ANALYSIS OF 3' UTR SHORTENING EVENTS IN BREAST CANCER

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ONUR BALOĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
MASTER OF SCIENCE
IN
BIOINFORMATICS

JANUARY 2013

ANALYSIS OF 3' UTR SHORTENING EVENTS IN BREAST CANCER

Submitted by **Onur BALOĞLU** in partial fulfillment of the requirements for the degree of **Master of Science, Bioinformatics Program**,
Middle East Technical University by,

Prof. Dr. Nazife Baykal _____
Director, Informatics Institute

Assist. Prof. Dr. Yeşim Aydın Son _____
Head of department, Medical Informatics, METU

Assoc. Prof. Dr. Tolga Can _____
Supervisor, Computer Engineering, METU

Examining Committee Members

Assoc. Prof. Dr. Mesut Muyan _____
METU, BIO

Assoc. Prof. Dr. Tolga Can _____
METU, CENG

Assist. Prof. Dr. Aybar Can Acar _____
METU, MIN

Assist. Prof. Dr. Ayşe Elif Erson Bensan _____
METU, BIO

Assist. Prof. Dr. Yeşim Aydın Son _____
METU, MIN

Date: 29.01.2013

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : **Onur Baloğlu**

Signature :

ABSTRACT

ANALYSIS OF 3' UTR SHORTENING EVENTS

IN BREAST CANCER

Baloğlu, Onur

M. Sc., Bioinformatics Program

Supervisor: Assoc. Prof. Dr. Tolga Can

January 2013, 42 pages

Cancer is the collective term used to describe a diverse group of diseases that share certain hallmarks, which in turn enables the affected cells to sustain an uncontrolled cell growth. Despite the increasing efforts and advances in cancer therapies, cancers are still responsible for approximately 10% of all the deaths worldwide. Furthermore, the increase in the average human lifespan will further contribute to the cancer incidences. This brings the necessity to focus our efforts on early detection and effective diagnosis methods. With the advances in high-throughput genomics technologies, gene expression signatures have gained attention as a novel method in cancer diagnostics. These signatures are identified by simply comparing the expression levels of genes in tumor and control samples. Here, we propose an alternative method based on the probe expression level measurement of 3'UTR of candidate genes. We chose breast cancer as a model and performed an in silico analysis on publicly available gene expression datasets of Affymetrix chips to analyse 3'UTR shortening during breast cancer situation. Overall, our analysis suggests that shortening of 3'UTR is a significant mechanism observed in breast cancer .

Keywords: *Microarray, 3'UTR, Alternative Polyadenylation ,Differential expression, breast cancer.*

ÖZ

MEME KANSERİNDE 3`UTR KISALMA

OLAYLARININ İNCELENMESİ

Baloğlu, Onur

Yüksek Lisans, Biyoenformatik

Tez Danışmanı: Assoc. Prof. Dr. Tolga Can

Ocak 2013, 42 sayfa

Kanser etkilenmiş olan hücrelerin kontrolsüz bir biçimde hücre bölünmesine neden olan belirli özellikler taşıyan bir grup hastalığı tanımlamak için kullanılan yaygın bir terimdir. Kanser terapisindeki gelişmelere rağmen hala günümüzde dünya çapındaki ölümlerin %10'u kanser nedenlidir. Ayrıca uzayan insan ömrü de kanser olaylarını arttırmıştır. Bu durum odağı erken tanıya ve ektili tetkiklere yöneltmiştir. Yüksek iş hacimli genomik teknolojilerindeki gelişmeler sayesinde, gen ekspresyonu profili kanser teşhisi için özgün bir yöntem olarak dikkati kazanmıştır. Bu yöntem basitçe tümörlü ve sağlıklı örneklerdeki gen expresiyonu seviyelerinin profillerinin karşılaştırılmasına dayanmaktadır. Biz ise bu teşhis tekniğine alternatif olabilecek bir teknik geliştirmek istemekteyiz. Bu noktada biz aday genlerin 3'UTR bölgesinin probe ekspresyonu ölçülmesini baz alan alternatif bir yöntem öneriyoruz Kanser olaylarında ki 3'UTR kısalmasını analiz etmek için model olarak oladığımız meme kanserine ait olan ve herkese açık durumdaki Affymetrix çiplerinin gen ekspresyon datasetleri ile sanal ortamda analizler yapıldı. Tümden düşünüldüğünde analizlerimiz 3'UTR kısalmasının meme kanserinde dikkate değer bir mekanizmasını ortaya koymuştur.

Anahtar kelimeler: *Microarray, 3'UTR , Alternative Polyadenylation, Differential expression, meme kanseri*

To my family

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude and sincerest appreciation to my supervisor Assoc. Prof. Dr. Tolga Can for his supervision, continuous advice, invaluable help and guidance throughout this research. He has supported me before the beginning of the programme with his knowledge and patience.

I am grateful to Assoc. Prof. Dr. Ayşe Elif Erson Bensan and Assoc. Prof. Dr. Mesut Muyan for their help and knowledge about cancer. They also guided me for writing this thesis with their knowledge and opinions.

I would like to thank to Damla Arslantunalı , H. Alper Döm, Erdoğan Pekcan Erkan, Alper Mutlu, Taner Tuncer and Begüm Akman Tuncer for their help and great friendship that made it easier to overcome difficulties.

I am also grateful to my supporting family. I thank my parents for devoting their life on me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Cancer is the name for diseases in which cells become abnormal and divide without control. According to several studies; in 2008, cancer accounted for 7.6 million deaths (around 13% of all deaths) worldwide ("WHO | World Health Organization," n.d.). Breast cancer is one of the top three cancer types for estimated new cases and which is the most important cause of cancer death in women in developing countries after lung cancer. (Herman, 1996; Howe et al., 2001; Kliewer et al., 1995; Parkin, Pisani, & Ferlay, 1993; Ziegler et al., 1993).

Like some diseases cancer can be the result of the some genetic disorders and some genes may be more effective in leading to cancer compared to the others. Inherited breast cancer is the case of about 5 to 10% of the whole breast cancer cases (Campeau, Foulkes, & Tischkowitz, 2008) and among these cases, 20 to 30% are caused by mutations in *BRCA1* and *BRCA2* genes which are responsible for transcriptional regulation and DNA repair mechanisms. *BRCA1* gene regulates the expression of some important genes in breast cancer such as *MYC, STAT1 , JAK1, laminin 3A and cyclin D1* (Dixelius et al., 2002).

Differential expression can be used for diagnosing cancer. This technique uses expression levels of genes before and after the questioned condition. Basically, experiments involve two types of samples which contain different cells: one from the control and one from the treated sample. Then, gene expression measurements are applied to find genes which are differentially expressed between the two samples. To understand the difference between samples for diagnosis or treatment, we need to

find up- and down- regulated genes between the control and test groups. The common practice is to process microarray data and perform some statistical tests by selecting a model which best suits the data.

An mRNA contains few structural elements. One of them is the three prime untranslated region which also named in short as 3'UTR.

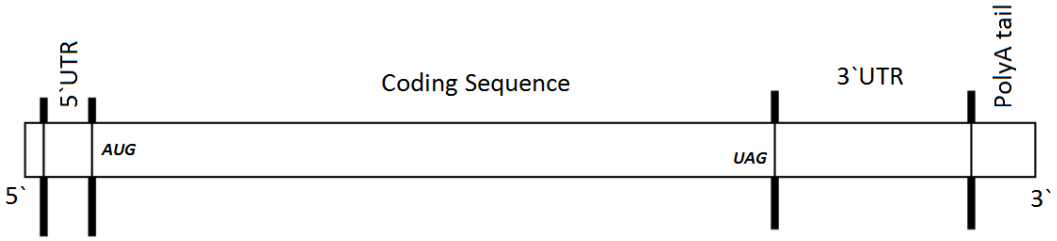

Figure 1: The structure of a typical human protein coding mRNA including the untranslated regions (UTRs) ("mRNA UTR Structure Exon Intron Cap - Molecular Biology Photo Gallery," n.d.).

As seen from Figure 1 above, the 3'UTRregion is just downstream of the stop codon and ends with a poly-A tail. In addition, there are several regulatory sequences at this region which include a polyadenylation signal, protein binding sites and miRNA binding sites. These have important roles for mRNA stability, localization, and translation (Zlotorynski & Agami, 2008).

In living organisms, genetic alterations may initiate cancer cells by activating proto-oncogenes. In cancer cells, oncogenes can be activated by widespread shortening of 3'UTRs which results from alternative cleavage and poyladenylation (APA) (Mayr & Bartel, 2009).

In this thesis, we analyze the differential shortening of 3'UTR via alternative polyadenylation in breast cancer cells as an alternative to the traditional differential expression analysis of cancer. Also, we try to assess whether differential shortening of the 3'UTR is observed between different breast cancer subtypes.

