cross-talking between differentially expressed and alternatively spliced genes in human liver cancer cells treated with berberine", PloS one, 10: e0143742.

[174] Merlo, L.M. Pepper, J.W. Reid, B.J. and Maley, C.C., (2006). "Cancer as an evolutionary and ecological process", Nature Reviews Cancer, 6: 924-935.

[175] Olivier, M. Petitjean, A. Marcel, V. Petre, A. Mounawar, M. Plymoth, A. De Fromentel, C. and Hainaut, P., (2009). "Recent advances in p53 research: an interdisciplinary perspective", Cancer gene therapy, 16: 1-12.

[176] Vogelstein, B. Papadopoulos, N. Velculescu, V.E. Zhou, S. Diaz, L.A. and Kinzler, K.W., (2013). "Cancer genome landscapes", science, 339: 1546-1558.

[177] Wu, T.-J. Schriml, L.M. Chen, Q.-R. Colbert, M. Crichton, D.J. Finney, R. Hu, Y. Kibbe, W.A. Kincaid, H. and Meerzaman, D., (2015). "Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis", Database, 2015.

[178] Markowetz, F., Cancer heterogeneity and evolution – the review to end all reviews,https://scientificbsides.wordpress.com/2014/11/24/cancer-heterogeneity-and-evolution-the-review-to-end-all-reviews/,5 September 2017.

[179] Fisher, R. Pusztai, L. and Swanton, C., (2013). "Cancer heterogeneity: implications for targeted therapeutics", British journal of cancer, 108: 479-485.

[180] Gerlinger, M. Rowan, A.J. Horswell, S. Larkin, J. Endesfelder, D. Gronroos, E. Martinez, P. Matthews, N. Stewart, A. and Tarpey, P., (2012). "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing", New England Journal of Medicine, 366: 883-892.

[181] Sottoriva, A. Spiteri, I. Shibata, D. Curtis, C. and Tavaré, S., (2013). "Single-molecule genomic data delineate patient-specific tumor profiles and cancer stem cell organization", Cancer research, 73: 41-49.

[182] Magee, J.A. Piskounova, E. and Morrison, S.J., (2012). "Cancer stem cells: impact, heterogeneity, and uncertainty", Cancer cell, 21: 283-296.

[183] Johann Jr, D.J. Rodriguez-Canales, J. Mukherjee, S. Prieto, D.A. Hanson, J.C. Emmert-Buck, M. and Blonder, J., (2009). "Approaching solid tumor heterogeneity on a cellular basis by tissue proteomics using laser capture microdissection and biological mass spectrometry", Journal of proteome research, 8: 2310-2318.

[184] Bonavia, R. Mukasa, A. Narita, Y. Sah, D.W. Vandenberg, S. Brennan, C. Johns, T.G. Bachoo, R. Hadwiger, P. and Tan, P., (2010). "Tumor heterogeneity is an active process maintained by a mutant EGFR-induced cytokine circuit in glioblastoma", Genes & development, 24: 1731-1745.

[185] Metastatic Cancer, National Cancer Enstitute, https://www.cancer.gov/types/metastatic-cancer, 18 November 2017.

[186] Valastyan, S. and Weinberg, R.A., (2011). "Tumor metastasis: molecular insights and evolving paradigms", Cell, 147: 275-292.

[187] Albritton, J.L. and Miller, J.S., (2017). "3D bioprinting: improving in vitro models of metastasis with heterogeneous tumor microenvironments", Disease models & mechanisms, 10: 3-14.

[188] Bos, P.D. Zhang, X.H.-F. Nadal, C. Shu, W. Gomis, R.R. Nguyen, D.X. Minn, A.J. van de Vijver, M.J. Gerald, W.L. and Foekens, J.A., (2009). "Genes that mediate breast cancer metastasis to the brain", Nature, 459: 1005-1009.

[189] Tabaries, S. Dong, Z. Annis, M. Omeroglu, A. Pepin, F. Ouellet, V. Russo, C. Hassanain, M. Metrakos, P. and Diaz, Z., (2011). "Claudin-2 is selectively enriched in and promotes the formation of breast cancer liver metastases through engagement of integrin complexes", Oncogene, 30: 1318-1328.

[190] Kawabata, R. Kimura, Y. Kawase, T. Kamigaki, S. Yamamura, J. Nakamura, Y. Munakata, S. Fukunaga, M. and Ohzato, H., (2014). "Long-term survival of a patient with esophageal metastasis from breast cancer treated with esophagectomy", Gan to kagaku ryoho. Cancer & chemotherapy, 41: 2024-2026.

[191] El, F.M.r. Kanab, R. Ameuraoui, T. Sidibe, F. Azegrare, M. Arifi, S. El, M.M. Amarti, A. and Mellas, N., (2017). "Parotid metastasis from carcinoma of the breast: report of a case and review of the literature", The Pan African medical journal, 27: 79-79.

[192] Li, H. Wu, X. and Cheng, X., (2016). "Advances in diagnosis and treatment of metastatic cervical cancer", Journal of gynecologic oncology, 27.

[193] Muro, K. Hamaguchi, T. Ohtsu, A. Boku, N. Chin, K. Hyodo, I. Fujita, H. Takiyama, W. and Ohtsu, T., (2004). "A phase II study of single-agent docetaxel in patients with metastatic esophageal cancer", Annals of Oncology, 15: 955-959.

[194] Shaheen, O. Ghibour, A. and Alsaid, B., (2017). "Esophageal Cancer Metastases to Unexpected Sites: A Systematic Review", Gastroenterology Research and Practice, 2017.

[195] Chino, O. Makuuchi, H. Ozawa, S. Shimada, H. Nishi, T. Yamamoto, S. Miyako, H. Ito, E. Kise, Y. and Hara, T., (2015). "Small intestinal metastasis from esophageal squamous cell carcinoma presenting with perforated peritonitis", Tokai J Exp Clin Med, 40: 63-68.

[196] Griffin, N. Gore, M.E. and Sohaib, S.A., (2007). "Imaging in metastatic renal cell carcinoma", American Journal of Roentgenology, 189: 360-370.

[197] Schlesinger-Raab, A. Treiber, U. Zaak, D. Hölzel, D. and Engel, J., (2008). "Metastatic renal cell carcinoma: results of a population-based study with 25 years follow-up", European Journal of Cancer, 44: 2485-2495.

[198] Bianchi, M. Sun, M. Jeldres, C. Shariat, S. Trinh, Q.-D. Briganti, A. Tian, Z. Schmitges, J. Graefen, M. and Perrotte, P., (2011). "Distribution of metastatic sites in renal cell carcinoma: a population-based analysis", Annals of Oncology, 23: 973-980.

[199] Alghamdi, A. and Tam, J., (2006). "Cardiac metastasis from a renal cell carcinoma", Canadian Journal of Cardiology, 22: 1231-1232.

[200] Ralli, M. Altissimi, G. Turchetta, R. and Rigante, M., (2017). "Metastatic renal cell carcinoma presenting as a paranasal sinus mass: the importance of differential diagnosis", Case reports in otolaryngology, 2017.

[201] Rouvinov, K. Neulander, E.Z. Kan, E. Asali, M. Ariad, S. and Mermershtain, W., (2017). "Testicular Metastasis from Renal Cell Carcinoma: A Case Report and Review of the Literature", Case reports in oncology, 10: 388-391.

[202] Lynne Eldridge, G.H., Where Does Lung Cancer Spread? Common Sites of Lung Cancer Metastases, https://www.verywell.com/where-does-lung-cancer-spread-2249368, 24 Nıovember 2017.

[203] Shin, D.-Y. Kim, C.H. Park, S. Baek, H. and Yang, S.H., (2014). "EGFR mutation and brain metastasis in pulmonary adenocarcinomas", Journal of Thoracic Oncology, 9: 195-199.

[204] Popper, H.H., (2016). "Progression and metastasis of lung cancer", Cancer and Metastasis Reviews, 35: 75-91.

[205] Kimakura, M. Abe, T. Nagahara, A. Fujita, K. Kiuchi, H. Uemura, M. and Nonomura, N., (2016). "Metastatic testicular cancer presenting with liver and kidney dysfunction treated with modified BEP chemotherapy combined with continuous hemodiafiltration and rasburicase", Anti-cancer drugs, 27: 364.

[206] Kim, S. Chun, M. Wang, H. Cho, S. Oh, Y.-T. Kang, S.-H. and Yang, J., (2007). "Bone metastasis from primary hepatocellular carcinoma: characteristics of soft tissue formation", Cancer research and treatment: official journal of Korean Cancer Association, 39: 104.

[207] Gao, J. Xie, L. Yang, W.-S. Zhang, W. Gao, S. Wang, J. and Xiang, Y.-B., (2012). "Risk factors of hepatocellular carcinoma--current status and perspectives", Asian Pac J Cancer Prev, 13: 743-752.

[208] Castelli, G. Pelosi, E. and Testa, U., (2017). "Liver Cancer: Molecular Characterization, Clonal Evolution and Cancer Stem Cells", Cancers, 9: 127.

[209] Su, S.-C. Ho, Y.-C. Liu, Y.-F. Reiter, R.J. Chou, C.-H. Yeh, C.-M. Lee, H.-L. Chung, W.-H. Hsieh, M.-J. and Yang, S.-F., (2017). "Association of melatonin membrane receptor 1A/1B gene polymorphisms with the occurrence and metastasis of hepatocellular carcinoma", Oncotarget, 8: 85655.

[210] Embuscado, E.E. Laheru, D. Ricci, F. Yun, K.J. de Boom Witzel, S. Seigel, A. Flickinger, K. Hidalgo, M. Bova, G.S. and Iacobuzio-Donahue, C.A., (2005). "Immortalizing the complexity of cancer metastasis: genetic features of lethal metastatic pancreatic cancer obtained from rapid autopsy", Cancer biology & therapy, 4: 548-554.

[211] Deeb, A. Haque, S.-U. and Olowokure, O., (2015). "Pulmonary metastases in pancreatic cancer, is there a survival influence?", Journal of gastrointestinal oncology, 6: E48.

[212] Boratyn-Nowicka, A. Blicharz-Dorniak, J. Wachuła, E. Kunikowska, J. and Kos-Kudła, B., (2014). "An atypical course of pancreatic neuroendocrine tumour manifesting as cardiac metastasis—a clinical case", Endokrynologia Polska, 65: 232-239.

[213] Rosati, L.M. Kummerlowe, M.N. Poling, J. Hacker-Prietz, A. Narang, A.K. Shin, E.J. Le, D.T. Fishman, E.K. Hruban, R.H. and Yang, S.C., (2017). "A rare case of esophageal metastasis from pancreatic ductal adenocarcinoma: a case report and literature review", Oncotarget, 8: 100942.

[214] Castle, E.P., Prostate cancer metastasis: Where does prostate cancer spread?, https://www.mayoclinic.org/diseases-conditions/prostate-cancer/expert-answers/prostate-cancer-metastasis/faq-20058270, 20 December 2017.

[215] Campara, Z. Simic, D. Aleksic, P. Spasic, A. and Milicevic, S., (2016). "Metastasis of Prostate Adenocarcinoma to the Testis", Medical Archives, 70: 318.

[216] Vatandoust, S. Price, T.J. and Karapetis, C.S., (2015). "Colorectal cancer: Metastases to a single organ", World journal of gastroenterology, 21: 11767.

[217] Riihimäki, M. Hemminki, A. Sundquist, J. and Hemminki, K., (2016). "Patterns of metastasis in colon and rectal cancer", Scientific reports, 6.

[218] Verma, N. Babu, S. Kushwaha, J.K. and Singhai, A., (2013). "Testicular metastasis of colorectal carcinoma: an unusual presentation", BMJ case reports, 2013: bcr2012007849.

[219] Tokuyama, S. Yoshioka, S. Fukunaga, M. Honda, S. Yukimoto, R. Okamoto, A. Saito, A. Konishi, K. Okada, K. and Ota, H., (2016). "A Case of Long-Term Survival of Resected Pancreatic Metastasis from Colon Cancer", Gan to kagaku ryoho. Cancer & chemotherapy, 43: 2459-2461.

[220] Riihimäki, M. Hemminki, A. Sundquist, K. Sundquist, J. and Hemminki, K., (2016). "Metastatic spread in patients with gastric cancer", Oncotarget, 7: 52307.

[221] Wexler, J.A., (2011). "Approach to the thyroid cancer patient with bone metastases", The Journal of Clinical Endocrinology & Metabolism, 96: 2296-2307.

[222] Appetecchia, M. Barnabei, A. Pompeo, V. Sentinelli, S. Baldelli, R. Corsello, S.M. and Torino, F., (2014). "Testicular and inguinal lymph node metastases of medullary thyroid cancer: a case report and review of the literature", BMC endocrine disorders, 14: 84.

[223] Davidson, M. Olsen, R.J. Ewton, A.A. and Robbins, R.J., (2017). "Pancreas metastases from papillary thyroid carcinoma: a review of the literature", Endocrine Practice.

[224] Kuwatani, M. Kawakami, H. Asaka, M. Marukawa, K. Matsuno, Y. and Hosaka, M., (2008). "Pancreatic metastasis from small cell carcinoma of the uterine cervix demonstrated by endoscopic ultrasonography-guided fine needle aspiration", Diagnostic cytopathology, 36: 840-842.

[225] Kurra, V. Krajewski, K.M. Jagannathan, J. Giardino, A. Berlin, S. and Ramaiya, N., (2013). "Typical and atypical metastatic sites of recurrent endometrial carcinoma", Cancer Imaging, 13: 113.

[226] Kapoor, K. Evans, M.C. Shkullaku, M. Schillinger, R. White, C.S. and Roque, D.M., (2016). "Biventricular metastatic invasion from cervical squamous cell carcinoma", BMJ case reports, 2016: bcr2016214931.

[227] Shinagare, A.B. Ramaiya, N.H. Jagannathan, J.P. Fennessy, F.M. Taplin, M.-E. and Van den Abbeele, A.D., (2011). "Metastatic pattern of bladder cancer: correlation with the characteristics of the primary tumor", American Journal of Roentgenology, 196: 117-122.

[228] Nakamura, E. Shimizu, M. Itoh, T. and Manabe, T., (2001). "Secondary tumors of the pancreas: clinicopathological study of 103 autopsy cases of Japanese patients", Pathology international, 51: 686-690.

[229] Neelakantan, D. Drasin, D.J. and Ford, H.L., (2015). "Intratumoral heterogeneity: clonal cooperation in epithelial-to-mesenchymal transition and metastasis", Cell adhesion & migration, 9: 265-276.

[230] Hanna Jr, M.G. Howard, J. and Vermorken, J., (2014). "Active specific immunotherapy: Using tumor heterogeneity to successfully fight cancer", Human vaccines & immunotherapeutics, 10: 3286-3296.

[231] Marino, N. Woditschka, S. Reed, L.T. Nakayama, J. Mayer, M. Wetzel, M. and Steeg, P.S., (2013). "Breast cancer metastasis: issues for the personalization of its prevention and treatment", The American journal of pathology, 183: 1084-1095.

[232] The R Project for Statistical Computing, https://www.r-project.org/, 10 December 2017.

[233] R Studio, https://www.rstudio.com/conference/, 2 October 2017.