As described above, breast cancer is one of the most dangerous cancer types and it can be caused by genetic disorders in both inherited and environmentally caused

cases. Therefore, we choose this type of cancer as a research area and our results may be applied to other types of cancer. The problem we attack is the identification of genes that have significant differences between control and treated samples in terms of their 3'UTR length. To solve this problem, we have developed methods to analyze Affymetrix chips at the probe level to identify the genes in which there is significant difference between the control and cancer samples' short 3'UTR expression and the long 3'UTR expression. For validation of the proposed approach, we have chosen one dataset, GSE7904 (Richardson et al., 2006) for primary analysis. Then, we tested our method on other datasets: GSE3744 (Alimonti et al., 2010) and GSE20711 (Dedeurwaerder et al., 2011) and assessed the consistency of our observations. The proposed method is limited to Affymetrix U133A and Affymetrix U133 Plus 2.0 chips. This constraint limits us to work on datasets that are produced on these two platforms. However, both platforms are the most popular platforms in NCBI GEO and about 25% of all the samples in GEO use these platforms. Our reference set has enough number of samples that can be used in cross-experiment analysis.

Most gene expression experiments in NCBI GEO either contain small number of samples or they do not have controls. This also limits the number of arrays that can be used for validation. The proposed method makes it possible to handle data and analyze them in some basic steps and give the results which can be easily read.

1.2 Background

1.2.1 Cancer background

Cancer is the result of the abnormal and uncontrolled cell division. Cancer cells are able to invade other tissues and can spread to other parts of the body via blood and lymph systems. There are more than 100 different types of cancer which are mostly named for the cell type or the organ in which they start. However, there are more than hundred types of cancers which begin in cells; therefore, we need to understand

this process better ("Comprehensive Cancer Information - National Cancer Institute," n.d.).

The human body has many cell types which grow and divide in a controlled way to produce more cells. This process happens when cells become old or damaged and need to produce new cells. However, in some cases this ancient process goes wrong. DNA is the genetic material of a cell which can become damaged or undergoes mutations that affect normal cell procedures during cell growth and division. When this situation occurs, cells do not die and new cells form which are not needed. These extra cells may form a tissue mass called a tumor ("Comprehensive Cancer Information - National Cancer Institute," n.d.).

These tumors can be divided in two groups;

- Benign tumors: these are not cancerous. Benign tumor cells do not spread to the other tissues. This type of tumors can often be removed and do not recur most of the time.
- Malignant tumors: these are cancerous. Malignant tumor cells spread to other tissues.

Also some cancer types, such as leukemia, do not form tumors ("Comprehensive Cancer Information - National Cancer Institute," n.d.).

1.2.2 Breast Cancer

After the lung cancer, the most important cause of cancer death in women is breast cancer in developing countries (Herman, 1996; Howe et al., 2001; Kliewer et al., 1995; Parkin et al., 1993; Ziegler et al., 1993).

There are some known risk factors for breast cancer such as genetic and familial causes as well as hormonal, lifestyle and environmental factors. Some other factors which increase the risk of breast cancer are; height among postmenopausal women (Van den Brandt, 2000) mammographically dense breasts, menopause age (less than 45 and more then 54), post menopause hormone, oral contraceptive, and alcohol use,

radiation exposure, menarche age (less than 12 and more than 14), high endogenous estrogen, prolactin and premenopausal insulin like growth factor levels. Factors that decrease the risk of breast cancer are physical activitiy, breast feeding and non-steroidal anti-inflammatory drug usage (Hankinson, Colditz, & Willett, 2004). But there are still lots of unknown factors which have effects over breast tumorigenesis.

Inherited breast cancer is the case in about 5% to 10% of the whole breast cancer cases (Phelan et al., 1996) and among these cases, 20% to 30% are caused by mutations in *BRCA1* and *BRCA2* genes which are responsible for transcriptional regulation and DNA repair mechanisms. *BRCA1* gene regulate the expression of some important genes in breast cancer such as *MYC, STAT1 , JAK1, laminin 3A and cyclin D1* (Dixelius et al., 2002).

1.2.3 Breast Cancer Subtypes

496 genes were identified by using 40 breast cancer patients and named as "intrinsic gene set" and used for subtype identification. This is done by searching genes with little variance within repeated tumor samples and high variance across different tumors (Charles M Perou & Børresen-Dale, 2011). By using this gene set, four tumor subtypes with a normal breast-like group were identified those are LumA, LumB, claudin-low, and HER2 subtypes. Those subtypes are called intrinsic subtypes because their marker genes had intrinsic properties (Abd El-Rehim et al., 2004; Carey et al., 2006; Hu et al., 2006; Parker et al., 2009; Sorlie et al., 2003; Sotiriou et al., 2003).

Luminal Subtypes

The most common breast cancers are ER-positive tumors and which fall into luminal subtypes. So called because they have a gene expression pattern reminiscent of the luminal epithelial component of the breast (C M Perou et al., 2000). There are two subtypes named as luminal A and luminal B. There are many relevant differences between these but it is not easy to distinguish a luminal A from luminal B, since the expression of the genes defining these groups are a continuum. Generally luminal A tumors have high expression of ER but luminal B tumors have low expression of ER

(Hu et al., 2006; Sorlie et al., 2003). In population based studies, the lum-A subtype is the most common breast cancer type. Approximately 40% of all breast tumors are lum-A type tumors and 10% are lum-B type tumors. (Carey et al., 2006; Millikan et al., 2008; Morris et al., 2007).

Luminal tumors in general are defined by a quartet of transcription factors that includes ER,GATA3, FOXA1, and XBP1 (Asselin-Labat et al., 2007; Carroll et al., 2005; Kouros-Mehr, Slorach, Sternlicht, & Werb, 2006; Usary et al., 2004).

HER2-enriched Subtypes

10% of all breast cancers (Carey et al., 2006) are the HER2-enriched tumors. This subtype shows high expression of HER2 and GRB7 genes (D. J. Slamon et al., n.d.; D. Slamon et al., 1989).

Basal-like Subtype

This subtype is also known as "triple-negative" tumors (Schneider et al., 2008), due to their IHC (immunohistochemical) pattern of being negative for ER, PR, and HER2, although this is not a definitive classification ~25% of basal-like tumors are not triple negative. The characteristic properties of basal-like subtype tumors are; low luminal genes expression, low HER2 gene cluster expression, high proliferation cluster expression, and high basal cluster genes expression.

There is a link between basal-like breast cancer and BRCA1 mutation, over 80% of the BRCA1 mutations result as basal-like subtype.

Claudin-low Subtype

Characteristic features of claudin-low subtype are the low expression of genes involved in tight junctions and cell-cell adhesion including claudin 3, 4, 7, Occludin, and E-cadherin and high expression of Vimentin, Snail1, Snail2, and Twist1. This lack of epithelial cell features and expression of mesenchymal trait is reminiscent of

features associated with stem cells (Lim et al., 2009). Claudin-low tumors are enriched for TIC (Tumor Initiating Cells) features including high ALDH1 (Creighton et al., 2009).

1.2.4 Alternative Polyadenylation (APA)

Polyadenylation is the addition of a poly(A) tail to a RNA molecule. This structure is important for the nuclear export, translation, and stability of mRNA. For 3'UTR shortening, alternative polyadenylation (APA) is an important process which is emerging as a widespread mechanism used to control gene expression. This mechanism allows multiple mRNA transcripts by a single gene. Also, in some cases which is important for this thesis, this mechanism changes the mRNA coding potential or not the code but the 3'UTR length. The change of 3'UTR length effect the availability of RNA binding protein sites and miRNA binding sites (Di Giammartino, Nishida, & Manley, 2011).

It has become evident that APA is extensively used to regulate gene expression, at least 50% of human genes encode multiple transcripts derived from APA (Ji, Lee, Pan, Jiang, & Tian, 2009).

APA sites can be located in two different forms; in one form APA sites are located in internal introns/exons and produce different protein isoforms which referred as CR-APA (coding region-APA) and in another form all APA sites are located in the 3' UTR and produce same protein with different length of mRNA which referred as UTR-APA. While CR-APA can affect gene expression qualitatively by producing distinct protein isoforms, UTR-APA can affect expression quantitatively (Di Giammartino et al., 2011). Not only expression quality but also mRNA's stability and translational properties are affected by the 3'UTR length (Mayr & Bartel, 2009; Zlotorynski & Agami, 2008). Also physiological conditions like cell growth and cancer like pathological events can influence the differential processing at multiple polyA sites.

1.2.5 Databases

NCBIGEO (Edgar, 2002a) is one of the most widely uses resources for gene expression data. GEO (Gene Expression Omnibus) is a public functional data vault which support MIAME-compliant data submissions. GEO freely distributes microarray, next-generation sequencing, and other forms of functional data outputs submitted by researchers. By the help of the this data storage, GEO helps users query and download gene expression datasets. Because GEO is public, anybody can access and download GEO datasets. One and biggest advantage of using a public dataset is that researchers can access huge amount of data which one cannot handle by his own.

1.2.6 Bioinformatics and Statistics

Bioinformatics is a research area which is application of computer technology to the biological information management by using programming and statistics.

Statistics is the study of the collection, organization, analysis, interpretation and presentation of data and deals with all aspects of this, including the planning of data collection in terms of the design of surveys and experiments (Moses, 1986).