[234] ArrayExpress, https://www.ebi.ac.uk/arrayexpress/, 8 December 2017.

[235] ExpressionAtlas, https://www.ebi.ac.uk/gxa/home, 11 December 2017.

[236] Habuka, M. Fagerberg, L. Hallström, B.M. Pontén, F. Yamamoto, T. and Uhlen, M., (2015). "The urinary bladder transcriptome and proteome defined by transcriptomics and antibody-based profiling", PloS one, 10: e0145301.

[237] Uhlén, M. Fagerberg, L. Hallström, B.M. Lindskog, C. Oksvold, P. Mardinoglu, A. Sivertsson, Å. Kampf, C. Sjöstedt, E. and Asplund, A., (2015). "Tissue-based map of the human proteome", science, 347: 1260419.

[238] Djureinovic, D. Hallström, B.M. Horie, M. Mattsson, J.S.M. La Fleur, L. Fagerberg, L. Brunnström, H. Lindskog, C. Madjar, K. and Rahnenführer, J., (2016). "Profiling cancer testis antigens in non–small-cell lung cancer", JCI insight, 1.

[239] Danielsson, A. Pontén, F. Fagerberg, L. Hallström, B.M. Schwenk, J.M. Uhlén, M. Korsgren, O. and Lindskog, C., (2014). "The human pancreas proteome defined by transcriptomics and antibody-based profiling", PloS one, 9: e115421.

[240] Fagerberg, L. Hallström, B.M. Oksvold, P. Kampf, C. Djureinovic, D. Odeberg, J. Habuka, M. Tahmasebpoor, S. Danielsson, A. and Edlund, K., (2014). "Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics", Molecular & Cellular Proteomics, 13: 397-406.

[241] Edqvist, P.-H.D. Fagerberg, L. Hallström, B.M. Danielsson, A. Edlund, K. Uhlén, M. and Pontén, F., (2015). "Expression of human skin-specific genes defined by transcriptomics and antibody-based profiling", Journal of Histochemistry & Cytochemistry, 63: 129-141.

[242] Sjöstedt, E. Fagerberg, L. Hallström, B.M. Häggmark, A. Mitsios, N. Nilsson, P. Pontén, F. Hökfelt, T. Uhlén, M. and Mulder, J., (2015). "Defining the human

brain proteome using transcriptomics and antibody-based profiling with a focus on the cerebral cortex", PloS one, 10: e0130028.

[243] Flegel, C. Vogel, F. Hofreuter, A. Wojcik, S. Schoeder, C. Kieć-Kononowicz, K. Brockmeyer, N.H. Müller, C.E. Becker, C. and Altmüller, J., (2016). "Characterization of non-olfactory GPCRs in human sperm with a focus on GPR18", Scientific reports, 6: 32255.

[244] Consortium, G., (2015). "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans", science, 348: 648-660.

[245] Lin, S. Lin, Y. Nery, J.R. Urich, M.A. Breschi, A. Davis, C.A. Dobin, A. Zaleski, C. Beer, M.A. and Chapman, W.C., (2014). "Comparison of the transcriptional landscapes between human and mouse tissues", Proceedings of the National Academy of Sciences, 111: 17224-17229.

[246] Brenda Tissue Ontology, https://www.ebi.ac.uk/ols/ontologies/bto, 22 September 2017.

[247] BioMart, Ensembl, http://www.ensembl.org/biomart/martview/6321b2eb34c3449b7b6f8b2db0ff83dc, 10 September 2017.

[248] Conesa, A. Madrigal, P. Tarazona, S. Gomez-Cabrero, D. Cervera, A. McPherson, A. Szcześniak, M.W. Gaffney, D.J. Elo, L.L. and Zhang, X., (2016). "A survey of best practices for RNA-seq data analysis", Genome biology, 17: 13.

[249] Li, B. Ruotti, V. Stewart, R.M. Thomson, J.A. and Dewey, C.N., (2009). "RNA-Seq gene expression estimation with read mapping uncertainty", Bioinformatics, 26: 493-500.

[250] Rahman, M. Jackson, L.K. Johnson, W.E. Li, D.Y. Bild, A.H. and Piccolo, S.R., (2015). "Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results", Bioinformatics, 31: 3666-3672.

[251] Wagner, G.P. Kin, K. and Lynch, V.J., (2012). "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples", Theory in biosciences, 131: 281-285.

[252] What the FPKM? A review of RNA-Seq expression units, https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/, 5 December 2017.

[253] Faraway, J.J., (2016). Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models: CRC press.

[254] Gibbons, M.R. Ross, S.A. and Shanken, J., (1989). "A test of the efficiency of a given portfolio", Econometrica: Journal of the Econometric Society: 1121-1152.

[255] Zhu, J. Chen, G. Zhu, S. Li, S. Wen, Z. Li, B. Zheng, Y. and Shi, L., (2016). "Identification of tissue-specific protein-coding and noncoding transcripts across 14 human tissues using RNA-sEq", Scientific reports, 6: 28400.

[256] Rumsey, D.J., (2007). Intermediate statistics for dummies: John Wiley & Sons.

[257] Rumsey, D.J. and Unger, D., (2015). U Can: statistics for dummies: John Wiley & Sons.

[258] Strelen, J.C., (2004). Min-max confidence intervals, http://web.informatik.uni-bonn.de/IV/strelen/Lehre/Veranstaltungen/sim/Folien/47MMCIs.pdf, 28 December 2017.

[259] Rumsey, D.J., How to Calculate A Confidence Interval For A Population Mean When You Know Its Standard Deviation, http://www.dummies.com/education/math/statistics/how-to-calculate-a-confidence-interval-for-a-population-mean-when-you-know-its-standard-deviation/, 16 January 2018.

[260] Morris, G.D.S.a.P.E., (2015). "Method: Building confidence in confidence intervals", Methods, 28: 476-479.

[261] Hahn, G.J. and Meeker, W.Q., (2011). Statistical intervals: a guide for practitioners: John Wiley & Sons.

[262] Markatou, M. Chen, Y. Afendras, G. and Lindsay, B.G., (2016). "Statistical Distances and Their Role in Robustness", arXiv preprint arXiv:1612.07408.

[263] Bezdek, J.C., (2013). Pattern recognition with fuzzy objective function algorithms: Springer Science & Business Media.

[264] Tari, L. Baral, C. and Kim, S., (2009). "Fuzzy c-means clustering with prior biological knowledge", Journal of biomedical informatics, 42: 74-81.

[265] Han, J. Pei, J. and Kamber, M., (2011). Data mining: concepts and techniques: Elsevier.

[266] Data Clustering Algorithms,Fuzzy c-means clustering algorithm, https://sites.google.com/site/dataclusteringalgorithms/fuzzy-c-means-clustering-algorithm, 5 October 2017.

[267] Chuang, K.-S. Tzeng, H.-L. Chen, S. Wu, J. and Chen, T.-J., (2006). "Fuzzy c-means clustering with spatial information for image segmentation", computerized medical imaging and graphics, 30: 9-15.

[268] Wold, H., Estimation of principal components and related models by iterative least squares. Multivariate Analysis. Edited by: Krishnaiah PR. 1966, New York: Academic Press.

[269] Testing For Normality, http://webspace.ship.edu/pgmarr/Geo441/Lectures/Lec%205%20-%20Normality%20Testing.pdf, 4 September 2017.

[270] Baxter, M., (2002). "Testing for Normality", Biometrics, 58: 1044.

[271] Chakravarti, I.M. and Laha, R.G., (1967). Handbook of methods of applied statistics, ed. Handbook of methods of applied statistics. John Wiley & Sons.

[272] Evans, D.L. Drew, J.H. and Leemis, L.M., (2008). "The distribution of the Kolmogorov–Smirnov, Cramer–von Mises, and Anderson–Darling test statistics for exponential populations with estimated parameters", Communications in Statistics—Simulation and Computation®, 37: 1396-1421.

[273] Engineering Statistics Handbook, http://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm, 16 January 2018.

[274] The Human Protein Atlas, https://www.proteinatlas.org/tissue, 8 November, 2017.

[275] School of Medicine and Health Sciences, Bioexpress, https://hive.biochemistry.gwu.edu/cgi-bin/prd/bioxpress/servlet.cgi, 27 September 2017.

[276] Pan, Q. Long, X. Song, L. Zhao, D. Li, X. Li, D. Li, M. Zhou, J. Tang, X. and Ren, H., (2016). "Transcriptome sequencing identified hub genes for hepatocellular carcinoma by weighted-gene co-expression analysis", Oncotarget, 7: 38487.

[277] Corley, S.M. MacKenzie, K.L. Beverdam, A. Roddam, L.F. and Wilkins, M.R., (2017). "Differentially expressed genes from RNA-Seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols", BMC genomics, 18: 399.

[278] The Database for Annotation, Visualization and Integrated Discovery (DAVID), https://david.ncifcrf.gov/, 9 November 2017.

[279] Benjamini, Y. and Hochberg, Y., (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing", Journal of the royal statistical society. Series B (Methodological): 289-300.

[280] NIH,The Cancer Genome Atlas Data Portal, https://tcga-data.nci.nih.gov/docs/publications/tcga/index.html, 25 November 2017.

[281] GeneCards, Human Gene Database, http://www.genecards.org/, 12 November 2017.

[282] WebGestalt (WEB-based Gene SeT AnaLysis Toolket), http://www.webgestalt.org/option.php, Accessed 24 November 2017.

[283] NetworkAnalyst, Network-based visual analytics for gene expression profiling, meta-analysis and interpretation, http://www.networkanalyst.ca/, 8 November 2017.

[284] mir2Disease Base, http://www.mir2disease.org/,10 September 2017 2017.

[285] Cytoscape, Network Data Integration, Analysis, and Visualization in a Box, http://www.cytoscape.org/index.html, 5 August 2017.

[286] Kryuchkova-Mostacci, N. and Robinson-Rechavi, M., (2016). "Tissue-specificity of gene expression diverges slowly between orthologs, and rapidly between paralogs", PLoS computational biology, 12: e1005274.

[287] Su, A.I. Wiltshire, T. Batalov, S. Lapp, H. Ching, K.A. Block, D. Zhang, J. Soden, R. Hayakawa, M. and Kreiman, G., (2004). "A gene atlas of the mouse and human protein-encoding transcriptomes", Proceedings of the National Academy of Sciences of the United States of America, 101: 6062-6067.

[288] Liang, S. Li, Y. Be, X. Howes, S. and Liu, W., (2006). "Detecting and profiling tissue-selective genes", Physiological genomics, 26: 158-162.

[289] Assis, R. and Bachtrog, D., (2013). "Neofunctionalization of young duplicate genes in Drosophila", Proceedings of the National Academy of Sciences, 110: 17409-17414.

[290] Bush, S.J. Kover, P.X. and Urrutia, A.O., (2015). "Lineage-specific sequence evolution and exon edge conservation partially explain the relationship between evolutionary rate and expression level in A. thaliana", Molecular ecology, 24: 3093-3106.

[291] The Human Protein Atlas,https://www.proteinatlas.org/, 11 December 2017.

[292] Enard, W. Khaitovich, P. Klose, J. Zöllner, S. Heissig, F. Giavalisco, P. Nieselt-Struwe, K. Muchmore, E. Varki, A. and Ravid, R., (2002). "Intra-and interspecific variation in primate gene expression patterns", science, 296: 340-343.

[293] Khaitovich, P. Muetzel, B. She, X. Lachmann, M. Hellmann, I. Dietzsch, J. Steigele, S. Do, H.-H. Weiss, G. and Enard, W., (2004). "Regional patterns of gene expression in human and chimpanzee brains", Genome research, 14: 1462-1473.

[294] Myers, A.J. Gibbs, J.R. Webster, J.A. Rohrer, K. Zhao, A. Marlowe, L. Kaleem, M. Leung, D. Bryden, L. and Nath, P., (2007). "A survey of genetic human cortical gene expression", Nature genetics, 39: 1494-1499.

[295] Yamashita, A. Goto, N. Nishiguchi, S. Shimada, K. Yamanishi, H. and Yasunaga, T., "Detection of tissue-specific genes and computational analysis of testis-specific gene expression regulatory regions".

[296] Baker, D.A. and Russell, S., (2011). "Role of testis-specific gene expression in sex-chromosome evolution of Anopheles gambiae", Genetics, 189: 1117-1120.

[297] Yamashita, A. Goto, N. Nishiguchi, S. Shimada, K. Yamanishi, H. ve Yasunaga, T., (2008). "Computational search for over-represented 8-mers within the 5'-regulatory regions of 634 mouse testis-specific genes", Gene, 427: 93-98.

[298] Naumova, O.Y. Lee, M. Rychkov, S.Y. Vlasova, N.V. and Grigorenko, E.L., (2013). "Gene expression in the human brain: the current state of the study of specificity and spatiotemporal dynamics", Child development, 84: 76-88.

[299] Nikoozad, Z. Ghorbanian, M.T. and Rezaei, A., (2014). "Comparison of the liver function and hepatic specific genes expression in cultured mesenchymal stem cells and hepatocytes", Iranian journal of basic medical sciences, 17: 27.

[300] Ko, Y. Ament, S.A. Eddy, J.A. Caballero, J. Earls, J.C. Hood, L. and Price, N.D., (2013). "Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain", Proceedings of the National Academy of Sciences, 110: 3095-3100.

[301] Willard-Mack, C.L., (2006). "Normal structure, function, and histology of lymph nodes", Toxicologic pathology, 34: 409-424.

[302] Anatomy and Physiology of the Female Reproductive System, https://courses.lumenlearning.com/ap2/chapter/anatomy-and-physiology-of-the-female-reproductive-system/, 22 Novemver 2017.

[303] Digestive System, The McGraw–Hill Companies, http://bio.bsu.by/t/temp/holik/24%20Digestive%20System.pdf, 13 Novemver 2017.

[304] Bacha, A.B. Karray, A. Daoud, L. Bouchaala, E. Ali, M.B. Gargouri, Y. and Ali, Y.B., (2011). "Biochemical properties of pancreatic colipase from the common stingray Dasyatis pastinaca", Lipids in health and disease, 10: 69.

[305] Jögi, A. Vaapil, M. Johansson, M. and Påhlman, S., (2012). "Cancer cell differentiation heterogeneity and aggressive behavior in solid tumors", Upsala journal of medical sciences, 117: 217-224.

[306] Snuderl, M. Fazlollahi, L. Le, L.P. Nitta, M. Zhelyazkova, B.H. Davidson, C.J. Akhavanfard, S. Cahill, D.P. Aldape, K.D. and Betensky, R.A., (2011). "Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma", Cancer cell, 20: 810-817.

[307] Nickel, G.C. Barnholtz-Sloan, J. Gould, M.P. McMahon, S. Cohen, A. Adams, M.D. Guda, K. Cohen, M. Sloan, A.E. and LaFramboise, T., (2012). "Characterizing mutational heterogeneity in a glioblastoma patient with double recurrence", PloS one, 7: e35262.

[308] Szerlip, N.J. Pedraza, A. Chakravarty, D. Azim, M. McGuire, J. Fang, Y. Ozawa, T. Holland, E.C. Huse, J.T. and Jhanwar, S., (2012). "Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with distinct growth factor response", Proceedings of the National Academy of Sciences, 109: 3041-3046.

[309] Kreso, A. O'brien, C.A. van Galen, P. Gan, O.I. Notta, F. Brown, A.M. Ng, K. Ma, J. Wienholds, E. and Dunant, C., (2013). "Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer", science, 339: 543-548.

[310] Diaz Jr, L.A. Williams, R.T. Wu, J. Kinde, I. Hecht, J.R. Berlin, J. Allen, B. Bozic, I. Reiter, J.G. and Nowak, M.A., (2012). "The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers", Nature, 486: 537-540.

[311] Cheang, M.C. van de Rijn, M. and Nielsen, T.O., (2008). "Gene expression profiling of breast cancer", Annu. Rev. pathmechdis. Mech. Dis., 3: 67-97.

[312] Li, H. Fan, X. and Houghton, J., (2007). "Tumor microenvironment: the role of the tumor stroma in cancer", Journal of cellular biochemistry, 101: 805-815.

[313] Gajewski, T.F. Schreiber, H. and Fu, Y.-X., (2013). "Innate and adaptive immune cells in the tumor microenvironment", Nature immunology, 14: 1014-1022.

[314] Gkretsi, V. Stylianou, A. Papageorgis, P. Polydorou, C. and Stylianopoulos, T., (2015). "Remodeling components of the tumor microenvironment to enhance cancer therapy", Frontiers in oncology, 5.

[315] Pagès, F. Berger, A. Camus, M. Sanchez-Cabo, F. Costes, A. Molidor, R. Mlecnik, B. Kirilovsky, A. Nilsson, M. and Damotte, D., (2005). "Effector memory

T cells, early metastasis, and survival in colorectal cancer", New England Journal of Medicine, 353: 2654-2666.

[316] Section 24.1,Tumor Cells and the Onset of Cancer, ed. Molecular Cell Biology. 4th edition, (2000), W. H. Freeman and Company.

[317] Chaffer, C.L. ve Weinberg, R.A., (2011). "A perspective on cancer cell metastasis", science, 331: 1559-1564.

[318] Health Line, Liver Metastasis, https://www.healthline.com/health/liver-metastases#overview1, 5 November 2017.

[319] Kreike, B. van Kouwenhove, M. Horlings, H. Weigelt, B. Peterse, H. Bartelink, H. and van de Vijver, M.J., (2007). "Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas", Breast Cancer Research, 9: R65.

[320] Kobayashi, H. Tanaka, Y. Yagi, J. Minato, N. and Tanabe, K., (2011). "Phase I/II study of adoptive transfer of γδ T cells in combination with zoledronic acid and IL-2 to patients with advanced renal cell carcinoma", Cancer Immunology, Immunotherapy, 60: 1075-1084.

[321] Zhang, L. Conejo-Garcia, J.R. Katsaros, D. Gimotty, P.A. Massobrio, M. Regnani, G. Makrigiannakis, A. Gray, H. Schlienger, K. and Liebman, M.N., (2003). "Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer", New England Journal of Medicine, 348: 203-213.

[322] Rodriguez, P.C. Quiceno, D.G. Zabaleta, J. Ortiz, B. Zea, A.H. Piazuelo, M.B. Delgado, A. Correa, P. Brayer, J. and Sotomayor, E.M., (2004). "Arginase I production in the tumor microenvironment by mature myeloid cells inhibits T-cell receptor expression and antigen-specific T-cell responses", Cancer research, 64: 5839-5849.

[323] Simpson, A.J. Caballero, O.L. Jungbluth, A. Chen, Y.-T. and Old, L.J., (2005). "Cancer/testis antigens, gametogenesis and cancer", Nature Reviews Cancer, 5: 615-625.

[324] da Silva, V.L. Fonseca, A.F. Fonseca, M. da Silva, T.E. Coelho, A.C. Kroll, J.E. de Souza, J.E.S. Stransky, B. de Souza, G.A. and de Souza, S.J., (2017). "Genome-wide identification of cancer/testis genes and their association with prognosis in a pan-cancer analysis", Oncotarget, 8: 92966.

[325] Scanlan, M.J. Simpson, A.J. and Old, L.J., (2004). "The cancer/testis genes: review, standardization, and commentary", Cancer Immunity Archive, 4: 1.

[326] Afsharpad, M. Nowroozi, M.R. Mobasheri, M.B. Ayati, M. Nekoohesh, L. Saffari, M. Zendehdel, K. and Modarressi, M.H., (2017). "Cancer-Testis Antigens as New Candidate Diagnostic Biomarkers for Transitional Cell Carcinoma of Bladder", Pathology & Oncology Research: 1-9.

[327] Takeda, R. Hirohashi, Y. Shen, M. Wang, L. Ogawa, T. Murai, A. Yamamoto, E. Kubo, T. Nakatsugawa, M. and Kanaseki, T., (2017). "Identification and functional analysis of variants of a cancer/testis antigen LEMD1 in colorectal

cancer stem-like cells", Biochemical and biophysical research communications, 485: 651-657.

[328] Axelsen, J.B. Lotem, J. Sachs, L. and Domany, E., (2007). "Genes overexpressed in different human solid cancers exhibit different tissue-specific expression profiles", Proceedings of the National Academy of Sciences, 104: 13122-13127.

[329] Mancino, M. Ametller, E. Gascón, P. and Almendro, V., (2011). "The neuronal influence on tumor progression", Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 1816: 105-118.

[330] Dolle, L. El Yazidi-Belkoura, I. Adriaenssens, E. Nurcombe, V. and Hondermarck, H., (2003). "Nerve growth factor overexpression and autocrine loop in breast cancer cells", Oncogene, 22: 5592-5601.

[331] Petrizzo, A. Caruso, F.P. Tagliamonte, M. Tornesello, M.L. Ceccarelli, M. Costa, V. Aprile, M. Esposito, R. Ciliberto, G. and Buonaguro, F.M., (2016). "Identification and Validation of HCC-specific Gene Transcriptional Signature for Tumor Antigen Discovery", Scientific reports, 6.

[332] Eisenberg, I. Eran, A. Nishino, I. Moggio, M. Lamperti, C. Amato, A.A. Lidov, H.G. Kang, P.B. North, K.N. and Mitrani-Rosenbaum, S., (2007). "Distinctive patterns of microRNA expression in primary muscular disorders", Proceedings of the National Academy of Sciences, 104: 17016-17021.

[333] Nasser, M.W. Datta, J. Nuovo, G. Kutay, H. Motiwala, T. Majumder, S. Wang, B. Suster, S. Jacob, S.T. and Ghoshal, K., (2008). "Down-regulation of micro-RNA-1 (miR-1) in lung cancer suppression of tumorigenic property of lung cancer cells and their sensitization to doxorubicin-induced apoptosis by miR-1", Journal of Biological Chemistry, 283: 33394-33405.

[334] Sokol, L. Caceres, G. Volinia, S. Alder, H. Nuovo, G.J. Liu, C.G. McGraw, K. Clark, J.A. Sigua, C.A. and Chen, D.T., (2011). "Identification of a risk dependent microRNA expression signature in myelodysplastic syndromes", British journal of haematology, 153: 24-32.

[335] Kudo, Y. Iizuka, S. Yoshida, M. Tsunematsu, T. Kondo, T. Subarnbhesaj, A. Deraz, E.M. Siriwardena, S.B. Tahara, H. and Ishimaru, N., (2012). "Matrix metalloproteinase-13 (MMP-13) directly and indirectly promotes tumor angiogenesis", Journal of Biological Chemistry, 287: 38716-38728.

APPENDIX-A

CALCULATION OF STATISTICAL DATA

In this study, determination of tissue-specific genes was carried out combination of two powerful applications. The first one is a robust specificity index, tau score, the second one is statistically significantly interval estimation as a rigorous procedure. To calculate statistical distance for each dataset optimized threshold is a certain necessity. For this reason, threshold values were calculated, and they are shown for each tissue sample in each dataset in Table A.1, Table A.2, Table A.3, Table A.4 and Table A.5, respectively.

Table A.1 Threshold ratios for EMTAB-1733

	colon	kidney	liver	pancreas	lung	prostate	brain	stomach	salivary. Gland
upper cluster	0.41	1.86	331.62	33321.87	6.03	1.62	0.25	650.39	1524.29
lower cluster	4.18	7.46	48.45	27.23	11.19	5.46	8.69	20.22	54.23
average of clusters	2.29	4.66	190.03	16674.55	8.61	3.54	4.47	335.30	789.26
number of upper cluster element	276.00	287.00	106.00	9.00	89.00	232.00	579.00	17.00	22.00
number of non-zero element	2305.00	2211.00	1948.00	1762.00	2227.00	2267.00	2196.00	2274.00	2048.00
threshold ratio	0.12	0.13	0.05	0.01	0.04	0.10	0.26	0.01	0.01
	spleen	lymph. Node	appendix	small. Intestine	adrenal. Gland	duodenum	adipose. Tissue	endometrium	bone. Marrow
upper cluster	5.01	3.45	2.68	10.17	0.40	15.99	0.70	2.94	328.98
lower cluster	1.38	1.37	1.60	13.38	5.66	12.39	2.37	1.39	19.77
average of clusters	3.20	2.41	2.14	11.78	3.03	14.19	1.54	2.17	174.37
number of upper cluster element	173.00	184.00	276.00	164.00	294.00	148.00	252.00	244.00	23.00
number of non-zero element	2146.00	2233.00	2141.00	2221.00	2013.00	2007.00	2100.00	2116.00	1887.00
threshold ratio	0.08	0.08	0.13	0.07	0.15	0.07	0.12	0.12	0.01
	placenta	testis	gall. Bladder	urinary. Bladder	thyroid	esophagus	Heart	skin	ovary
upper cluster	17.05	1.18	0.53	0.47	0.40	9.63	22.84	2.10	121.14
lower cluster	8.50	16.69	2.13	1.60	3.39	22.49	20.09	13.35	1.29
average of clusters	12.78	8.94	1.33	1.03	1.90	16.06	21.47	7.73	61.21
number of upper cluster element	119.00	827.00	556.00	541.00	248.00	156.00	113.00	236.00	18.00
number of non-zero element	2228.00	2646.00	2209.00	2047.00	2131.00	2122.00	2076.00	2134.00	1920.00

Table A.2 Threshold ratios for EMTAB-2836

	adipose. Tissue	adrenal. Gland	bone. Marrow	cerebral. Cortex	colon	duodenum	endometrium	esophagus
upper cluster	0.17	0.28	0.45	0.06	0.08	60.58	0.62	18.66
lower cluster	2.09	4.51	9.00	6.94	2.68	12.20	0.81	17.49
average of clusters	1.13	2.40	4.73	3.50	1.38	36.39	0.71	18.08
number of upper cluster element	240.00	345.00	198.00	607.00	343.00	107.00	644.00	117.00
number of non-zero element	1799.00	1886.00	1391.00	1963.00	2000.00	2141.00	1888.00	1854.00
threshold ratio	0.13	0.18	0.14	0.31	0.17	0.05	0.34	0.06
	oviduct	gall. Bladder	heart	kidney	liver	lung	lymph. node	ovary
upper cluster	0.33	0.06	41.57	3.80	939.98	2.35	0.12	16.24
lower cluster	3.37	1.88	13.53	6.26	32.71	8.65	1.30	1.07
average of clusters	1.85	0.97	27.55	5.03	486.34	5.50	0.71	8.65
number of upper cluster element	583.00	793.00	74.00	268.00	34.00	137.00	541.00	67.00
number of non-zero element	2092.00	2044.00	1570.00	2035.00	1454.00	2020.00	1748.00	1701.00
threshold ratio	0.28	0.39	0.05	0.13	0.02	0.07	0.31	0.04
	pancreas	placenta	prostate	rectum	salivary. Gland	skeletal. Muscle	small. Intestine	smooth. Muscle
upper cluster	36487.46	12.26	0.49	0.06	1951.12	447.87	80.08	0.02
lower cluster	33.48	2.57	5.75	3.06	34.89	32.82	14.01	0.67
average of clusters	18260.47	7.42	3.12	1.56	993.00	240.35	47.05	0.34
number of upper cluster element	6.00	137.00	182.00	371.00	12.00	42.00	79.00	967.00
number of non-zero element	1527.00	1883.00	2115.00	2079.00	1806.00	1092.00	2095.00	1948.00
threshold ratio	0.00	0.07	0.09	0.18	0.01	0.04	0.04	0.50
	spleen	stomach	testis	thyroid	tonsil	urinary. Bladder	appendix	skin
upper cluster	0.43	46.07	0.33	1.45	1.50	0.73	0.36	10.47
lower cluster	1.51	14.82	12.49	3.80	4.62	1.28	1.66	11.92
average of clusters	0.97	30.44	6.41	2.62	3.06	1.01	1.01	11.19
number of upper cluster element	604.00	41.00	824.00	201.00	265.00	452.00	475.00	212.00
number of non-zero element	1886.00	2071.00	2729.00	1708.00	1894.00	1926.00	2125.00	1976.00
threshold ratio	0.32	0.02	0.30	0.12	0.14	0.23	0.22	0.11

Table A.3 Threshold ratios for EMTAB-5214

	adrenal. Gland	amygdala	aorta	atrium	whole. Blood	breast	caudate. Nucleus	cerebellar. Hemisphere
upper cluster	4.15	1.21	0.61	10.74	10.67	0.78	1.67	3.46
lower cluster	2.50	0.43	1.24	220.88	964.28	1.14	0.34	0.38
average of clusters	3.32	0.82	0.93	115.81	487.48	0.96	1.00	1.92
number of upper cluster element	120.00	667.00	336.00	26.00	6.00	538.00	497.00	525.00
number of non-zero element	1804.00	1900.00	1713.00	1750.00	1353.00	2007.00	1954.00	1856.00
threshold ratio	0.07	0.35	0.20	0.01	0.00	0.27	0.25	0.28
	cerebellum	cerebral. Cortex	coronary. Artery	cortex. Of kidney	ectocervix	endocervix	gastroesophageal. Junction	esophagus. Mucosa
upper cluster	3.47	2.18	0.54	1.79	0.49	0.62	0.43	17.14
lower cluster	0.69	0.60	1.29	1.88	1.84	1.89	0.89	191.59
average of clusters	2.08	1.39	0.91	1.83	1.16	1.25	0.66	104.37
number of upper cluster element	429.00	563.00	410.00	394.00	188.00	272.00	507.00	56.00
number of non-zero element	1945.00	2005.00	1787.00	2007.00	1931.00	2038.00	1829.00	1891.00
threshold ratio	0.22	0.28	0.23	0.20	0.10	0.13	0.28	0.03
	esophagus.muscularis.mucosa	oviduct	greater. omentum	left. ventricle	hippocampus	hypothalamus	liver	lower.leg.skin
upper cluster	0.41	0.72	0.88	8.62	1.43	1.46	23.37	10.93
lower cluster	1.19	1.87	1.21	43.48	0.44	0.46	63.37	23.03
average of clusters	0.80	1.29	1.05	26.05	0.93	0.96	43.37	16.98
number of upper cluster element	419.00	391.00	308.00	49.00	639.00	764.00	137.00	173.00
number of non-zero element	1780.00	2121.00	1878.00	1528.00	1944.00	2018.00	1752.00	2019.00
threshold ratio	0.24	0.18	0.16	0.03	0.33	0.38	0.08	0.09