P-value:

P-value is the probability of obtaining a test statistic which assumes that the null hypothesis is true. For p-value 0.05 or 0.01 are the boundary values which are indicated that the observed results would be highly unlikely under the null hypothesis and also named as significance level α. In other words rejects the null hypothesis if p-value is under 0.05 or 0.01 (Goodman, 1999).

Standard Deviation:

When observation values lie close to the mean, the dispersion is less than the values when values are scattered. For that a term called variance is used for measuring the

dispersion relative to the scatter of the values about mean. Also standard deviation shows variation and it is simply square root of the variance (Daniel, 2005a). That is the dispersion exists from the expected value. Data points are very close to the mean if standard deviation is low and data points are spread out if standard deviation is high.

Student t-test with Welch Correction:

In statistics, Student's t-distribution is a continuous probability distribution which arises when estimating the mean of a normally distributed population with small sample size and unknown standard deviation (Senn & Richardson, 1994).

A Student t-test follows Student's t distribution if the null hypothesis is supported. If the value of a scaling term is known, normal distribution is followed by a test statistic. If the scaling term is unknown and estimation based on data is used, the test statistic follows a Student's t distribution (Daniel, 2005b).

Welch Correction in statistics is an adaptation of Student's t-test. This correction is used when two samples having possibly unequal variances (Welch, 1947).

CHAPTER 2

MATERIALS AND METHODS

2.1 The Gene Expression Omnibus (GEO) Database

The Gene Expression Omnibus (GEO) is a public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community. In addition to data storage, a collection of web-based interfaces and applications are available to help users query and download the studies and gene expression patterns stored in GEO.

Three datasets from GEO was used for the analysis of 3'UTR shortening events. Those datasets are GSE3744, GSE7904, and GSE20711. First two datasets were used for developing analysis needs for APA. The third dataset, GSE20711 was used for further analysis and independent validation of the first analysis by means of using totally different samples.

All three datasets have similar basic properties and they are from same platform named as GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array) and all of them were handled as raw data format which had .CEL file types.

GSE7904 was the primary dataset for building algorithm which had total 62 samples but 7 of them were used for control group and 18 of them were used for cancer group which are samples of basal subtype of breast cancer. GSE3744 was used for obtain more results which dataset had 47 samples and 7 of them were control group and rest is for cancer group. Here control group and 18 samples out of the 40 were same as the GSE7904 database samples. Finally GSE20711 were used for making analysis

with completely different dataset which had 2 control samples, 26 basal subtype samples, 26 HER2 subtype samples, and 13 LumA subtype samples.

The GEO database has a flexible and open design that is responsive to developing trends.

Raw data must be provided by the microarray submitters either within the sample record data tables or as external supplementary data files.

2.2 Methods

At the first step, GSE7904 dataset was tested for validation of the proposed 3'UTR shortening identification technique. Then, results were converted to readable format. Genes which have more differences between proximal and distal probes expression levels for normal and tumor samples were selected. At first, genes with difference rates over 3.0 were selected to make it simple and easy to work. After comparing the test data with GSE3744 data results, difference rate threshold 2.0 was used in our analyses.

After the analysis for just one subtype of the breast cancer which is the basal subtype, the proposed algorithm was applied on GSE20711, which is a dataset with not only basal subtype samples but also luminal A , luminal B, and HER2 subtypes of breast cancer. At this point differences between subtypes were observed for difference threshold of 2.0.

2.3 The Proposed Method for Identification of Differential 3'UTR Shortening

To obtain polyA information for all the genes, polyA database file was read initially and then genes with less than 2 polyA sites were ignored. For each polyA site, for a gene site Name, position on the chromosomes, number of supporting ESTs, and strand information are recorded. Also all of the refseq IDs associated with a gene

which have more than one refseq IDs were gotten from the polyA database file. Including those terms, 5679 unique genes with more than one polyA site were used in the current version of the polyA database(Zhang, Hu, Recce, & Tian, 2005).

Unigene names and corresponding Affymetrix probe set IDs were read from the Affymetrix Annotation file. Refseq IDs and unigene IDs were matched in the annotation. In the annotation file the 11th column contains the unigene id and 24th column contains refseq ids separated by "///". For a unigene ID in polyA database, probe set of first occurrence of that unigene ID was used if refseq IDs in that annotation matches in the polyA database file. So, out of 5679 genes, 5217 genes were found on the chip annotation file.

For each probe in the probe sets, probe locations were read from probe set information file. To handle a correct probe location, 13 was subtract from the location given in the file since the location in the file is the middle position of a 25mer on the target gene. Probe locations are read relative to the transcription start site.

At the final stage the probe alignment file was read and probe sequences onto genome positions were mapped. Information which was handled from that reading was matched with the genome positions of polyA sites and whether there was a probe set whose probes were split into two sets by a polyA site. For further procedures, only the first half of the alignment file which contains the mappings of the probes onto genomic positions were used.

Formed alignment file contained 21 columns for each alignment line. 10th column was probe set IDs and 9th column was strand information. Also polyA site in the polyA database and the probe set were on the same strand. 19th column contained block sizes separated by columns. These blocks were subsequences on the 3' UTR. The last column contained chromosome start positions of these blocks of sequences.

Then a mapping of each block into actual chromosome positions were created. These chromosome positions were used to find out whether polyA site splits the probe sequences in a probe set into two nonempty subsets. Probe sequences to the upstream of the polyA site are named "Valid Probes", since they will be able to be used to

measure expression and Probe sequences to the downstream are called "Invalid Probes."

Detailed information about split probe sets were found in the output. So that output can be used by the programs that read expression intensities from CEL formatted files.

After that process, the average probe intensities of control and cancer group for the valid and invalid subsets of a probe set were analyzed by examining the split probe set data with valid and invalid probe sequences. Then that output was used to identify the genes with a difference higher than the specific fold change threshold between the expression levels of valid and invalid probe subsets with using unigene dictionary to report gene names.

At the final step gene lists were sorted according to that threshold values and top genes were selected for further experiments. Then selected genes were analyzed further to measure significance. For this purpose, student t-test with Welch-correction was used, since, the data was unpaired and cancer and control groups had unequal variances. Also for statistical tests GraphPad Prism 5 statistic program was used.

CHAPTER 3

RESULTS

In this chapter, we present the results of our probe level analysis on the three GEO datasets described in the previous chapter.

3.1 Primary Set Results of the 3'UTR Shortening Method

For primary tests GSE7904 dataset from GEO was used. This dataset contains originally 62 samples. But 7 control samples and 18 basal type breast cancer samples were selected for analysis.

Unpaired t test with Welch's correction method was performed on the data. All results were found to be between cancer proximal - cancer distal ratio and control proximal - control distal ratio.

As seen from Table 1; there is a significant difference between proximal and distal probe expression levels in cancer samples for each gene but in control samples that difference can not be observed. Because in control samples distal and proximal mean differences are relatively smaller than the mean values of distal and proximal probe expresion levels so their p-values will be greater than 0.05 which is the border for significance. But in cancer situation mean differences between distal and proximal are significantly different from each other.

Table 1: Top gene results from primary dataset GSE7904

	Control prox		Control dist		Cancer prox		Cancer dist		Control prox/distal		Cancer prox/distal	
genes	**min max**	**Sd mean**	**min max**	**Sd mean**	**min max**	**Sd mean**	**min max**	**Sd mean**	**Mean diff**	**p-value**	**Mean diff**	**p-value**
AURKA	139.8 242.4	33.42 171.3	104.0 188.4	32.6 152.3	316.0 1411	267.9 749.0	166.6 434.0	70.83 279.9	19 ±17.67	0.3052	469.0 ±65.32	<0.0001
BGN	171.4 318.9	58.91 228.3	143.0 207.3	21.54 181.0	124.6 1623	453.5 524.0	83.0 215.0	36.32 152.3	47.22 ± 25.61	0.1147	371.7 ±107.2	0.0030
DENR	107.0 177.0	22.92 128.5	87.33 152.3	22.77 111.2	89.0 683.7	129.3 183.8	79.14 191.7	26.54 111.4	17.29 ±12.21	0.1846	72.45 ±31.11	0.0317
LFRN1	195.3 290.8	41.22 238.8	182.3 249.8	23.37 218.6	180.0 930.3	271.0 484.4	159.8 299.0	42.96 235.1	20.20 ±17.91	0.2885	249.3 ±66.56	0.0018
RAB39	101.3 215.3	41.96 151.7	111.3 166.0	22.60 130.1	133.0 798.0	185.4 416.4	93.17 185.2	25.17 134.8	21.63 ±18.01	0.2604	281.6 ±44.11	<0.0001
SLC16A3	214.4 501.5	120.3 329.2	248.2 392.8	51.56 282.1	308.3 1575	319.7 733.3	170.4 322.0	41.09 248.1	47.10 ±49.47	0.3689	485.3 ±75.97	<0.0001
TOP2A	99.50 226.1	45.95 148.0	72.00 149.7	26.79 109.0	201.8 3602	857.9 1259	164.0 641.5	127.3 314.9	38.98 ±20.11	0.0845	944.4 ±204.4	0.0002

When comparing distal and proximal expression level difference between cancer and control groups it can be seen that distal expression levels did not change as much as proximal levels (Figure 2). And also as seen from Figure 2; distal expression levels have an increase rather than control distal levels but their mean are relatively close to each other when comparing with proximal levels. Also distal levels did not increase much but proximal levels increase times of control proximal levels.

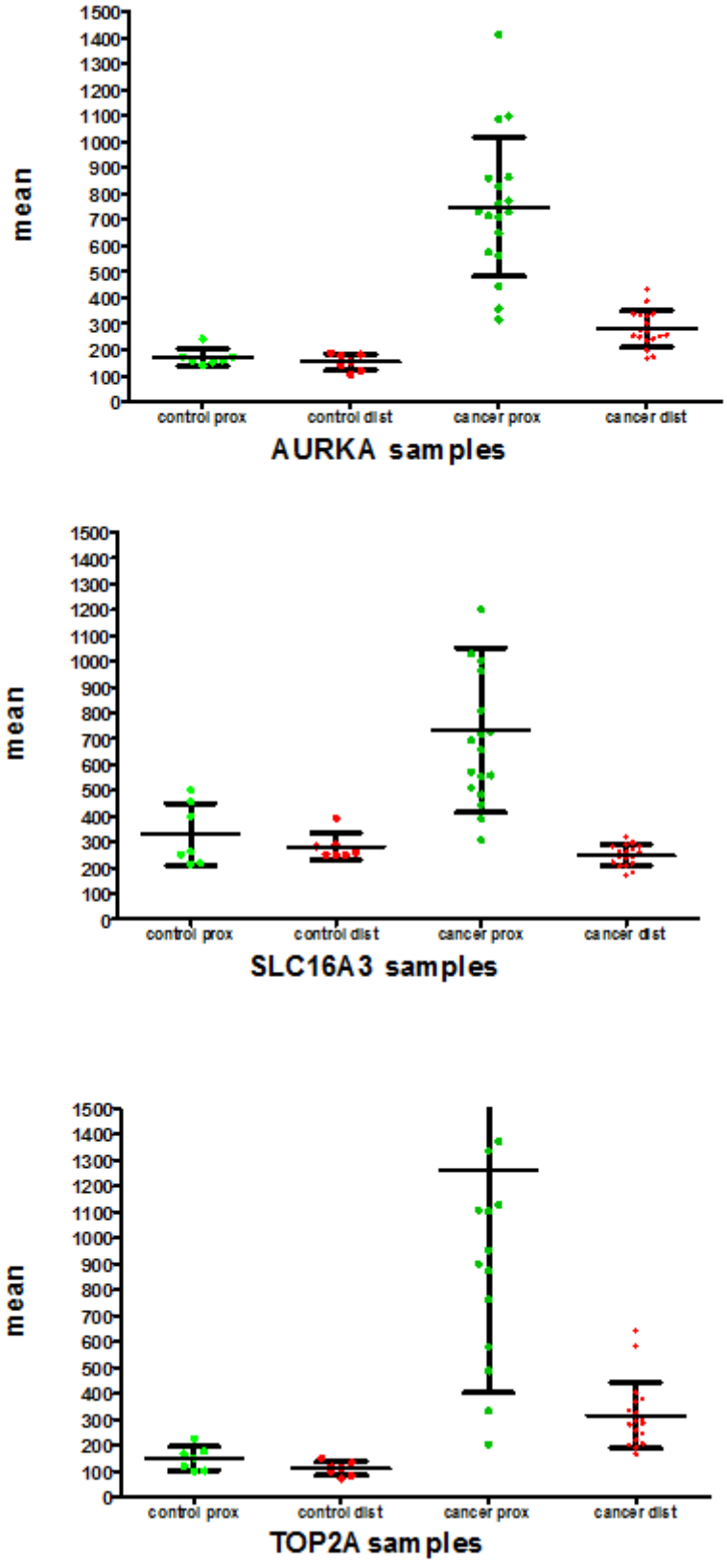

Figure 2: Means of top gene results from primary dataset GSE7904

3.2 Comparing Test Data Results with an Additional Data for Basal Type of Breast Cancer: GSE3744

To validate the first dataset results, another dataset from GEO (GSE3744) was used. That dataset was also obtained from the same research group and contained 7 control samples with 40 cancer samples. Control samples were the same as in dataset GSE7904.

Table 2: The highest five results of GSE3744 dataset

3744	Control prox		Control dist		Cancer prox		Cancer dist		Control prox/distal		Cancer prox/distal	
genes	min max	Sd mean	min max	Sd mean	min max	Sd mean	min max	Sd mean	Mean diff	p-value	Mean diff	p-value
AURKA	139.8 242.4	33.42 171.3	104.0 188.4	32.6 152.3	189.0 1411	280.5 680.8	133.6 434.0	77.93 268.9	19 ± 17.67	0.3052	411.9 ±46.03	<0.0001
BGN	171.4 318.9	58.91 228.3	143.0 207.3	21.54 181.0	123.0 1623	362.8 488.3	83.0 234.5	38.11 154.8	47.22 ± 25.61	0.1147	333.5± 57.68	<0.0001
DENR	107.0 177.0	22.92 128.5	87.33 152.3	22.77 111.2	89.0 683.7	172.4 251.8	75.57 207.9	31.16 123.2	17.29 ±12.21	0.1846	128.6 ±27.70	<0.0001
LFRN1	195.3 290.8	41.22 238.8	182.3 249.8	23.37 218.6	164.3 1088	262.2 476.8	146.3 366.0	45.91 225.5	20.20 ±17.91	0.2885	251.4 ±42.62	<0.0001
RAB39	101.3 215.3	41.96 151.7	111.3 166.0	22.60 130.1	69.33 842.8	209.5 394.7	69.33 194.0	27.42 134.3	21.63 ±18.01	0.2604	260.4 ±33.41	<0.0001
SLC16A3	214.4 501.5	120.3 329.2	248.2 392.8	51.56 282.1	308.3 1819	369.0 749.1	154.3 414.2	56.01 252.8	47.10 ±49.47	0.3689	496.4 ±59.01	<0.0001
TOP2A	99.50 226.1	45.95 148.0	72.00 149.7	26.79 109.0	201.8 3602	807.0 1032	113.0 651.5	140.9 296.1	38.98 ±20.11	0.0845	735.6 ±129.5	<0.0001