Table A.3 Threshold ratios for EMTAB-5214 (cont'd)

	lung	minor. salivary. Gland	nucleus. accumbens	ovary	pancreas	pituitary. Gland	prostate	putamen
upper cluster	3.24	7.84	1.92	0.89	16.75	10.30	2.42	1.58
lower cluster	1.72	98.44	0.31	1.67	14511.71	648.62	1.92	0.34
average of clusters	2.48	53.14	1.11	1.28	7264.23	329.46	2.17	0.96
number of upper cluster element	344.00	26.00	508.00	223.00	10.00	6.00	197.00	609.00
number of non-zero element	2000.00	2047.00	1953.00	1798.00	1855.00	2157.00	2090.00	1888.00
threshold ratio	0.17	0.01	0.26	0.12	0.01	0.00	0.09	0.32
	sigmoid. Colon	skeletal. muscle	small. intestine. Peyers.patch	spleen	stomach	subcutaneous. adipose. Tissue	substantia. nigra	suprapubic. Skin
upper cluster	0.56	17.48	5.74	3.40	7.86	0.88	1.08	9.80
lower cluster	0.92	238.21	60.75	4.37	135.22	0.52	0.54	23.72
average of clusters	0.74	127.84	33.25	3.89	71.54	0.70	0.81	16.76
number of upper cluster element	442.00	43.00	61.00	404.00	26.00	605.00	655.00	165.00
number of non-zero element	1900.00	1397.00	2091.00	1921.00	2009.00	1794.00	1919.00	2023.00
threshold ratio	0.23	0.03	0.03	0.21	0.01	0.34	0.34	0.08
	testis	thyroid	tibial. Artery	tibial. nerve	transverse. Colon	urinary. bladder	uterus	vagina
upper cluster	17.43	2.30	0.45	1.81	3.29	0.60	0.51	6.54
lower cluster	59.28	0.80	0.60	0.51	1.47	2.35	1.52	69.08
average of clusters	38.35	1.55	0.53	1.16	2.38	1.48	1.02	37.81
number of upper cluster element	213.00	387.00	438.00	362.00	304.00	200.00	178.00	57.00
number of non-zero element	2685.00	1937.00	1598.00	1897.00	2033.00	1911.00	1825.00	2017.00
threshold ratio	0.08	0.20	0.27	0.19	0.15	0.10	0.10	0.03

Table A.4 Threshold ratios for EMTAB-3358

	amygdala	artery	bone. Marrow	brain. Meninx	brain	breast	caudate. Nucleus	cerebellum
upper cluster	0.54	14.61	3.76	0.82	1.30	1.78	0.93	0.99
lower cluster	0.00	0.01	0.00	0.00	0.00	0.05	0.00	0.00
average of clusters	0.27	7.31	1.88	0.41	0.65	0.91	0.46	0.49
number of upper cluster element	620.00	133.00	291.00	595.00	489.00	317.00	576.00	392.00
number of non-zero element	760.00	475.00	662.00	985.00	552.00	592.00	795.00	645.00
threshold ratio	0.82	0.28	0.44	0.60	0.89	0.54	0.72	0.61
	colon	diencephalon	thalamus	Dura mater	epididymis	gall. Bladder	Globus. pallidus	left. ventricle
upper cluster	6.66	0.65	0.52	0.94	4.90	4.45	0.47	3.98
lower cluster	0.00	0.00	0.00	1.17	2.44	0.14	0.00	0.04
average of clusters	3.33	0.32	0.26	1.05	3.67	2.29	0.23	2.01
number of upper cluster element	227.00	610.00	657.00	258.00	266.00	242.00	637.00	175.00
number of non-zero element	634.00	813.00	791.00	1037.00	1447.00	1026.00	740.00	531.00
threshold ratio	0.36	0.75	0.83	0.25	0.18	0.24	0.86	0.33
	heart	hippocampus	kidney	left. Atrium	locus. coeruleus	lung	lymph. Node	medulla. oblongata
upper cluster	3.67	0.54	4.12	3.31	0.66	9.32	2.29	1.34
lower cluster	2.36	0.00	0.00	0.13	0.00	4.27	0.00	0.07
average of clusters	3.02	0.27	2.06	1.72	0.33	6.80	1.14	0.71
number of upper cluster element	174.00	695.00	227.00	268.00	642.00	150.00	235.00	511.00
number of non-zero element	932.00	834.00	1105.00	657.00	1060.00	900.00	585.00	945.00
threshold ratio	0.19	0.83	0.21	0.41	0.61	0.17	0.40	0.54
	middle. frontal gyrus	middle. temporal. Gyrus	mitral. Valve	occipital. Cortex	occipital. Lobe	olfactory. Apparatus	ovary	pancreas
upper cluster	0.50	0.55	2.26	0.46	0.57	0.53	0.69	204.07
lower cluster	0.00	0.00	0.29	0.00	0.16	0.06	0.00	18434.15
average of clusters	0.25	0.28	1.27	0.23	0.37	0.30	0.35	9319.11
number of upper cluster element	617.00	680.00	238.00	507.00	574.00	676.00	641.00	10.00
number of non-zero element	783.00	947.00	835.00	725.00	840.00	899.00	1046.00	551.00

Table A.4 Threshold ratios for EMTAB-3358 (cont'd)

	parietal. Lobe	parotid. Gland	penis	pineal. Gland	pituitary. Gland	placenta	prostate	pulmonary. Valve
upper cluster	0.49	101.29	7.26	3.92	3.37	10.84	1.73	3.38
lower cluster	0.00	186455.87	17.60	0.06	0.50	0.00	0.51	0.21
average of clusters	0.25	93278.58	12.43	1.99	1.93	5.42	1.12	1.80
number of upper cluster element	619.00	4.00	93.00	250.00	241.00	148.00	249.00	218.00
number of non-zero element	833.00	686.00	966.00	853.00	919.00	1014.00	1176.00	885.00
threshold ratio	0.74	0.01	0.10	0.29	0.26	0.15	0.21	0.25
	putamen	seminal.Vesicle	smooth.Muscle	spinal. Cord	spleen	submandibular. Gland	substantia.nigra	testis
upper cluster	0.86	15.15	0.60	0.59	2.39	76.94	1.43	25.37
lower cluster	0.00	3.24	0.12	0.00	0.06	191010.96	0.10	6.81
average of clusters	0.43	9.20	0.36	0.30	1.22	95543.95	0.77	16.09
number of upper cluster element	521.00	60.00	506.00	687.00	274.00	1.00	175.00	363.00
number of non-zero element	810.00	755.00	759.00	861.00	781.00	842.00	2063.00	1446.00
threshold ratio	0.64	0.08	0.67	0.80	0.35	0.00	0.08	0.25
	tongue	tricuspid. Valve	cervix	uterus	vagina	vas. deferens	appendix	zone.of.skin
upper cluster	31.66	1.87	5.85	1.19	0.63	10.87	3.70	14.85
lower cluster	153.95	0.00	0.25	3.36	0.23	27.47	6.80	0.42
average of clusters	92.81	0.94	3.05	2.27	0.43	19.17	5.25	7.63
number of upper cluster element	65.00	261.00	143.00	119.00	521.00	40.00	198.00	184.00
number of non-zero element	1145.00	488.00	1172.00	1105.00	584.00	807.00	945.00	300.00
threshold ratio	0.06	0.53	0.12	0.11	0.89	0.05	0.21	0.61

Table A.5 Threshold ratios for EMTAB-4344

	adipose. Tissue	adrenal. Gland	brain	heart	kidney	liver	lung	ovary
upper cluster	2.60	9.09	5.58	14.38	3.63	34.02	9.61	1.55
lower cluster	50.04	51.07	0.08	37.46	6.99	1351.06	59.45	1.59
average of clusters	26.32	30.08	2.83	25.92	5.31	692.54	34.53	1.57
number of upper cluster element	25.00	33.00	488.00	85.00	198.00	22.00	48.00	273.00
number of non-zero element	1576.00	1429.00	1412.00	1149.00	1540.00	1247.00	1520.00	1453.00
threshold ratio	0.02	0.02	0.35	0.07	0.13	0.02	0.03	0.19
	pancreas	sigmoid. Colon	small. Intestine	spleen	testis			
upper cluster	13.13	1.72	11.47	3.55	15.83			
lower cluster	10124.85	0.51	23.86	0.25	3.84			
average of clusters	5068.99	1.11	17.66	1.90	9.84			
number of upper cluster element	13.00	380.00	214.00	454.00	795.00			
number of non-zero element	1553.00	1558.00	1125.00	1437.00	2154.00			
threshold ratio	0.01	0.24	0.19	0.32	0.37			

R programming language and its very useful packages are used all analysis of thesis. Besides, all graphs are generated using R. There is an example of generating a network (Figure 7.8) Firstly, data were prepared:

```
all_geneTissuePair <- read.csv("all_geneTissuePair.csv", stringsAsFactors = FALSE)
all_geneTissuePair<-tbl_df(all_geneTissuePair)
includedList1733 <-read.csv("includedList1733.csv", stringsAsFactors = FALSE)
includedList2836 <-read.csv("includedList2836.csv", stringsAsFactors = FALSE)
includedList5214 <-read.csv("includedList5214.csv", stringsAsFactors = FALSE)
includedList4344 <-read.csv("includedList4344.csv", stringsAsFactors = FALSE)
includedList3358 <-read.csv("includedList3358.csv", stringsAsFactors = FALSE)
includedList1733<-tbl_df(includedList1733)
includedList2836<-tbl_df(includedList2836)
includedList5214<-tbl_df(includedList5214)
includedList4344<-tbl_df(includedList4344)
includedList3358<-tbl_df(includedList3358)
includedList1733 <- mutate(includedList1733, Ensembl.Gene.ID=EnsemblGeneID)
joined <- includedList1733 %>%
 select(Ensembl.Gene.ID,Tau.Score) %>%
 full_join( select(includedList2836, Ensembl.Gene.ID,Tau.Score), by="Ensembl.Gene.ID
", suffix=c(".1733",".2836")) %>% full_join( select(includedList5214, Ensembl.Gene.ID,T
au.Score), by="Ensembl.Gene.ID") %>% full_join( select(includedList4344, Ensembl.Ge
ne.ID,Tau.Score), by="Ensembl.Gene.ID",suffix=c(".5214",".4344")) %>%full_join( selec
t(includedList3358, Ensembl.Gene.ID,Tau.Score), by="Ensembl.Gene.ID") %>% rename
(Tau.Score.3358 = Tau.Score)
```

This part prepares data to be used by tidygraph and ggraph. First, joined data frame, which is wide and full of NA's. It is converted into narrow format:

```
joined_gather <- gather(joined,"sample","tauScore",2:6) %>%
        mutate(sample=str_replace(sample,"Tau.Score.","E-MTAB-"))
```

Now we can join all_geneTissuePair with joined_gather so that we have gene id, tissue, expressionLevel, sample name and tauScore in one table:

```
joined_genepair <- all_geneTissuePair %>%
 select(-X) %>%
 rename("sample"="file","Ensembl.Gene.ID"="gene") %>%
 mutate(sample=substr(sample,1,11)) %>%
 mutate(Ensembl.Gene.ID=as.character(Ensembl.Gene.ID)) %>%
 left_join(joined_gather)
```

Preparing a data frame with two columns, from and to, which will be edges of the network. We tried to use the whole joined_genepair table with changing column names but it was confused, it created edges between genes and tissues. So, it just kept two columns (and planning to join the data later). Sampling 4000 interactions only, since 19000 is slow to graph.

UPDATE1: the columns does not need to be named from and to. We just kept Ensembl.Gene.ID and sample columns and it worked. Node name becomes the first column (Ensembl.Gene.ID)

UPDATE2: Ensembl.Gene.ID, sample, tauScore, tissue columns are expected in this order, so first two columns should be the nodes we want to connect:

```
set.seed(1)
relations <- joined_genepair %>%
        select(Ensembl.Gene.ID,sample,tauScore,tissue) %>%
        sample_n(4000)
```

With "graph from data frame" functions (part of igraph library), we can generate igraph objects first then we convert it into graph tbl (part of tidygraph library):

```
 g <- graph_from_data_frame(relations)
g2 <- as_tbl_graph(g)
```

Below, we used create_layout to determine x, y coordinates of nodes so that ggraph() does not need to calculate the layout each time we plot. Thus, run the code below once, after that next chunk with ggraph will be quicker to run. More info about available layout algorithms are described in igraph layouts page. Available layouts are:

- drl

- kk : kamada kawai
- fr: fruchterman reingold
- graphopt
- lgl
- bipartite (didn't work, needs types argument)
- mds
- sugiyama
- gem (this was not listed in igraph page, I found it by typing layout. in console. extremely slow and didnt finish after long time)
- grid

drl layout gives best result to show genes that are specific in more than one sample. Next best one is kk layout. sugiyama algoritm reveals bipartite nature of the graph.

```
layout <- create_layout(g2, layout = 'kk')
```

Draw the network: By tidygraph we can use dplyr verbs on nodes or edges. We added two columns for nodes, indicating if the node is sample or gene type. ggraph() colors sample type nodes with forestgreen (node size is 10) and gene type nodes with firebrick color with node size being 1.

```
g2 %>% activate(nodes) %>%
 mutate(sample=grepl("MTAB",name) , gene=grepl("ENSG",name)) %>%
 ggraph('manual',node.position=layout) +
 geom_edge_link(aes(color=tauScore, alpha=0.2)) +
 geom_node_point(aes(filter = gene), colour = 'firebrick', size = 1) +
 geom_node_point(aes(filter = sample), colour = 'forestgreen', size = 10) +
 geom_node_text(aes(filter = sample, label = name), vjust = 0.4) +
 theme_graph()
```

After that, if we run the code we will obtain Figure 7.8.

APPENDIX-B

GENERATED RESULTS

There are two main objectives in thesis study. First one is generating tissue-specific gene list via a robust and rigorous computational approach. Second one is investigation of intra-tumoral heterogeneity. Some of the results produced are given in long lists below. Table B.1 shows number of tissue-specific genes in unrelated tissues in particular cancer differentially expressed genes to gain basic insight about cancer heterogeneity.