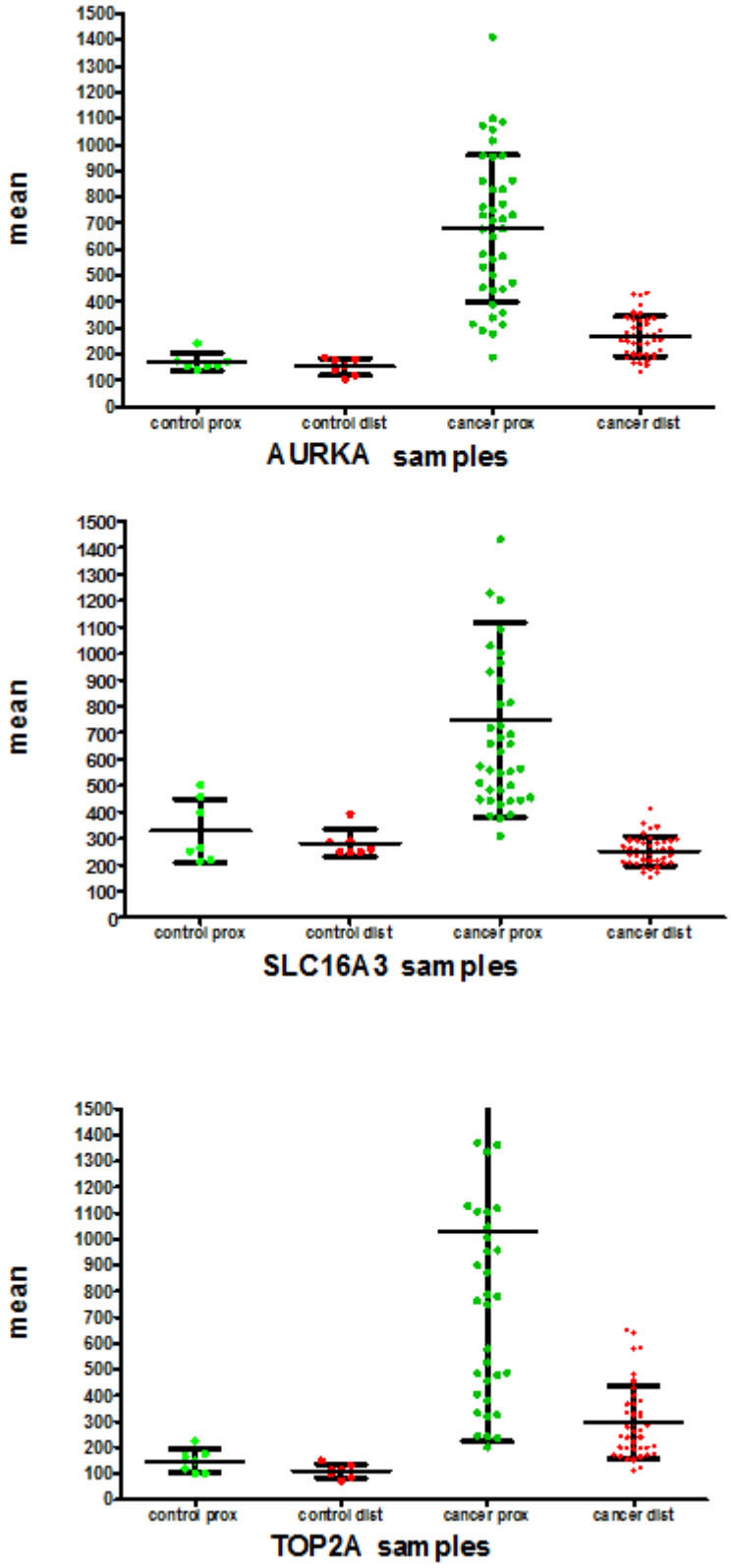

Figure 3: Means of the highest five results of GSE3744 dataset

As seen from the Table 2; both GSE7904 and GSE3744 had nearly similar results. Just like GSE7904, GSE3744 result showed that P-values of genes indicate significant shortening of 3'UTR. In addition their standard deviation values were relatively small like previous results. That means data values were close to the mean and most data were covered in close range area of the mean. So GSE3744 results support the GSE7904 results.

When Figure 3 is examined, it can be seen that there is also increase of proximal probe expression levels in cancer samples similar to the previous results of dataset GSE7904 but there is not a significant increase in distal probe expression levels.

Table 3: Comparison of top genes for GSE7904 and GSE3744

genes	7904				3744			
	CONTROL		CANCER		CONTROL		CANCER	
	P-value	Mean diff.	P-value	Mean diff.	P-value	Mean diff.	P-value	Mean diff.
AURKA	19.00 ±17.67	0.3052	469.0 ±65.32	<0.0001	19.00 ± 17.67	0.3052	411.9 ±46.03	<0.0001
SLC16A3	47.10 ±49.47	0.3689	485.3 ±75.97	<0.0001	47.10 ±49.47	0.3689	496.4 ±59.01	<0.0001
TOP2A	38.98 ±20.11	0.0845	944.4 ±204.4	0.0002	38.98 ±20.11	0.0845	735.6 ±129.5	<0.0001
BGN	47.22 ± 25.61	0.1147	371.7 ±107.2	0.0030	47.22 ± 25.61	0.1147	333.5± 57.68	<0.0001
DENR	17.29 ±12.21	0.1846	72.45 ±31.11	0.0317	17.29 ±12.21	0.1846	128.6 ±27.70	<0.0001
LFRN1	20.20 ±17.91	0.2885	249.3 ±66.56	0.0018	20.20 ±17.91	0.2885	251.4 ±42.62	<0.0001
RAB39	21.63 ±18.01	0.2604	281.6 ±44.11	<0.0001	21.63 ±18.01	0.2604	260.4 ±33.41	<0.0001

In summary, according to the tables and figures above, the GSE7904 and GSE3744 dataset results are similar to each other and in both datasets proximal probe expression levels are increased at cancer tissues but there was not a significant change in distal expression levels.

3.3 Additional Tests for Comparing Basal Types of Different Dataset by Using GSE20711

Up to this point, two nearly identical datasets were used to create an argument. For further tests a completely new dataset was used. This dataset was also obtained from the GEO database and marked as GSE20711 which contained basal subtype of breast cancer and also three other subtypes HER2, LumA, and LumB subtypes. However, it was not possible to analyze LumB subtypes because some of the samples were measured on an unsupported platform. Also this dataset had a problem that only two control samples were used during the experiment which caused some lack of statistical information during analysis, especially standard deviation results of the control group were not meaningful.

For basal subtype samples of GSE20711 dataset, as seen from Table 4, for all top genes, significant results were collected for cancer proximal distal expression level differences. According to their standard deviation values, it can be said that data results were close to the mean value because they had small SD values both in cancer and control samples. Also when comparing control proximal and distal mean differences between cancer proximal and distal mean differences it can be seen that there was not significant difference in control mean differences but with the increase of the proximal expression level there were signifant differences for cancer samples for selected genes.

Table 4: Basal Subtype Results of GSE20711

basal	Control prox		Control dist		Cancer prox		Cancer dist		Control prox/distal		Cancer prox/distal	
genes	min max	Sd mean	min max	Sd mean	min max	Sd mean	min max	Sd mean	Mean diff	p-value	Mean diff	p-value
AURKA	79.75 80.25	0.35 80	53 73.40	14.42 63.20	156.3 944.7	219.7 559.9	92.20 433.3	85.48 222.0	16.80 ±10.20	0.3475	337.8 ±46.24	<0.0001
BGN	70.62 114.9	31.30 92.75	82.33 115.0	23.10 98.67	73.25 257.2	44.10 120.5	71.00 220.3	38.41 123.6	-5.915 ±27.51	0.8652	-3.171 ±13.42	0.8146
DENR	104.3 124.3	14.14 114.3	77.57 107.0	20.81 92.29	87.33 214.0	33.80 148.7	44.80 186.9	38.22 116.1	22.05 ±17.79	0.4323	32.59 ±10.01	0.0020
RAB39	64.25 96.75	22.98 80.50	77.57 107.0	20.81 92.29	64.75 210.3	31.81 91.40	44.80 186.9	38.22 116.1	17.79 ±17.04	0.4863	30.63 ±6.900	<0.0001
SLC16A3	87.80 199.7	79.10 143.7	79.40 88.40	6.364 83.90	116.6 1370	349.8 554.4	62.50 116.0	15.31 84.38	59.84 ±56.12	0.4796	470.0 ±48.67	<0.0001
TOP2A	77.71 116.0	27.08 96.86	70.67 92.00	141.0 607.5	340.3 2070	469.2 872.7	141.0 607.5	131.7 298.30	15.52 ±21.92	0.6077	574.4 ±95.57	<0.0001

Also according to Figure 4, previous results which were obtained from Table 4 were understood more clearly. As seen from graphs there was not a significant difference between distal expression levels of cancer and control samples but when analyzing the proximal levels a huge increase at cancer situation was occurred and cause significant difference between cancer and control samples.