Table B.1 Tissue-specific genes in irrelevant tissues to a particular cancer in DEG

Liver cancer		Prostate cancer		Esophageal cancer		Papillary renal cell carcinoma	
others	**number**	**others**	**number**	**others**	**number**	**others**	**number**
adipose	13	adipose	25	adipose	60	adipose	10
adrenal gland	20	adrenal gland	50	adrenal gland	72	adrenal gland	16
appendix	17	appendix	26	appendix	106	appendix	32
b. marrow	9	b. marrow	26	b. marrow	86	b. marrow	35
brain	102	brain	184	brain	322	brain	101
breast	6	breast	9	breast	16	breast	3
colon	19	colon	39	colon	100	colon	30
epididymis	7	epididymis	12	epididymis	27	epididymis	5
esophagus	20	esophagus	60	heart	113	esophagus	14
heart	26	heart	92	kidney	121	heart	29
kidney	32	kidney	70	liver	140	liver	44
lung	15	liver	77	lung	125	lung	25
lymph node	1	lung	54	lymph node	59	lymph node	13
ovary	13	lymph node	8	ovary	63	ovary	7
oviduct	9	ovary	37	oviduct	43	oviduct	9
pancreas	23	oviduct	23	pancreas	85	pancreas	17

Table B.1 Tissue-specific genes in irrelevant tissues to a particular cancer in DEG (cont'd)

penis	6	pancreas	33	penis	18	penis	2	
placenta	27	penis	12	placenta	88	placenta	19	
prostate	14	placenta	41	prostate	34	prostate	9	
rectum	8	rectum	12	rectum	38	rectum	6	
salivary gland	12	salivary gland	22	salivary gland	28	salivary gland	7	
seminal vesicle	4	seminal vesicle	17	seminal vesicle	25	seminal vesicle	5	
skin	18	skin	52	skin	42	skin	18	
small intestine	26	small intestine	50	small intestine	150	small intestine	22	
spleen	12	spinal cord	3	spinal cord	3	spleen	36	
stomach	10	spleen	40	spleen	144	stomach	12	
testis	73	stomach	29	stomach	61	testis	93	
thyroid	6	testis	156	testis	340	thyroid	7	
tibial nerve	4	thyroid	26	thyroid	36	tibial nerve	7	
tongue	22	tibial nerve	10	tibial nerve	17	tongue	28	
uterus	2	tongue	60	tongue	44	uterus	1	
vagina	7	uterus	1	uterus	19	vagina	8	
vas deferens	6	vagina	25	vagina	21	vas deferens	6	
whole blood	2	vas deferens	20	vas deferens	20	whole blood	16	

Table B.1 Tissue-specific genes in irrelevant tissues to a particular cancer in DEG (cont'd)

Lung squamous cell carcinoma		Pancreas adenocarcinoma		Stomach cancer		Thyroid cancer	
others	**number**	**others**	**number**	**others**	**number**	**others**	**number**
adipose	51	adipose	2	adipose	15	adipose	11
adrenal gland	100	adrenal gland	11	adrenal gland	34	adrenal gland	25
appendix	94	appendix	3	appendix	6	appendix	9
b. marrow	118	b.marrow	3	b.marrow	15	b.marrow	25
brain	406	brain	31	brain	127	brain	100
breast	16	breast	2	breast	6	breast	4
colon	84	colon	1	colon	12	colon	17
epididymis	34	epididymis	2	epididymis	7	epididymis	10
esophagus	121	esophagus	1	heart	37	esophagus	38
heart	138	heart	2	kidney	76	heart	37
kidney	163	kidney	10	liver	21	liver	37
lung	187	liver	12	lung	26	lung	34
lymph node	57	lung	5	lymph node	31	lymph node	49
ovary	71	lymph node	2	ovary	3	ovary	2
oviduct	88	ovary	4	oviduct	23	oviduct	20
pancreas	79	oviduct	1	pancreas	13	pancreas	19
penis	22	pancreas	1	penis	14	penis	18

Table B.1 Tissue-specific genes in irrelevant tissues to a particular cancer in DEG (cont'd)

placenta	96	penis	1	placenta	12	placenta	4	
prostate	48	placenta	1	prostate	27	prostate	24	
rectum	29	rectum	2	rectum	21	rectum	11	
salivary gland	43	salivary gland	6	salivary gland	4	salivary gland	6	
seminal vesicle	17	seminal vesicle	6	seminal vesicle	14	seminal vesicle	8	
skin	153	skin	2	skin	9	skin	5	
small intestine	156	small intestine	13	small intestine	43	small intestine	40	
spleen	7	spinal cord	3	spinal cord	7	spleen	39	
stomach	175	spleen	1	spleen	2	stomach	16	
testis	63	stomach	2	stomach	8	testis	21	
thyroid	519	testis	3	testis	73	thyroid	69	
tibial nerve	50	thyroid	2	thyroid	11	tibial nerve	5	
tongue	23	tibial nerve	11	tibial nerve	10	tongue	34	
uterus	120	tongue	3	tongue	43	uterus	4	
vagina	11	uterus	3	uterus	6	vagina	18	
vas deferens	61	vagina	31	vagina	10	vas deferens	4	
whole blood	17	vas deferens	2	vas deferens	9	whole blood	10	

Table B.1 Tissue-specific genes in irrelevant tissues to a particular cancer in DEG (cont'd)

Uterine cancer		Cervical cancer		Urinary bladder cancer	
others	**number**	**others**	**number**	**others**	**number**
adipose	58	adipose	34	adipose	70
adrenal gland	69	adrenal gland	53	adrenal gland	67
appendix	67	appendix	62	appendix	82
b.marrow	78	b.marrow	56	b.marrow	87
brain	293	brain	247	brain	328
breast	20	breast	10	breast	17
colon	66	colon	46	colon	56
epididymis	32	epididymis	21	epididymis	23
esophagus	72	esophagus	98	heart	54
heart	114	heart	91	kidney	136
kidney	97	kidney	62	liver	106
lung	95	liver	55	lung	107
lymph node	120	lung	79	lymph node	96
ovary	30	lymph node	31	ovary	53
oviduct	82	ovary	63	oviduct	79
pancreas	67	oviduct	31	pancreas	44
penis	69	pancreas	42	penis	57
placenta	12	penis	7	placenta	22

Table B.1 Tissue-specific genes in irrelevant tissues to a particular cancer in DEG (cont'd)

prostate	85	placenta	68	prostate	87	
rectum	32	rectum	30	rectum	35	
salivary gland	17	salivary gland	8	salivary gland	21	
seminal vesicle	29	seminal vesicle	21	seminal vesicle	32	
skin	20	skin	16	skin	18	
small intestine	68	small intestine	110	small intestine	72	
spleen	101	spinal cord	88	spinal cord	85	
stomach	4	spleen	4	spleen	3	
testis	123	stomach	86	stomach	141	
thyroid	29	testis	33	testis	29	
tibial nerve	352	thyroid	266	thyroid	336	
tongue	37	tibial nerve	30	tibial nerve	39	
uterus	20	tongue	15	tongue	28	
vagina	71	uterus	83	uterus	71	
vas deferens	40	vagina	17	vagina	15	
whole blood	18	vas deferens	48	vas deferens	30	
		Whole blood	34	Whole blood	46	

We can see the number of specific-genes in each unrelated cancer type showing in Figure 7.15 based on TCGA data for eleven different solid tumors in Table B.2. Moreover, Table B.3 indicate selected genes after all criteria.

Table B.2 Tissue-specific genes in irrelevant tissues to a single cancer type (according TCGA cancer expression data)

blca		brca		coad		esca		kirc	
tissue	number	tissue	number	tissue	number	tissue	number	tissue	number
liver	15	testis	28	liver	9	small.intestine	8	liver	15
placenta	10	liver	18	salivary.gland	6	liver	6	bone.marrow	8
salivary.gland	7	esophagus	11	brain	5	testis	6	pancreas	6
skeletal.muscle	6	brain	10	esophagus	5	skeletal.muscle	3	testis	6
heart	5	salivary.gland	9	pancreas	4	pancreas	2	skeletal.muscle	4
lung	5	bone.marrow	8	skeletal.muscle	4	pituitary.gland	2	brain	3
pancreas	4	pancreas	5	lung	3	salivary.gland	2	heart	3
testis	4	skin	5	stomach	3	bone.marrow	1	small.intestine	3
bone.marrow	3	small.intestine	5	bone.marrow	2	endometrium	1	stomach	3
minor.salivary.gland	3	cerebral.cortex	4	placenta	2	gall.bladder	1	adrenal.gland	2
prostate	3	lung	4	testis	2	heart	1	placenta	2
kidney	2	placenta	4	adrenal.gland	1	kidney	1	prostate	2
stomach	2	kidney	3	cerebral.cortex	1	oviduct	1	salivary.gland	2
tongue	2	adrenal.gland	2	endometrium	1	placenta	1	skin	2
gall.bladder	1	heart	2	gall.bladder	1	skin	1	cerebral.cortex	1
oviduct	1	pituitary.gland	2	heart	1	stomach	1	minor.salivary.gland	1
pineal.gland	1	prostate	2	kidney	1	submandibular.gland	1	ovary	1
pituitary.gland	1	cortex.of.kidney	1	minor.salivary.gland	1	thyroid	1	oviduct	1
skin	1	duodenum	1	oviduct	1			pineal.gland	1
small.intestine	1	endometrium	1	prostate	1			small.intestine.Peyers.patch	1
submandibular.gland	1	gall.bladder	1	submandibular.gland	1			spleen	1
thyroid	1	minor.salivary.gland	1	thyroid	1			thyroid	1
		ovary	1	tongue	1				
		oviduct	1						
		pineal.gland	1						
		stomach	1						
		submandibular. Gland	1						

Table B.2 Tissue-specific genes in irrelevant tissues to a single cancer type according TCGA cancer expression data (cont'd)

kirp		lihc		luad		paad		thca		ucec	
tissue	number	tissue	number	tissue	number	tissue	number	tissue	number	tissue	number
testis	6	testis	20	liver	30	esophagus	8	liver	7	liver	10
liver	5	small. intestine	10	testis	23	liver	8	salivary. Gland	7	testis	8
brain	3	pancreas	8	brain	8	heart	5	testis	5	esophagus	7
heart	3	heart	6	esophagus	8	small. intestine	4	lung	4	heart	6
pancreas	3	kidney	4	pancreas	8	bone. marrow	3	bone. Marrow	3	skeletal. Muscle	5
skeletal.Muscle	3	minor. salivary. gland	4	skeletal. Muscle	8	salivary. gland	3	esophagus	3	brain	4
bone. marrow	2	placenta	4	small. intestine	8	skeletal. Muscle	3	placenta	3	pancreas	4
placenta	2	salivary. gland	4	placenta	7	lung	2	skeletal. Muscle	3	salivary. gland	3
salivary. gland	2	bone. marrow	3	salivary. gland	7	skin	2	adrenal. Gland	1	lung	2
skin	2	brain	3	bone. marrow	6	testis	2	brain	1	minor.salivary.gland	2
small. intestine	2	lung	3	heart	5	minor. salivary.gland	1	cerebral. Cortex	1	bone.marrow	1
stomach	2	prostate	3	kidney	3	placenta	1	kidney	1	kidney	1
whole. Blood	2	esophagus	2	gall. Bladder	2	stomach	1	pancreas	1	pineal.gland	1
adrenal. Gland	1	pituitary. Gland	2	minor. salivary.gland	2			pituitary. Gland	1	pituitary.gland	1
cerebral.Cortex	1	skeletal. Muscle	2	pineal. Gland	2			prostate	1	placenta	1
oviduct	1	adrenal. Gland	1	pituitary. Gland	2			stomach	1	prostate	1
prostate	1	cerebral. Cortex	1	skin	2			submandibular.Gland	1	skin	1
spleen	1	endometrium	1	stomach	2					small.intestine	1
thyroid	1	ovary	1	tongue	2					tongue	1
		stomach	1	adrenal. Gland	1						
		tongue	1	cerebral. Cortex	1						
				endometrium	1						
				oviduct	1						
				prostate	1						
				submandibular.Gland	1						
				thyroid	1						

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data)

Specific genes	Related tissue	Cancer type	Number of cancer observed	Consensus	Specific genes	Related tissue	Cancer type	Number of cancer observed	Consensus
ENSG00000105048	skeletal.muscle	blca	9	2	ENSG00000075886	testis	luad	3	5
ENSG00000170373	salivary.gland	blca	9	2	ENSG00000137948	testis	luad	3	5
ENSG00000170373	salivary.gland	brca	9	2	ENSG00000143452	testis	luad	3	5
ENSG00000105048	skeletal.muscle	coad	9	2	ENSG00000166796	testis	luad	3	5
ENSG00000170373	salivary.gland	coad	9	2	ENSG00000196260	lung	thca	3	5
ENSG00000105048	skeletal.muscle	esca	9	2	ENSG00000143450	testis	ucec	3	5
ENSG00000170373	salivary.gland	esca	9	2	ENSG00000143452	testis	ucec	3	5
ENSG00000105048	skeletal.muscle	kirc	9	2	ENSG00000196260	lung	ucec	3	5
ENSG00000105048	skeletal.muscle	kirp	9	2	ENSG00000007350	testis	blca	3	4
ENSG00000105048	skeletal.muscle	lihc	9	2	ENSG00000129988	liver	blca	3	4
ENSG00000170373	salivary.gland	lihc	9	2	ENSG00000171560	liver	blca	3	4
ENSG00000105048	skeletal.muscle	luad	9	2	ENSG00000229314	liver	blca	3	4
ENSG00000170373	salivary.gland	luad	9	2	ENSG00000110169	liver	brca	3	4
ENSG00000105048	skeletal.muscle	paad	9	2	ENSG00000143194	testis	brca	3	4
ENSG00000170373	salivary.gland	paad	9	2	ENSG00000176566	testis	brca	3	4
ENSG00000170373	salivary.gland	thca	9	2	ENSG00000204019	testis	brca	3	4
ENSG00000105048	skeletal.muscle	ucec	9	2	ENSG00000229314	liver	brca	3	4
ENSG00000170373	salivary.gland	ucec	9	2	ENSG00000111700	liver	coad	3	4
ENSG00000242252	bone.marrow	blca	9	1	ENSG00000198610	liver	coad	3	4
ENSG00000242252	bone.marrow	brca	9	1	ENSG00000139209	liver	esca	3	4
ENSG00000242252	bone.marrow	kirc	9	1	ENSG00000204019	testis	esca	3	4
ENSG00000242252	bone.marrow	kirp	9	1	ENSG00000148965	liver	kirc	3	4
ENSG00000242252	bone.marrow	lihc	9	1	ENSG00000171560	liver	kirc	3	4

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data) (cont'd)

ENSG00000242252	bone.marrow	luad	9	1	ENSG00000007350	testis	lihc	3	4
ENSG00000242252	bone.marrow	paad	9	1	ENSG00000143194	testis	lihc	3	4
ENSG00000242252	bone.marrow	thca	9	1	ENSG00000176566	testis	lihc	3	4
ENSG00000242252	bone.marrow	ucec	9	1	ENSG00000110169	liver	luad	3	4
ENSG00000129991	heart	blca	8	4	ENSG00000111700	liver	luad	3	4
ENSG00000151224	liver	blca	8	4	ENSG00000129988	liver	luad	3	4
ENSG00000129991	heart	brca	8	4	ENSG00000139209	liver	luad	3	4
ENSG00000151224	liver	brca	8	4	ENSG00000143194	testis	luad	3	4
ENSG00000129991	heart	coad	8	4	ENSG00000148965	liver	luad	3	4
ENSG00000151224	liver	coad	8	4	ENSG00000171560	liver	luad	3	4
ENSG00000151224	liver	esca	8	4	ENSG00000176566	testis	luad	3	4
ENSG00000129991	heart	kirp	8	4	ENSG00000198610	liver	luad	3	4
ENSG00000129991	heart	lihc	8	4	ENSG00000204019	testis	luad	3	4
ENSG00000129991	heart	luad	8	4	ENSG00000229314	liver	luad	3	4
ENSG00000151224	liver	luad	8	4	ENSG00000111700	liver	paad	3	4
ENSG00000129991	heart	paad	8	4	ENSG00000198610	liver	paad	3	4
ENSG00000151224	liver	paad	8	4	ENSG00000139209	liver	thca	3	4
ENSG00000151224	liver	thca	8	4	ENSG00000007350	testis	ucec	3	4
ENSG00000129991	heart	ucec	8	4	ENSG00000110169	liver	ucec	3	4
ENSG00000151224	liver	ucec	8	4	ENSG00000129988	liver	ucec	3	4
ENSG00000134551	salivary.gland	blca	8	2	ENSG00000148965	liver	ucec	3	4
ENSG00000231887	salivary.gland	blca	8	2	ENSG00000155622	placenta	blca	3	3
ENSG00000134551	salivary.gland	brca	8	2	ENSG00000170498	placenta	blca	3	3
ENSG00000231887	salivary.gland	brca	8	2	ENSG00000155622	placenta	brca	3	3
ENSG00000134551	salivary.gland	coad	8	2	ENSG00000096088	stomach	coad	3	3
ENSG00000231887	salivary.gland	coad	8	2	ENSG00000112818	small.intestine	esca	3	3
ENSG00000134551	salivary.gland	kirc	8	2	ENSG00000168955	small.intestine	esca	3	3
ENSG00000231887	salivary.gland	kirc	8	2	ENSG00000077522	heart	kirc	3	3
ENSG00000134551	salivary.gland	kirp	8	2	ENSG00000148677	heart	kirc	3	3

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data) (cont'd)

ENSG00000231887	salivary.gland	kirp	8	2	ENSG00000077522	heart	kirp	3	3
ENSG00000134551	salivary.gland	lihc	8	2	ENSG00000148677	heart	kirp	3	3
ENSG00000231887	salivary.gland	lihc	8	2	ENSG00000077522	heart	lihc	3	3
ENSG00000134551	salivary.gland	luad	8	2	ENSG00000077935	testis	lihc	3	3
ENSG00000231887	salivary.gland	luad	8	2	ENSG00000096088	stomach	lihc	3	3
ENSG00000134551	salivary.gland	thca	8	2	ENSG00000112818	small.intestine	lihc	3	3
ENSG00000231887	salivary.gland	thca	8	2	ENSG00000168955	small.intestine	lihc	3	3
ENSG00000101441	salivary.gland	blca	7	2	ENSG00000170498	placenta	lihc	3	3
ENSG00000101441	salivary.gland	brca	7	2	ENSG00000077935	testis	luad	3	3
ENSG00000101441	salivary.gland	coad	7	2	ENSG00000170498	placenta	luad	3	3
ENSG00000101441	salivary.gland	luad	7	2	ENSG00000096088	stomach	paad	3	3
ENSG00000101441	salivary.gland	paad	7	2	ENSG00000112818	small.intestine	paad	3	3
ENSG00000101441	salivary.gland	thca	7	2	ENSG00000168955	small.intestine	paad	3	3
ENSG00000101441	salivary.gland	ucec	7	2	ENSG00000077935	testis	ucec	3	3
ENSG00000085465	oviduct	blca	7	1	ENSG00000148677	heart	ucec	3	3
ENSG00000162078	salivary.gland	blca	7	1	ENSG00000155622	placenta	ucec	3	3
ENSG00000085465	oviduct	brca	7	1	ENSG00000130598	skeletal.muscle	blca	3	2
ENSG00000137745	bone.marrow	brca	7	1	ENSG00000159173	skeletal.muscle	blca	3	2
ENSG00000162078	salivary.gland	brca	7	1	ENSG00000173991	heart	blca	3	2
ENSG00000085465	oviduct	coad	7	1	ENSG00000243137	placenta	blca	3	2
ENSG00000137745	bone.marrow	coad	7	1	ENSG00000143631	skin	brca	3	2
ENSG00000162078	salivary.gland	coad	7	1	ENSG00000157851	brain	brca	3	2
ENSG00000085465	oviduct	esca	7	1	ENSG00000169059	testis	brca	3	2
ENSG00000137745	bone.marrow	esca	7	1	ENSG00000170477	esophagus	brca	3	2
ENSG00000085465	oviduct	kirc	7	1	ENSG00000241794	esophagus	brca	3	2
ENSG00000137745	bone.marrow	kirc	7	1	ENSG00000241794	esophagus	coad	3	2
ENSG00000085465	oviduct	kirp	7	1	ENSG00000101842	testis	kirc	3	2
ENSG00000162078	salivary.gland	lihc	7	1	ENSG00000101842	testis	kirp	3	2
ENSG00000085465	oviduct	luad	7	1	ENSG00000157851	brain	kirp	3	2

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data) (cont'd)

ENSG00000137745	bone.marrow	luad	7	1	ENSG00000243137	placenta	kirp	3	2
ENSG00000162078	salivary.gland	luad	7	1	ENSG00000130598	skeletal.muscle	lihc	3	2
ENSG00000137745	bone.marrow	paad	7	1	ENSG00000169059	testis	lihc	3	2
ENSG00000162078	salivary.gland	paad	7	1	ENSG00000173991	heart	lihc	3	2
ENSG00000137745	bone.marrow	thca	7	1	ENSG00000143631	skin	luad	3	2
ENSG00000162078	salivary.gland	ucec	7	1	ENSG00000157851	brain	luad	3	2
ENSG00000185615	pancreas	blca	6	5	ENSG00000159173	skeletal.muscle	luad	3	2
ENSG00000185615	pancreas	brca	6	5	ENSG00000169059	testis	luad	3	2
ENSG00000185615	pancreas	coad	6	5	ENSG00000241794	esophagus	luad	3	2
ENSG00000185615	pancreas	lihc	6	5	ENSG00000130598	skeletal.muscle	paad	3	2
ENSG00000185615	pancreas	luad	6	5	ENSG00000143631	skin	paad	3	2
ENSG00000185615	pancreas	ucec	6	5	ENSG00000170477	esophagus	paad	3	2
ENSG00000010318	testis	blca	6	4	ENSG00000243137	placenta	thca	3	2
ENSG00000130649	liver	blca	6	4	ENSG00000101842	testis	ucec	3	2
ENSG00000138115	liver	blca	6	4	ENSG00000159173	skeletal.muscle	ucec	3	2
ENSG00000170835	pancreas	blca	6	4	ENSG00000170477	esophagus	ucec	3	2
ENSG00000204983	pancreas	blca	6	4	ENSG00000173991	heart	ucec	3	2
ENSG00000257017	liver	blca	6	4	ENSG00000137142	brain	brca	3	1
ENSG00000130649	liver	brca	6	4	ENSG00000137142	brain	luad	3	1
ENSG00000138115	liver	brca	6	4	ENSG00000137142	brain	ucec	3	1
ENSG00000204983	pancreas	brca	6	4	ENSG00000124490	testis	brca	2	5
ENSG00000170835	pancreas	coad	6	4	ENSG00000124678	testis	brca	2	5
ENSG00000204983	pancreas	coad	6	4	ENSG00000170627	testis	brca	2	5
ENSG00000170835	pancreas	esca	6	4	ENSG00000204140	pancreas	brca	2	5
ENSG00000010318	testis	kirc	6	4	ENSG00000039600	testis	esca	2	5
ENSG00000138115	liver	kirc	6	4	ENSG00000170627	testis	esca	2	5
ENSG00000257017	liver	kirc	6	4	ENSG00000073754	spleen	kirc	2	5
ENSG00000010318	testis	kirp	6	4	ENSG00000073754	spleen	kirp	2	5
ENSG00000138115	liver	kirp	6	4	ENSG00000171864	testis	lihc	2	5

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data) (cont'd)

ENSG00000257017	liver	kirp	6	4	ENSG00000039600	testis	luad	2	5
ENSG00000170835	pancreas	lihc	6	4	ENSG00000124678	testis	luad	2	5
ENSG00000010318	testis	luad	6	4	ENSG00000124490	testis	paad	2	5
ENSG00000130649	liver	luad	6	4	ENSG00000171864	testis	thca	2	5
ENSG00000170835	pancreas	luad	6	4	ENSG00000204140	pancreas	ucec	2	5
ENSG00000204983	pancreas	luad	6	4	ENSG00000142515	prostate	blca	2	4
ENSG00000257017	liver	luad	6	4	ENSG00000163631	liver	blca	2	4
ENSG00000130649	liver	paad	6	4	ENSG00000167751	prostate	blca	2	4
ENSG00000138115	liver	paad	6	4	ENSG00000005421	liver	brca	2	4
ENSG00000010318	testis	thca	6	4	ENSG00000070915	kidney	brca	2	4
ENSG00000130649	liver	thca	6	4	ENSG00000104760	liver	brca	2	4
ENSG00000138115	liver	thca	6	4	ENSG00000151790	liver	brca	2	4
ENSG00000204983	pancreas	thca	6	4	ENSG00000156096	liver	brca	2	4
ENSG00000257017	liver	thca	6	4	ENSG00000214107	testis	brca	2	4
ENSG00000010318	testis	ucec	6	4	ENSG00000228278	liver	brca	2	4
ENSG00000130649	liver	ucec	6	4	ENSG00000238269	testis	brca	2	4
ENSG00000170835	pancreas	ucec	6	4	ENSG00000099937	liver	coad	2	4
ENSG00000204983	pancreas	ucec	6	4	ENSG00000136881	liver	esca	2	4
ENSG00000257017	liver	ucec	6	4	ENSG00000151790	liver	esca	2	4
ENSG00000105697	liver	blca	6	3	ENSG00000172023	pancreas	esca	2	4
ENSG00000014257	prostate	brca	6	3	ENSG00000111305	testis	kirc	2	4
ENSG00000105697	liver	brca	6	3	ENSG00000136881	liver	kirc	2	4
ENSG00000014257	prostate	coad	6	3	ENSG00000142515	prostate	kirc	2	4
ENSG00000105697	liver	kirc	6	3	ENSG00000172023	pancreas	kirc	2	4
ENSG00000105697	liver	kirp	6	3	ENSG00000180210	liver	kirc	2	4
ENSG00000014257	prostate	lihc	6	3	ENSG00000111305	testis	kirp	2	4
ENSG00000014257	prostate	luad	6	3	ENSG00000100253	kidney	lihc	2	4
ENSG00000105697	liver	luad	6	3	ENSG00000164266	pancreas	lihc	2	4
ENSG00000014257	prostate	thca	6	3	ENSG00000167751	prostate	lihc	2	4

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data) (cont'd)

ENSG00000014257	prostate	ucec	6	3	ENSG00000214107	testis	lihc	2	4
ENSG00000105697	liver	ucec	6	3	ENSG00000005421	liver	luad	2	4
ENSG00000115138	pituitary.gland	blca	6	2	ENSG00000091583	liver	luad	2	4
ENSG00000170231	small.intestine	blca	6	2	ENSG00000100253	kidney	luad	2	4
ENSG00000170369	salivary.gland	blca	6	2	ENSG00000104760	liver	luad	2	4
ENSG00000115138	pituitary.gland	brca	6	2	ENSG00000128040	testis	luad	2	4
ENSG00000140519	esophagus	brca	6	2	ENSG00000151655	liver	luad	2	4
ENSG00000170231	small.intestine	brca	6	2	ENSG00000156096	liver	luad	2	4
ENSG00000170369	salivary.gland	brca	6	2	ENSG00000163631	liver	luad	2	4
ENSG00000205420	esophagus	brca	6	2	ENSG00000164266	pancreas	luad	2	4
ENSG00000140519	esophagus	coad	6	2	ENSG00000180210	liver	luad	2	4
ENSG00000170369	salivary.gland	coad	6	2	ENSG00000228278	liver	luad	2	4
ENSG00000205420	esophagus	coad	6	2	ENSG00000238269	testis	luad	2	4
ENSG00000115138	pituitary.gland	esca	6	2	ENSG00000091583	liver	paad	2	4
ENSG00000170231	small.intestine	esca	6	2	ENSG00000099937	liver	paad	2	4
ENSG00000170369	salivary.gland	esca	6	2	ENSG00000128040	testis	thca	2	4
ENSG00000170231	small.intestine	kirc	6	2	ENSG00000070915	kidney	ucec	2	4
ENSG00000170231	small.intestine	kirp	6	2	ENSG00000151655	liver	ucec	2	4
ENSG00000140519	esophagus	lihc	6	2	ENSG00000100593	placenta	blca	2	3
ENSG00000115138	pituitary.gland	luad	6	2	ENSG00000160182	stomach	blca	2	3
ENSG00000140519	esophagus	luad	6	2	ENSG00000074803	kidney	brca	2	3
ENSG00000170369	salivary.gland	luad	6	2	ENSG00000114113	small.intestine	brca	2	3
ENSG00000205420	esophagus	luad	6	2	ENSG00000115850	small.intestine	brca	2	3
ENSG00000140519	esophagus	paad	6	2	ENSG00000135346	placenta	brca	2	3
ENSG00000170231	small.intestine	paad	6	2	ENSG00000172232	bone.marrow	brca	2	3
ENSG00000205420	esophagus	paad	6	2	ENSG00000180176	adrenal.gland	brca	2	3
ENSG00000115138	pituitary.gland	thca	6	2	ENSG00000100593	placenta	coad	2	3
ENSG00000170369	salivary.gland	thca	6	2	ENSG00000180176	adrenal.gland	coad	2	3
ENSG00000205420	esophagus	thca	6	2	ENSG00000104537	small.intestine	esca	2	3

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data) (cont'd)

ENSG00000115138	pituitary.gland	ucec	6	2	ENSG00000121410	liver	esca	2	3		
ENSG00000140519	esophagus	ucec	6	2	ENSG00000163295	small.intestine	esca	2	3		
ENSG00000205420	esophagus	ucec	6	2	ENSG00000121410	liver	kirc	2	3		
ENSG00000170369	submandibular.gland	blca	6	1	ENSG00000143954	pancreas	kirc	2	3		
ENSG00000043355	brain	brca	6	1	ENSG00000169495	placenta	kirc	2	3		
ENSG00000170369	submandibular.gland	brca	6	1	ENSG00000229183	stomach	kirc	2	3		
ENSG00000043355	brain	coad	6	1	ENSG00000256713	stomach	kirc	2	3		
ENSG00000170369	submandibular.gland	coad	6	1	ENSG00000143954	pancreas	kirp	2	3		
ENSG00000170369	submandibular.gland	esca	6	1	ENSG00000169495	placenta	kirp	2	3		
ENSG00000043355	brain	kirc	6	1	ENSG00000178919	thyroid	kirp	2	3		
ENSG00000043355	brain	lihc	6	1	ENSG00000229183	stomach	kirp	2	3		
ENSG00000043355	brain	luad	6	1	ENSG00000256713	stomach	kirp	2	3		
ENSG00000170369	submandibular.gland	luad	6	1	ENSG00000070019	small.intestine	lihc	2	3		
ENSG00000170369	submandibular.gland	thca	6	1	ENSG00000074803	kidney	lihc	2	3		
ENSG00000043355	brain	ucec	6	1	ENSG00000114113	small.intestine	lihc	2	3		
ENSG00000138161	pancreas	blca	5	5	ENSG00000163295	small.intestine	lihc	2	3		
ENSG00000138161	pancreas	brca	5	5	ENSG00000166866	small.intestine	lihc	2	3		
ENSG00000138161	pancreas	kirc	5	5	ENSG00000070019	small.intestine	luad	2	3		
ENSG00000154040	testis	kirc	5	5	ENSG00000104537	small.intestine	luad	2	3		
ENSG00000138161	pancreas	kirp	5	5	ENSG00000115850	small.intestine	luad	2	3		
ENSG00000154040	testis	kirp	5	5	ENSG00000135346	placenta	luad	2	3		
ENSG00000154040	testis	lihc	5	5	ENSG00000160182	stomach	luad	2	3		
ENSG00000138161	pancreas	luad	5	5	ENSG00000166866	small.intestine	luad	2	3		