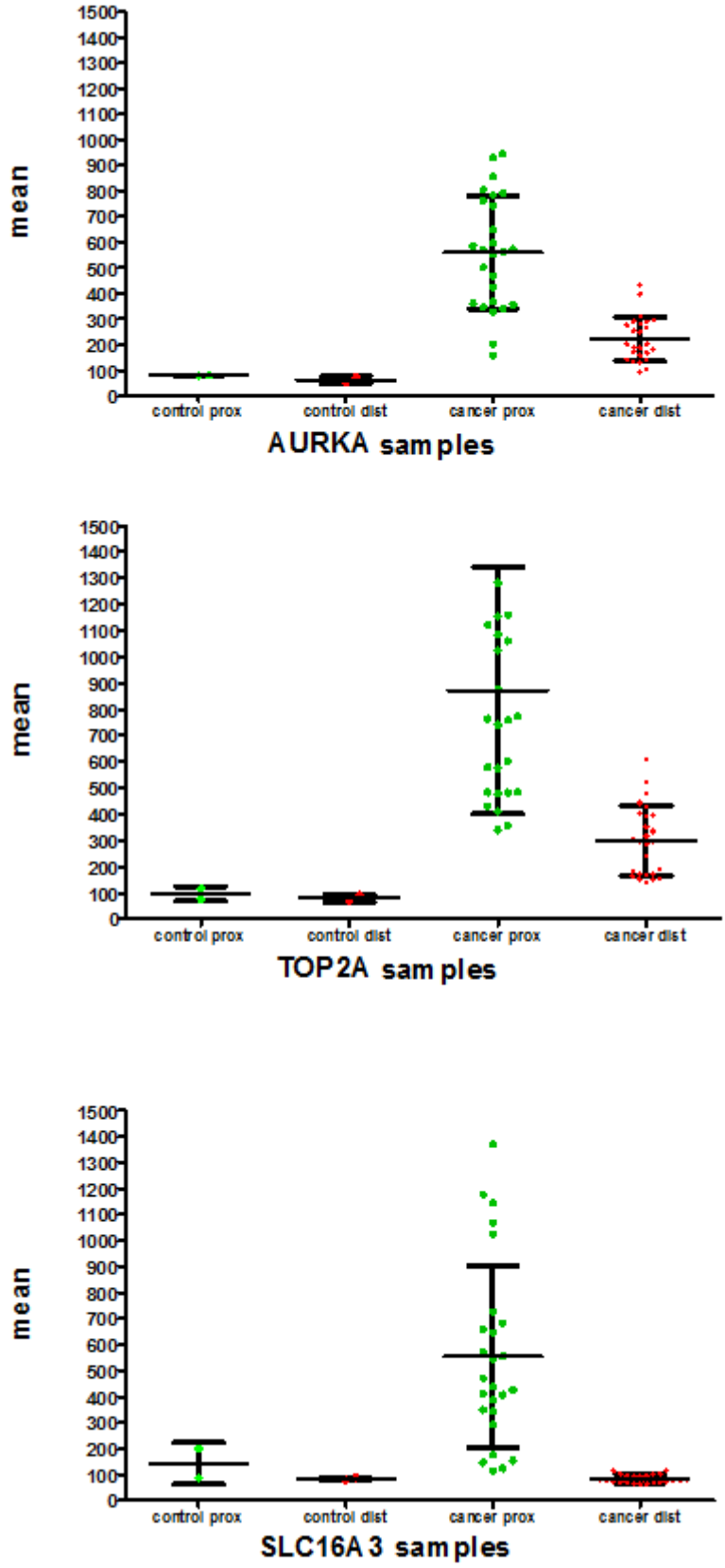

Figure 4:Means of GSE20711 dataset for basal subtype

3.4 Additional Tests for Comparing Some Types of Breast Cancer for GSE20711: LumA and HER2 subtypes.

As seen from Table 5; for LumA subtype samples of GSE20711 dataset, AURKA, SLC16A3, and TOP2A genes had similar results with basal subtype. In addition there were some genes which had increase at proximal expression level during cancer situation like C4A. Also as seen from Figure 5; there were significant increase in cancer samples proximal levels but not in distal samples and for ZNF214 which had no significant results for basal subtype, again there were a significant increase in cancer samples proximal probe expression levels but not in distal samples.

Table 5:GSE20711 LumA subtype results

Luma	Control prox		Control dist		Cancer prox		Cancer dist		Control prox/distal		Cancer prox/distal	
genes	min max	Sd mean	min max	Sd mean	min max	Sd mean	min max	Sd mean	Mean diff	p-value	Mean diff	p-value
AURKA	79.75 80.25	0.3536 80.00	53.00 73.40	14.42 63.20	87.00 302.8	62.69 173.6	66.50 161.6	23.31 92.91	16.80 ±10.20	0.3475	80.67 ±18.55	0.0006
SLC16A3	87.80 199.7	79.10 143.7	79.40 88.40	6.364 83.90	166.8 460.8	92.06 303.9	65.75 98.00	9.792 77.22	59.84 ±56.12	0.4796	226.7 ±25.68	<0.0001
TOP2A	77.71 116.0	27.08 96.86	77.57 107.0	20.81 92.29	79.00 396.9	96.39 205.6	44.80 186.9	38.22 116.1	15.52 ±21.92	0.6077	85.29 ±30.29	0.0114
C4A	396.7 1663	895.6 1030	113.0 147.0	24.04 130.0	609.9 7553	1931 3019	118.5 365.0	67.21 180.8	899.9 ±633.5	0.3905	2838 ±535.9	0.0002
MRP63	530.5 613.5	58.69 572.0	320.3 332.8	8.839 326.6	746.3 2741	618.6 1154	214.2 525.0	93.08 328.8	245.4 ±41.97	0.1078	825.7 ±173.5	0.0005
ZNF214	64.40 66.20	1.273 65.30	57.00 60.67	2.595 58.84	75.80 148.3	22.05 108.0	45.00 70.00	7.494 57.50	6.465 ±2.044	0.1949	50.51 ±6.459	<0.0001

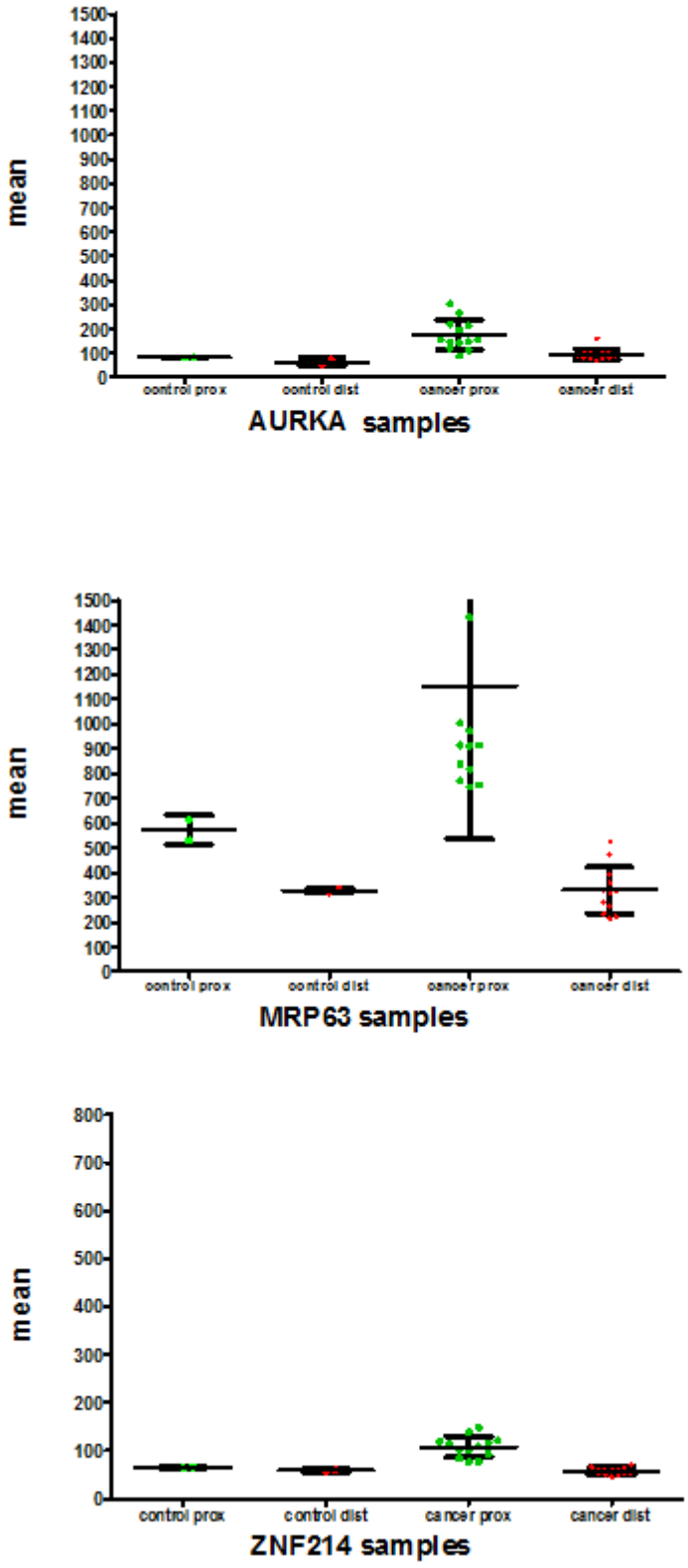

Figure 5:Means of GSE20711 dataset for LumA subtype

Table 6: GSE20711 HER2 subtype results

Her2	Control prox		Control dist		Cancer prox		Cancer dist		Control prox/distal		Cancer prox/distal	
genes	min max	Sd mean	min max	Sd mean	min max	Sd mean	min max	Sd mean	Mean diff	p-value	Mean diff	p-value
AURKA	79.75 80.25	0.3536 80.00	53.00 73.40	14.42 63.20	179.0 1847	343.2 522.1	84.33 1032	183.0 217.9	16.80 ±10.20	0.3475	304.2 ±76.28	0.0003
SLC16A3	87.80 199.7	79.10 143.7	79.40 88.40	6.364 83.90	122.6 1699	378.5 569.8	63.00 130.4	14.97 83.86	59.84 ±56.12	0.4796	486.0 ±74.29	<0.0001
TOP2A	77.71 116.0	27.08 96.86	70.67 92.00	15.08 81.34	198.3 3722	1043 1083	122.5 1447	358.4 396.8	15.52 ±21.92	0.6077	685.7 ±216.3	0.0035
CDC6	63.00 65.00	1.414 64.00	54.50 76.50	15.56 65.50	77.00 1085	272.0 241.0	56.50 441.0	77.61 118.3	-1.500 ±11.05	0.9141	122.7 ±55.46	0.0349
MRP63	530.5 613.5	58.69 572.0	320.3 332.8	8.839 326.6	532.0 2751	577.6 1217	185.0 464.2	78.69 326.3	245.4 ±41.97	0.1078	890.3 ±114.3	<0.0001
NSDHL	129.1 131.0	0.329 130.1	53.50 55.00	1.061 54.25	135.4 575.8	136.9 295.1	40.50 67.50	6.038 55.65	75.81 ±1.203	0.0101	239.5 ±26.88	<0.0001
TCF3	290.6 345.0	38.47 317.8	70.25 80.25	7.071 75.25	435.4 1583	321.9 875.1	69.00 263.4	36.44 96.07	242.6 ±27.66	0.0723	779.0 ±63.53	<0.0001

As seen from Table 6; for Her2 subtype samples of GSE20711 dataset, AURKA, SLC16A3, and TOP2A genes had similar results with basal subtype like LumA results. In addition there were some genes which had increase at proximal expression level during cancer situation like CDC6. Also as seen from Figure 6; there were significant increase in cancer samples proximal levels but not in distal samples and for NSDHL which had no significant results for basal subtype, again there were a significant increase in cancer samples proximal probe expression levels but not in distal samples.