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data) (cont'd)

ENSG00000154040	testis	luad	5	5	ENSG00000172232	bone.marrow	luad	2	3
ENSG00000154040	testis	thca	5	5	ENSG00000178919	thyroid	luad	2	3
ENSG00000106927	liver	blca	5	4	ENSG00000189182	skin	blca	2	2
ENSG00000167749	prostate	blca	5	4	ENSG00000105141	skin	brca	2	2
ENSG00000106927	liver	brca	5	4	ENSG00000106689	brain	brca	2	2
ENSG00000118194	heart	brca	5	4	ENSG00000118113	bone.marrow	brca	2	2
ENSG00000167749	prostate	brca	5	4	ENSG00000135346	pituitary.gland	brca	2	2
ENSG00000171564	liver	brca	5	4	ENSG00000164047	bone.marrow	brca	2	2
ENSG00000106927	liver	coad	5	4	ENSG00000164076	cerebral.cortex	brca	2	2
ENSG00000171564	liver	coad	5	4	ENSG00000170367	salivary.gland	brca	2	2
ENSG00000167749	prostate	kirc	5	4	ENSG00000188293	esophagus	brca	2	2
ENSG00000171564	liver	kirc	5	4	ENSG00000189182	skin	brca	2	2
ENSG00000167749	prostate	kirp	5	4	ENSG00000164076	cerebral.cortex	coad	2	2
ENSG00000118194	heart	lihc	5	4	ENSG00000177238	skeletal.muscle	coad	2	2
ENSG00000167749	prostate	lihc	5	4	ENSG00000105141	skin	esca	2	2
ENSG00000106927	liver	luad	5	4	ENSG00000162344	gall.bladder	esca	2	2
ENSG00000118194	heart	luad	5	4	ENSG00000175121	skin	kirc	2	2
ENSG00000171564	liver	luad	5	4	ENSG00000175121	skin	kirp	2	2
ENSG00000118194	heart	paad	5	4	ENSG00000171431	small.intestine	lihc	2	2
ENSG00000171564	liver	paad	5	4	ENSG00000106689	brain	luad	2	2
ENSG00000106927	liver	ucec	5	4	ENSG00000118113	bone.marrow	luad	2	2
ENSG00000118194	heart	ucec	5	4	ENSG00000135346	pituitary.gland	luad	2	2
ENSG00000101470	skeletal.muscle	blca	5	2	ENSG00000162344	gall.bladder	luad	2	2
ENSG00000114854	heart	blca	5	2	ENSG00000171431	small.intestine	luad	2	2
ENSG00000183091	skeletal.muscle	blca	5	2	ENSG00000177238	skeletal.muscle	luad	2	2
ENSG00000196296	skeletal.muscle	blca	5	2	ENSG00000188293	esophagus	paad	2	2
ENSG00000171401	esophagus	brca	5	2	ENSG00000164047	bone.marrow	thca	2	2
ENSG00000101470	skeletal.muscle	coad	5	2	ENSG00000170367	salivary.gland	thca	2	2
ENSG00000183091	skeletal.muscle	coad	5	2	ENSG00000177791	skeletal.muscle	thca	2	2

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data) (cont'd)

ENSG00000114854	heart	esca	5	2	ENSG00000177791	skeletal.muscle	ucec	2	2
ENSG00000183091	skeletal.muscle	esca	5	2	ENSG00000070729	pineal.gland	blca	2	1
ENSG00000196296	skeletal.muscle	esca	5	2	ENSG00000104827	placenta	blca	2	1
ENSG00000086967	skeletal.muscle	kirc	5	2	ENSG00000123500	bone.marrow	blca	2	1
ENSG00000131095	brain	kirc	5	2	ENSG00000136688	tongue	blca	2	1
ENSG00000196296	skeletal.muscle	kirc	5	2	ENSG00000189052	placenta	blca	2	1
ENSG00000086967	skeletal.muscle	kirp	5	2	ENSG00000213030	placenta	blca	2	1
ENSG00000131095	brain	kirp	5	2	ENSG00000244094	tongue	blca	2	1
ENSG00000196296	skeletal.muscle	kirp	5	2	ENSG00000137869	ovary	brca	2	1
ENSG00000114854	heart	lihc	5	2	ENSG00000173404	brain	brca	2	1
ENSG00000131095	brain	lihc	5	2	ENSG00000185527	pineal.gland	brca	2	1
ENSG00000086967	skeletal.muscle	luad	5	2	ENSG00000096088	lung	coad	2	1
ENSG00000101470	skeletal.muscle	luad	5	2	ENSG00000123500	bone.marrow	coad	2	1
ENSG00000131095	brain	luad	5	2	ENSG00000134443	brain	coad	2	1
ENSG00000171401	esophagus	luad	5	2	ENSG00000091138	small.intestine	esca	2	1
ENSG00000183091	skeletal.muscle	luad	5	2	ENSG00000181617	minor.salivary.gland	kirc	2	1
ENSG00000196296	skeletal.muscle	luad	5	2	ENSG00000078898	minor.salivary.gland	lihc	2	1
ENSG00000101470	skeletal.muscle	paad	5	2	ENSG00000091138	small.intestine	lihc	2	1
ENSG00000114854	heart	paad	5	2	ENSG00000096088	lung	lihc	2	1
ENSG00000171401	esophagus	paad	5	2	ENSG00000137285	brain	lihc	2	1
ENSG00000086967	skeletal.muscle	thca	5	2	ENSG00000137869	ovary	lihc	2	1
ENSG00000101470	skeletal.muscle	thca	5	2	ENSG00000181617	minor.salivary.gland	lihc	2	1
ENSG00000131095	brain	thca	5	2	ENSG00000070729	pineal.gland	luad	2	1
ENSG00000171401	esophagus	thca	5	2	ENSG00000078898	minor.salivary.gland	luad	2	1
ENSG00000086967	skeletal.muscle	ucec	5	2	ENSG00000104827	placenta	luad	2	1

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data) (cont'd)

ENSG00000114854	heart	ucec	5	2	ENSG00000136688	tongue	luad	2	1
ENSG00000171401	esophagus	ucec	5	2	ENSG00000137285	brain	luad	2	1
ENSG00000183091	skeletal.muscle	ucec	5	2	ENSG00000173404	brain	luad	2	1
ENSG00000109205	minor.salivary.gla nd	blca	5	1	ENSG00000189052	placenta	luad	2	1
ENSG00000162078	minor.salivary.gla nd	blca	5	1	ENSG00000198074	small.intestine	luad	2	1
ENSG00000109205	minor.salivary.gla nd	brca	5	1	ENSG00000213030	placenta	luad	2	1
ENSG00000166670	endometrium	brca	5	1	ENSG00000244094	tongue	luad	2	1
ENSG00000109205	minor.salivary.gla nd	coad	5	1	ENSG00000134443	brain	ucec	2	1
ENSG00000166670	endometrium	coad	5	1	ENSG00000185527	pineal.gland	ucec	2	1
ENSG00000166670	endometrium	esca	5	1	ENSG00000198074	small.intestine	ucec	2	1
ENSG00000131095	cerebral.cortex	kirc	5	1	ENSG00000101435	testis	brca	1	5
ENSG00000131095	cerebral.cortex	kirp	5	1	ENSG00000104901	testis	brca	1	5
ENSG00000109205	minor.salivary.gla nd	lihc	5	1	ENSG00000173838	testis	brca	1	5
ENSG00000131095	cerebral.cortex	lihc	5	1	ENSG00000175877	testis	brca	1	5
ENSG00000162078	minor.salivary.gla nd	lihc	5	1	ENSG00000182459	testis	brca	1	5
ENSG00000166670	endometrium	lihc	5	1	ENSG00000147378	testis	kirc	1	5
ENSG00000131095	cerebral.cortex	luad	5	1	ENSG00000120440	testis	kirp	1	5
ENSG00000162078	minor.salivary.gla nd	luad	5	1	ENSG00000171790	testis	kirp	1	5
ENSG00000166670	endometrium	luad	5	1	ENSG00000163467	testis	lihc	1	5
ENSG00000162078	minor.salivary.gla nd	paad	5	1	ENSG00000170890	pancreas	lihc	1	5
ENSG00000131095	cerebral.cortex	thca	5	1	ENSG00000171804	testis	luad	1	5

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data) (cont'd)

ENSG00000109205	minor.salivary.gland	ucec	5	1	ENSG00000168484	lung	thca	1	5
ENSG00000162078	minor.salivary.gland	ucec	5	1	ENSG00000185955	testis	thca	1	5
ENSG00000122852	lung	blca	4	5	ENSG00000140505	liver	blca	1	4
ENSG00000133661	lung	blca	4	5	ENSG00000158874	liver	blca	1	4
ENSG00000122852	lung	brca	4	5	ENSG00000180432	liver	blca	1	4
ENSG00000133661	lung	brca	4	5	ENSG00000109181	liver	brca	1	4
ENSG00000169347	pancreas	brca	4	5	ENSG00000142698	testis	brca	1	4
ENSG00000198033	testis	brca	4	5	ENSG00000147465	adrenal.gland	brca	1	4
ENSG00000169347	pancreas	coad	4	5	ENSG00000169344	kidney	brca	1	4
ENSG00000198033	testis	kirc	4	5	ENSG00000175336	liver	brca	1	4
ENSG00000133661	lung	lihc	4	5	ENSG00000180043	testis	brca	1	4
ENSG00000169347	pancreas	lihc	4	5	ENSG00000196553	testis	brca	1	4
ENSG00000198033	testis	lihc	4	5	ENSG00000227234	testis	brca	1	4
ENSG00000169347	pancreas	luad	4	5	ENSG00000234068	testis	brca	1	4
ENSG00000198033	testis	luad	4	5	ENSG00000243543	testis	brca	1	4
ENSG00000122852	lung	paad	4	5	ENSG00000120054	liver	coad	1	4
ENSG00000122852	lung	thca	4	5	ENSG00000123561	liver	coad	1	4
ENSG00000133661	lung	ucec	4	5	ENSG00000140093	liver	coad	1	4
ENSG00000139547	liver	blca	4	4	ENSG00000186007	testis	coad	1	4
ENSG00000171557	liver	blca	4	4	ENSG00000055957	liver	kirc	1	4
ENSG00000185303	lung	blca	4	4	ENSG00000113600	liver	kirc	1	4
ENSG00000106327	liver	brca	4	4	ENSG00000132693	liver	kirc	1	4
ENSG00000139547	liver	brca	4	4	ENSG00000153002	pancreas	kirc	1	4
ENSG00000171557	liver	brca	4	4	ENSG00000203859	adrenal.gland	kirc	1	4
ENSG00000185303	lung	brca	4	4	ENSG00000213512	liver	kirc	1	4
ENSG00000198054	testis	brca	4	4	ENSG00000103375	pancreas	lihc	1	4
ENSG00000125207	testis	coad	4	4	ENSG00000132972	testis	lihc	1	4

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data) (cont'd)

ENSG00000125207	testis	esca	4	4	ENSG00000148795	adrenal.gland	lihc	1	4
ENSG00000139547	liver	esca	4	4	ENSG00000234511	testis	lihc	1	4
ENSG00000106327	liver	kirc	4	4	ENSG00000072080	liver	luad	1	4
ENSG00000115386	pancreas	kirc	4	4	ENSG00000116785	liver	luad	1	4
ENSG00000171557	liver	kirc	4	4	ENSG00000151704	kidney	luad	1	4
ENSG00000231852	adrenal.gland	kirc	4	4	ENSG00000157131	liver	luad	1	4
ENSG00000106327	liver	kirp	4	4	ENSG00000164744	testis	luad	1	4
ENSG00000115386	pancreas	kirp	4	4	ENSG00000174015	testis	luad	1	4
ENSG00000171557	liver	kirp	4	4	ENSG00000197768	testis	luad	1	4
ENSG00000231852	adrenal.gland	kirp	4	4	ENSG00000243480	pancreas	luad	1	4
ENSG00000115386	pancreas	lihc	4	4	ENSG00000255974	liver	luad	1	4
ENSG00000198054	testis	lihc	4	4	ENSG00000261701	liver	luad	1	4
ENSG00000106327	liver	luad	4	4	ENSG00000131187	liver	thca	1	4
ENSG00000115386	pancreas	luad	4	4	ENSG00000244414	liver	thca	1	4
ENSG00000125207	testis	luad	4	4	ENSG00000137090	testis	ucec	1	4
ENSG00000139547	liver	luad	4	4	ENSG00000186910	liver	ucec	1	4
ENSG00000198054	testis	luad	4	4	ENSG00000047936	lung	blca	1	3
ENSG00000231852	adrenal.gland	luad	4	4	ENSG00000165685	kidney	blca	1	3
ENSG00000125207	testis	paad	4	4	ENSG00000165887	heart	blca	1	3
ENSG00000185303	lung	paad	4	4	ENSG00000203857	placenta	blca	1	3
ENSG00000185303	lung	thca	4	4	ENSG00000204941	placenta	blca	1	3
ENSG00000231852	adrenal.gland	thca	4	4	ENSG00000137270	placenta	brca	1	3
ENSG00000198054	testis	ucec	4	4	ENSG00000174885	small.intestine	brca	1	3
ENSG00000042832	thyroid	blca	4	3	ENSG00000134812	stomach	coad	1	3
ENSG00000143839	kidney	blca	4	3	ENSG00000145384	small.intestine	esca	1	3
ENSG00000160181	stomach	blca	4	3	ENSG00000167580	kidney	esca	1	3
ENSG00000163283	placenta	blca	4	3	ENSG00000165409	thyroid	kirc	1	3
ENSG00000042832	thyroid	brca	4	3	ENSG00000169876	small.intestine	kirc	1	3
ENSG00000160181	stomach	brca	4	3	ENSG00000229859	stomach	kirc	1	3

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data) (cont'd)