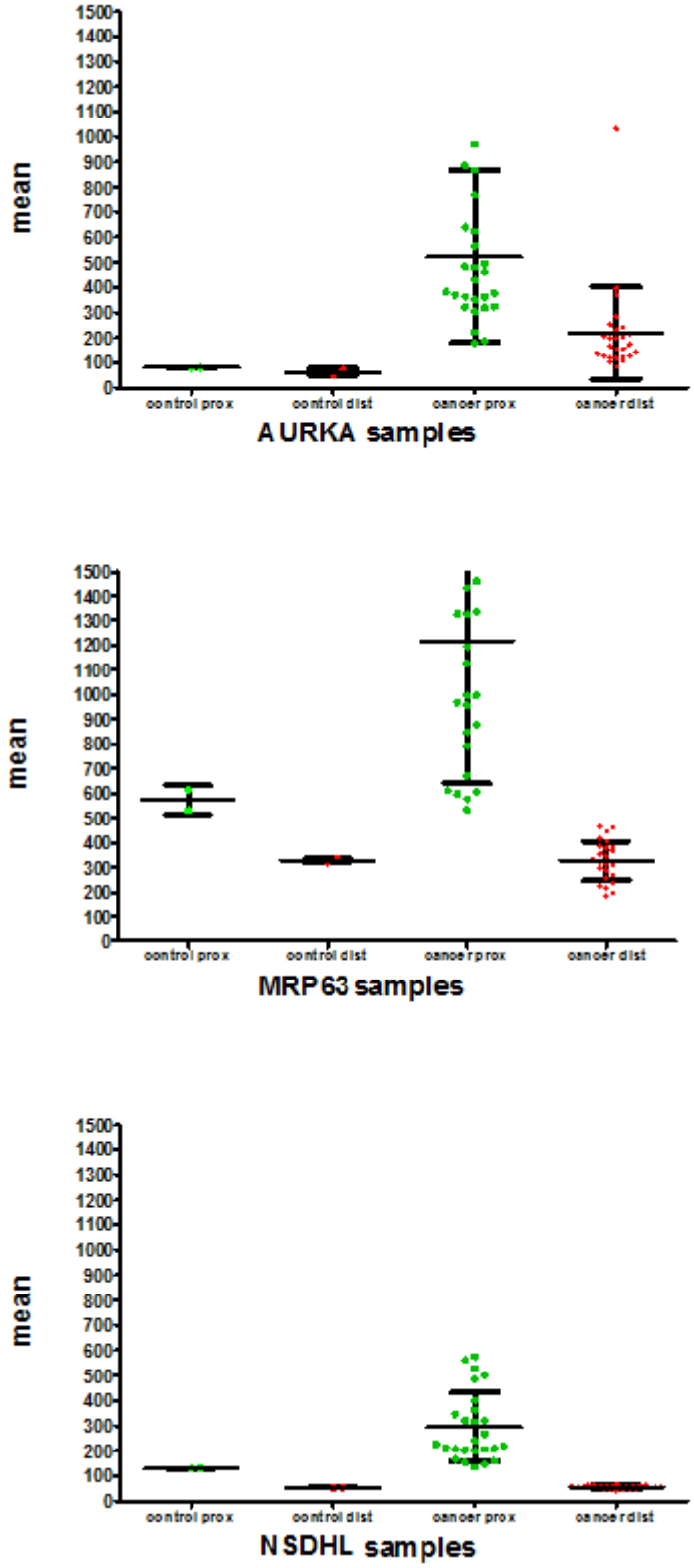

Figure 6: Means of GSE20711 dataset for HER2 subtype

According to the tables and graphs above, the top genes; AURKA, SLC16A3, and TOP2A had significant results that there were increase in expression levels of proximal probes, especially in basal subtype samples of breast cancer. Also for basal subtype samples of GSE20711 set, all genes have similar behavior with previous results except BGN gene so that gene cannot be classified as a marker gene. For LumA and HER2 samples, they had their own genes with significant results which may be used for subtype selection during analysis.

Table 7: Top 20 genes of GSE20711

GENES	BASAL		HER2		LUMA	
	p-value	Cancer/control ratio	p-value	Cancer/control ratio	p-value	Cancer/control ratio
AURKA	<0.0001	1,94	0.0003	1,92	0.0006	1,42
SLC16A3	<0,0001	3,84	<0.0001	3,88	<0.0001	2,29
TOP2A	<0,0001	2,44	0.0035	2,25	0.0114	1,48
BGN	0,8146	1,05	0,8232	1,11	0.1778	0,99
DENR	0,0020	1,08	0.0502	0,94	0,0723	0,95
RAB39	<0,0001	1,19	<0.0001	1,16	0,0016	1,09
C4A	0,0030	0,57	0,0007	0,72	0.0002	2,16
MRP63	<0.0001	2,20	<0.0001	2,06	0.0005	0,91
ZNF214	0,1495	1,08	0,0035	1,26	<0.0001	1,70
CDC6	<0.0001	1,06	0.0349	1,93	0,7044	1,02
NSDHL	<0.0001	2,34	<0.0001	2,25	0,0018	1,49
TCF3	<0.0001	2,18	<0.0001	2,24	<0.0001	1,67

According to the Table 7; in all subtypes top three genes (AURKA, SLC16A3, and TOP2A) had significant results. And also each subtype had its own specific genes which had significant results for only itself but not with the others. Also as seen from table, these specific genes have smaller difference values for other subtypes. This difference is ratio of cancer proximal/distal ratio and control proximal/distal ratio.

Table 8: Results of top genes for differential expression analysis from GEO for GSE7904

Gene symbol	Affymetrix probeset ID	P-VALUE	Significance diff.
AURKA	204092_s_at	1.28E-10	YES
SLC16A3	217691_x_at	1.15E-05	YES
TOP2A	201291_s_at	1.57E-09	YES
DENR	238982_at	7.91E-01	NO
LFRN1	232486_at	6.53E-04	YES
RAB39	1554800_at	1.70E-01	NO

Table 9: Results of top genes for differential expression analysis from GEO for GSE3744

Gene symbol	Affymetrix probeset ID	P-VALUE	Significance diff.
AURKA	204092_s_at	1.14E-10	YES
SLC16A3	217691_x_at	6.49E-04	YES
TOP2A	201291_s_at	4.10E-15	YES
DENR	238982_at	1.27E-01	NO
LFRN1	232486_at	7.19E-01	NO
RAB39	1554800_at	1.93E-01	NO

Table 10: Results of top genes for differential expression analysis from GEO for GSE20711 for basal

Gene symbol	Affymetrix probeset ID	P-VALUE	Significance diff.
AURKA	204092_s_at	1.84E-06	YES
SLC16A3	217691_x_at	1.60E-01	NO
TOP2A	201291_s_at	3.50E-07	YES
DENR	238982_at	9.53E-01	NO
LFRN1	232486_at	3.70E-01	NO
RAB39	1554800_at	7.68E-01	NO

Table 11: Results of top genes for differential expression analysis from GEO for GSE20711 for HER2

Gene symbol	Affymetrix probeset ID	P-VALUE	Significance diff.
AURKA	204092_s_at	6.91E-05	YES
SLC16A3	217691_x_at	1.22E-01	NO
TOP2A	201291_s_at	2.98E-04	YES
CDC6	203967_at	5.06E-02	NO
MRP63	204387_x_at	9.45E-01	NO
NSDHL	215093_at	3.90E-02	YES
TCF3	213730_x_at	1.17E-01	NO

Table 12: Results of top genes for differential expression analysis from GEO for GSE20711 for LumA

Gene symbol	Affymetrix probeset ID	P-VALUE	Significance diff.
AURKA	204092_s_at	1.40E-01	NO
SLC16A3	217691_x_at	2.21E-01	NO
TOP2A	201291_s_at	2.41E-01	NO
C4A	214428_x_at	1.44E-01	NO
MRP63	204387_x_at	1.98E-01	NO
ZNF14	220497_at	4.62E-01	NO

According to Tables 12 to 16; when analyzing some top genes identified by the proposed 3' UTR shortening assay, it can be said that most top genes had significant results of differential expression results for first two datasets from GEO2R (Edgar, 2002b). Most genes which were analyzed for further researches had significant p-values both for 3'UTR shortening and differential expression results. According to these results especially SLC16A3, AURKA, and TOP2A genes can be used for breast cancer 3'UTR shortening analysis.