ENSG00000186652	placenta	brca	4	3	ENSG00000068985	testis	lihc	1	3
ENSG00000042832	thyroid	coad	4	3	ENSG00000125999	lung	lihc	1	3
ENSG00000143839	kidney	coad	4	3	ENSG00000131142	small.intestine	lihc	1	3
ENSG00000160181	stomach	coad	4	3	ENSG00000137869	placenta	lihc	1	3
ENSG00000163283	placenta	coad	4	3	ENSG00000168903	small.intestine	lihc	1	3
ENSG00000042832	thyroid	esca	4	3	ENSG00000204511	kidney	lihc	1	3
ENSG00000079112	small.intestine	esca	4	3	ENSG00000111701	small.intestine	luad	1	3
ENSG00000163283	placenta	esca	4	3	ENSG00000120211	placenta	luad	1	3
ENSG00000079112	small.intestine	kirc	4	3	ENSG00000130528	heart	luad	1	3
ENSG00000186652	placenta	kirc	4	3	ENSG00000146678	liver	luad	1	3
ENSG00000079112	small.intestine	kirp	4	3	ENSG00000110244	small.intestine	paad	1	3
ENSG00000143839	kidney	lihc	4	3	ENSG00000159251	heart	paad	1	3
ENSG00000186652	placenta	lihc	4	3	ENSG00000231924	placenta	thca	1	3
ENSG00000079112	small.intestine	luad	4	3	ENSG00000159763	salivary.gland	blca	1	2
ENSG00000160181	stomach	luad	4	3	ENSG00000101425	bone.marrow	brca	1	2
ENSG00000186652	placenta	luad	4	3	ENSG00000130287	brain	brca	1	2
ENSG00000163283	placenta	paad	4	3	ENSG00000167419	salivary.gland	brca	1	2
ENSG00000143839	kidney	thca	4	3	ENSG00000167768	skin	brca	1	2
ENSG00000169385	bone.marrow	blca	4	2	ENSG00000169906	duodenum	brca	1	2
ENSG00000198125	heart	blca	4	2	ENSG00000177108	cerebral.cortex	brca	1	2
ENSG00000125780	esophagus	brca	4	2	ENSG00000183347	esophagus	brca	1	2
ENSG00000163209	esophagus	brca	4	2	ENSG00000187533	salivary.gland	brca	1	2
ENSG00000169385	bone.marrow	brca	4	2	ENSG00000188508	skin	brca	1	2
ENSG00000169474	esophagus	brca	4	2	ENSG00000196361	brain	brca	1	2
ENSG00000170465	esophagus	brca	4	2	ENSG00000196415	bone.marrow	brca	1	2
ENSG00000125780	esophagus	coad	4	2	ENSG00000204655	brain	brca	1	2
ENSG00000169474	esophagus	coad	4	2	ENSG00000130595	skeletal.muscle	kirc	1	2
ENSG00000169385	bone.marrow	kirc	4	2	ENSG00000141086	pancreas	kirc	1	2
ENSG00000204539	skin	kirc	4	2	ENSG00000185028	heart	kirc	1	2

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data) (cont'd)

ENSG00000204539	skin	kirp	4	2	ENSG00000223609	bone.marrow	kirc	1	2
ENSG00000125780	esophagus	lihc	4	2	ENSG00000239839	bone.marrow	kirc	1	2
ENSG00000198125	heart	lihc	4	2	ENSG00000090932	brain	kirp	1	2
ENSG00000163209	esophagus	luad	4	2	ENSG00000163554	bone.marrow	lihc	1	2
ENSG00000169385	bone.marrow	luad	4	2	ENSG00000172016	pancreas	lihc	1	2
ENSG00000169474	esophagus	luad	4	2	ENSG00000104879	heart	luad	1	2
ENSG00000170465	esophagus	luad	4	2	ENSG00000113889	kidney	luad	1	2
ENSG00000198125	heart	luad	4	2	ENSG00000125571	skin	luad	1	2
ENSG00000163209	esophagus	paad	4	2	ENSG00000131050	salivary.gland	luad	1	2
ENSG00000169474	esophagus	paad	4	2	ENSG00000196091	skeletal.muscle	luad	1	2
ENSG00000170465	esophagus	paad	4	2	ENSG00000197561	bone.marrow	luad	1	2
ENSG00000198125	heart	paad	4	2	ENSG00000214711	esophagus	luad	1	2
ENSG00000204539	skin	paad	4	2	ENSG00000105675	stomach	thca	1	2
ENSG00000125780	esophagus	ucec	4	2	ENSG00000124467	placenta	thca	1	2
ENSG00000163209	esophagus	ucec	4	2	ENSG00000171201	salivary.gland	thca	1	2
ENSG00000170465	esophagus	ucec	4	2	ENSG00000204544	esophagus	thca	1	2
ENSG00000204539	skin	ucec	4	2	ENSG00000197893	heart	ucec	1	2
ENSG00000160181	gall.bladder	blca	4	1	ENSG00000159763	lung	blca	1	1
ENSG00000125780	tongue	brca	4	1	ENSG00000159763	minor.salivary.gland	blca	1	1
ENSG00000160181	gall.bladder	brca	4	1	ENSG00000139352	brain	brca	1	1
ENSG00000125780	tongue	coad	4	1	ENSG00000169344	cortex.of.kidney	brca	1	1
ENSG00000160181	gall.bladder	coad	4	1	ENSG00000169906	small.intestine	brca	1	1
ENSG00000100448	bone.marrow	kirc	4	1	ENSG00000177108	brain	brca	1	1
ENSG00000100448	bone.marrow	kirp	4	1	ENSG00000183206	testis	brca	1	1
ENSG00000100448	bone.marrow	lihc	4	1	ENSG00000186297	cerebral.cortex	brca	1	1
ENSG00000125780	tongue	lihc	4	1	ENSG00000189182	zone.of.skin	brca	1	1
ENSG00000160181	gall.bladder	luad	4	1	ENSG00000198183	lung	brca	1	1

Table B.3 Selected genes passing all criteria (after analysis of tissue-specific genes and TCGA cancer expression data) (cont'd)

ENSG00000100448	bone.marrow	paad	4	1	ENSG00000204655	cerebral.cortex	brca	1	1
ENSG00000125780	tongue	ucec	4	1	ENSG00000066405	lung	coad	1	1
ENSG00000099399	testis	blca	3	5	ENSG00000169474	brain	coad	1	1
ENSG00000143450	testis	blca	3	5	ENSG00000182968	brain	coad	1	1
ENSG00000046774	testis	brca	3	5	ENSG00000183145	brain	coad	1	1
ENSG00000075886	testis	brca	3	5	ENSG00000102195	pituitary.gland	esca	1	1
ENSG00000137948	testis	brca	3	5	ENSG00000184502	stomach	esca	1	1
ENSG00000143452	testis	brca	3	5	ENSG00000102001	pineal.gland	kirc	1	1
ENSG00000166796	testis	brca	3	5	ENSG00000138039	ovary	kirc	1	1
ENSG00000196260	lung	coad	3	5	ENSG00000169876	small.intestine. Peyers.patch	kirc	1	1
ENSG00000099399	testis	esca	3	5	ENSG00000205927	brain	kirc	1	1
ENSG00000137948	testis	esca	3	5	ENSG00000206047	bone.marrow	kirc	1	1
ENSG00000046774	testis	lihc	3	5	ENSG00000240247	bone.marrow	kirc	1	1
ENSG00000075886	testis	lihc	3	5	ENSG00000158578	whole.blood	kirp	1	1
ENSG00000099399	testis	lihc	3	5	ENSG00000206172	whole.blood	kirp	1	1
ENSG00000143450	testis	lihc	3	5	ENSG00000056291	placenta	lihc	1	1
ENSG00000166796	testis	lihc	3	5	ENSG00000106128	pituitary.gland	lihc	1	1
ENSG00000046774	testis	luad	3	5	ENSG00000107187	pituitary.gland	lihc	1	1
					ENSG00000168757	testis	lihc	1	1
					ENSG00000170289	pineal.gland	luad	1	1
					ENSG00000184486	brain	luad	1	1
					ENSG00000138083	brain	ucec	1	1

Six different selected genes passing all criteria after integration of tissue-specific genes and TCGA cancer expression data presented in Figure 7.17 are explained in Table B.4 for their related GO terms to understand their function and association with intra-tumoral heterogeneity and cancer cell mechanisms.

Table B.4 Related GO terms in biological processes for six selected genes

Ensembl Gene ID	Gene Name	tissue	cancer count	consensus	Non specific cancer	Functional annotation (GO term /Biological Processes)
ENSG00000151224	methionine adenosyltransferase 1A(MAT1A)	liver	8	4	brca	sulfur amino acid , selenium, one-carbon metabolic process; methionine catabolic process; methylation
ENSG00000151224	methionine adenosyltransferase 1A(MAT1A)	liver	8	4	coad	sulfur amino acid , selenium, one-carbon metabolic process; methionine catabolic process; methylation
ENSG00000151224	methionine adenosyltransferase 1A(MAT1A)	liver	8	4	esca	sulfur amino acid , selenium, one-carbon metabolic process; methionine catabolic process; methylation
ENSG00000151224	methionine adenosyltransferase 1A(MAT1A)	liver	8	4	luad	sulfur amino acid , selenium, one-carbon metabolic process; methionine catabolic process; methylation
ENSG00000151224	methionine adenosyltransferase 1A(MAT1A)	liver	8	4	paad	sulfur amino acid , selenium, one-carbon metabolic process; methionine catabolic process; methylation
ENSG00000151224	methionine adenosyltransferase 1A(MAT1A)	liver	8	4	thca	sulfur amino acid , selenium, one-carbon metabolic process; methionine catabolic process; methylation
ENSG00000151224	methionine adenosyltransferase 1A(MAT1A)	liver	8	4	ucec	sulfur amino acid , selenium, one-carbon metabolic process; methionine catabolic process; methylation
ENSG00000151224	methionine adenosyltransferase 1A(MAT1A)	liver	8	4	blca	sulfur amino acid , selenium, one-carbon metabolic process; methionine catabolic process; methylation
ENSG00000129991	troponin I3, cardiac type(TNNI3)	heart	8	4	brca	vasculogenesis; skeletal and smooth muscle contraction; cellular calcium ion homeostasis; heart development; negative regulation of ATPase activity

Table B.4 Related GO terms in biological processes for six selected genes (cont'd)

ENSG00000129991	troponin I3, cardiac type(TNNI3)	heart	8	4	coad		vasculogenesis; skeletal and smooth muscle contraction; cellular calcium ion homeostasis; heart development; negative regulation of ATPase activity
ENSG00000129991	troponin I3, cardiac type(TNNI3)	heart	8	4	kirp		vasculogenesis; skeletal and smooth muscle contraction; cellular calcium ion homeostasis; heart development; negative regulation of ATPase activity
ENSG00000129991	troponin I3, cardiac type(TNNI3)	heart	8	4	lihc		vasculogenesis; skeletal and smooth muscle contraction; cellular calcium ion homeostasis; heart development; negative regulation of ATPase activity
ENSG00000129991	troponin I3, cardiac type(TNNI3)	heart	8	4	luad		vasculogenesis; skeletal and smooth muscle contraction; cellular calcium ion homeostasis; heart development; negative regulation of ATPase activity
ENSG00000129991	troponin I3, cardiac type(TNNI3)	heart	8	4	paad		vasculogenesis; skeletal and smooth muscle contraction; cellular calcium ion homeostasis; heart development; negative regulation of ATPase activity
ENSG00000129991	troponin I3, cardiac type(TNNI3)	heart	8	4	ucec		vasculogenesis; skeletal and smooth muscle contraction; cellular calcium ion homeostasis; heart development; negative regulation of ATPase activity
ENSG00000129991	troponin I3, cardiac type(TNNI3)	heart	8	4	blca		vasculogenesis; skeletal and smooth muscle contraction; cellular calcium ion homeostasis; heart development; negative regulation of ATPase activity

Table B.4 Related GO terms in biological processes for six selected genes (cont'd)

ENSG00000133661	surfactant protein D(SFTPD)	lung	4	5	blca	regulation of cytokine production; respiratory gaseous exchange; negative regulation of T cell proliferation
ENSG00000133661	surfactant protein D(SFTPD)	lung	4	5	brca	regulation of cytokine production; respiratory gaseous exchange; negative regulation of T cell proliferation
ENSG00000133661	surfactant protein D(SFTPD)	lung	4	5	lihc	regulation of cytokine production; respiratory gaseous exchange; negative regulation of T cell proliferation
ENSG00000133661	surfactant protein D(SFTPD)	lung	4	5	ucec	regulation of cytokine production; respiratory gaseous exchange; negative regulation of T cell proliferation
ENSG00000170465	keratin 6C(KRT6C)	esophagus	4	2	brca	intermediate filament cytoskeleton organization
ENSG00000170465	keratin 6C(KRT6C)	esophagus	4	2	luad	intermediate filament cytoskeleton organization
ENSG00000170465	keratin 6C(KRT6C)	esophagus	4	2	paad	intermediate filament cytoskeleton organization
ENSG00000170465	keratin 6C(KRT6C)	esophagus	4	2	ucec	intermediate filament cytoskeleton organization
ENSG00000147378	fetal and adult testis expressed 1(FATE1)	testis	1	5	kirc	integral component of membrane
ENSG00000196361	ELAV like RNA binding protein 3(ELAVL3)	brain	1	2	brca	nervous system development; cell differentiation

CURRICULUM VITAE

PERSONAL INFORMATION

Name Surname : Hatice Büşra KONUK

Date of birth and place : 14.09.1993 / Üsküdar

Foreign Languages : English

E-mail : hbsrknk@gmail.com/ hbkonuk@gtu.edu.tr

EDUCATION

Degree	Department	University	Date of Graduation
Undergraduate	Bioengineering Department	Yıldız Technical University	2015
High School	-	Nevzat Ayaz Anatolian High School	2011

WORK EXPERIENCE

Year	Corporation/Institute	Enrollment
2016- Present	Gebze Technical University	Research Assistant

PUBLISHMENTS

Conference Papers

1. HB Konuk, A Yılmaz, Rigorous Identification of Tissue Specific Genes in silico: Revealing the Interplay between Cancer Specific Expression and Tissue Specific Expression, International University of Sarajevo, International Congress on Advances in Bioscience and Biotechnology, 25-29 October, 2017, *Poster Presentation*.

2. HB Konuk, MR Cesur, A Yılmaz, Identification of Tissue Specific Genes and Assessment of Their Intersection with Differentialy Expressed Genes in Cancer Data to Understand Tumor Heterogeneity in silico, METU Northern Cyprus Campus, 10 th The International Symposium on Health Informatics and Bioinformatics (HIBIT), 28-30 June, 2017, *Poster Presentation*.

3. HB Konuk, MR Cesur, A Yılmaz, What is the relationship between cancer DEGs and tissue specific genes?, Gebze Teknik Üniversitesi, Lisansüstü Araştırmalar Sempozyumu ve Tanıtım Günleri, 17-18 Mayıs,2017, *Oral Presentation*.

4 HB Konuk, A Yılmaz, Robust and Rigorous Classification of Tissue Specific Genes, Ahi Evran Üniversitesi, International DNA Day and Genome Congress (IDDGC17), 24-28 Nisan,2017, *Oral Presentation*.

AWARDS

1. Hatice Büşra Konuk, Rigorous Identification of Tissue Specific Genes in silico: Revealing the Interplay between Cancer Specific Expression and Tissue Specific Expression, The Best Visual presentations, International Congress on Advances in Bioscience and Biotechnology (ICABB), Sarajevo, 2017.

2. Genç Araştırmacı Ödülü Kapsamında Kayıt Bursu, MOKAD Burs Komitesi, 6. Multidisipliner Kanser Araştırma Kongresi, 27-30 Ekim 2016, Konya, Türkiye.