3.5 Functions of the candidate genes

Table 13: Functions of the top genes

GENE SYMBOL	CODED PROTEIN	FUNCTION
AURKA	Aurora A kinase	formation of microtubules and stabilization at the spindle pole during G2/M transition (Hannak, Kirkham, Hyman, & Oegema, 2001)
SLC16A3	monocarboxylate transporter 4	Has role in TCA cycle and signaling in immune system pathways (Hu et al., 2009)
TOP2A	DNA topoisomerase 2-alpha	controls DNA's topologic states during transcription ("TOP2A topoisomerase (DNA) II alpha 170kDa [Homo sapiens] - Gene - NCBI," n.d.)
NSDHL	Sterol-4-alpha-carboxylate 3-dehydrogenase	Localized in the endoplasmic reticulum and involved in cholesterol biosynthesis. Its mutation cause x-linked dominant disorder of lipid metabolism ("NSDHL NAD(P) dependent steroid dehydrogenase-like [Homo sapiens] - Gene - NCBI," n.d.)
TCF3	Transcription factor 3	Member of E-protein family that activates transcription by binding regulatory E-box sequences. Also involved in some chromosomal translocations ("TCF3 transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47) [Homo sapiens] - Gene - NCBI," n.d.).
MRP63	mitochondrial ribosomal protein 63	Has role protein synthesis within the mitochondrion ("MRP63 mitochondrial ribosomal protein 63 [Homo sapiens] - Gene - NCBI," n.d.).
CDC6	Cell division cycle 6	Has a regulatory role at the initiation of the DNA replication. Localized cell nucleus during G1 phase ("CDC6 cell division cycle 6 [Homo sapiens] - Gene - NCBI," n.d.).

3.5.1 AURKA

This gene encodes Aurora A kinase protein which is also known as serine/threonine-protein kinase 6. Aurora A enzyme has important role during cell division during the transition of G2 phase to M phase. During this transition this enzyme is responsible for formation of microtubules and stabilization at the spindle pole (Hannak et al., 2001). This protein is also in interaction with BRCA1 and p53. Aurora A activity is controlled by p53 in different levels like phosphorylation by mean of negative regulation (Crane, Gadea, Littlepage, Wu, & Ruderman, 2004). Also Aurora-A localizes to the centrosome during G2 phase to M phase transition and it regulates this phenomenon. If there is a loss of this transition checkpoint it will results with the loss of BRCA1 activation. Also biochemically BRCA1 is physically phosphorylated by Aurora-A in vivo (Ouchi et al., 2004). This enzyme is also in interaction with some more proteins like a metastasis suppressor nucleoside diphpsphate kinase A which is encoded by NME1 gene (Du & Hannon, 2002). Because of its interactions with genes which have important roles in cell division, AURKA gene corruption may result the inactivation of some genes like BRCA1.

After 3'UTR shortening AURKA has been involved and positively implicated in tumor resistance and progression to therapy express (Lembo, Di Cunto, & Provero, 2012) . Also in specific mRNAs 3'UTR shortening correlates with poor prognosis in breast cancer (Wang et al., 2010).

3.5.2 SLC16A3

This gene encodes monocarboxylate transporter 4 protein which is a member of a transporter family (MCT). This enzyme has roles in hemosthasis, metabolism of carbohydrates, TCA cycle and signaling in immune system pathways. Compared to primary tumors SLC16A3 expression is higher in breast cancer distant metastasis (Hu et al., 2009). SLC16A3 has several miRNA binding sites. One of them is miR-339-5p binding site. miR-339-5p labeled as a potential biomarker because it inhibits breast cancer cell migration and invasion (Wu et al., 2010).

3.5.3 TOP2A

This gene encodes DNA topoisomerase 2-alpha enzyme which controls DNA's topologic states during transcription. Also this enzyme has role during chromosome condensation, chromatid separation, and DNA transcription and replication ("TOP2A topoisomerase (DNA) II alpha 170kDa [Homo sapiens] - Gene - NCBI," n.d.). Also it has interaction with p53 like AURKA. C-terminal basic region of p53 regulatory is necessary for interaction with DNA topoisomerase 2 (Cowell et al., 2000). Also TOP2A has interaction with CDC5L which encodes cell division cycle 5-like protein. This enzyme has regulatory role during G2 to M phase transition (Ajuh et al., 2000).

HuR is a RNA-binding protein which is ubiquitous regulator of the translation and stability of bound transcripts (López de Silanes, Zhan, Lal, Yang, & Gorospe, 2004). HuR increase the translation of TOP2A gene by binding to the 3'UTR of the gene and overexpressed in cancers. Recruitment of TOP2A transcripts are triggered by reducing the HuR levels. It has been shown that microRNA miR-548c-3p interacts with TOP2A 3'UTR and repressed the TOP2A translation by antagonizing of HuR action. Also lowering of TOP2A expression under the control of HuR or miR-548c-3p decreased DNA damage after treatment (Srikantan et al., 2011). As seen from Figure 7, there are two special sites at 3'UTR of the TOP2A. One of them is miR-548c-3p site (red) and other is HuR CLIP site (blue) and shortening at 3'UTR site may result lack of one or more these sites.

Figure 7 : TOP2A mRNA

CHAPTER 4

CONCLUSION

The results and contributions of this thesis can be listed as follows:

Differential expression is an effective method but there are some limitations for this technique. The most important one is unstable results of low expressed genes. Also RNA isolation and cDNA synthesis are required for differential expression and it makes it possible only comparing tumor and healthy samples, even in this condition, false positive and false negative results can be handled. In addition to these limitations differential expression procedures are also time consuming and costly.

On the other side, 3'UTR shortening analysis does not use gene expression levels of genes, but investigates differences between expression levels of probes of a gene. It does not matter if a gene is expressed at high or low levels one as long as differences between the probes are observed. One limitation for 3'UTR shortening analysis is if all probes found one side of polyadenylation site, it is impossible to make analysis to differentiate between proximal and distal expression levels.

According to the results for GSE7904 and GSE3744 datasets which have basal subtype of breast cancer samples, there are especially 5 genes that can be used for analysis of length difference between cancer and normal samples. For statistically

AURKA, SLC16A3, and TOP2A have p-values lower than 0.05 with low standard deviations, which means that their results are significant and because of the low SD, individual sample results are close to the mean. Also for all these genes which has significant results, there is a significant increase of their proximal probe expression levels during cancer situation which means when shorter 3'UTR is observed, the sample may have basal type of breast cancer.

The aforementioned genes have interactions with other genes like p53, BRCA1, and CDC5L which have key roles during cell cycle. Also SLC16A3 has role during TCA cycle. All have importance not only breast cancer but also all types of cancer. Also some other genes which are used as candidate genes of subtype specific analysis have also critical roles in some phenomena like cell cycle and transcription. So our candidate genes can be used for breast cancer 3'UTR shortening analysis as marker genes.

For the further experiments GSE20711 dataset was used. According to that dataset results, again candidate genes had significant results with low SD values for basal subtype. Also that dataset contain HER2 and LumA subtype samples and these two also have significant results of those genes. But when their own top genes were analyzed there were differences. All subtypes have their own top genes but those three genes are found in all subtypes.

Also there are some more significant genes which can be used for diagnosis for breast cancer which can be separated into the two groups. First group is the genes which are just for if the sample has breast cancer or not analysis just like AURKA because these genes gave significant results for all three subtypes. Second group is the genes that are for finding the subtypes because these are found only one or two subtypes with significant results like CDC6, ZNF214, and NSHDL.

In addition, those tests were made for only HGU133plus2 chipset that limits the data we can analyze. But we still found significant results in spite of that limitation. For further studies, more chipsets can be added to the algorithm more genes could we be analyzed.

Finally, we must compare these results with differential expression results. Because we need proof for compliance of 3'UTR analysis results and differential expression results. According to the results, it can be said that we have high compliance. Again especially AURKA, SLC16A3, and TOP2A have also significant results for differential expression for first two datasets. Also for subtypes they have non-significant results for differential expression which means 3'UTR shortening may found different results from differential expression results that are caused by the unstable results or false and true negative results.

So; according to the differential expression results of datasets which were handled from GEO, it can be said that 3'UTR shortening analysis results are highly coherent and there are specific genes like AURKA, SLC16A3, and TOP2A which can be used for analysis of breast cancer. Also it is possible that subtypes can be separated from each other during analysis with 3'UTR shortening technique by using subtype specific genes. In addition; because of the costs, time consuming procedures and risks of false and true negative results, differential expression analysis is not a perfect tool. Rather than that technique 3'UTR shortening can be used for simple analysis steps.

REFERENCES

Abd El-Rehim, D. M., Pinder, S. E., Paish, C. E., Bell, J., Blamey, R. W., Robertson, J. F. R., Nicholson, R. I., et al. (2004). Expression of luminal and basal cytokeratins in human breast carcinoma. *The Journal of pathology*, *203*(2), 661–71. doi:10.1002/path.1559

Ajuh, P., Kuster, B., Panov, K., Zomerdijk, J. C., Mann, M., & Lamond, A. I. (2000). Functional analysis of the human CDC5L complex and identification of its components by mass spectrometry. *The EMBO journal*, *19*(23), 6569–81. doi:10.1093/emboj/19.23.6569

Alimonti, A., Carracedo, A., Clohessy, J. G., Trotman, L. C., Nardella, C., Egia, A., Salmena, L., et al. (2010). Subtle variations in Pten dose determine cancer susceptibility. *Nature genetics*, *42*(5), 454–8. doi:10.1038/ng.556

Asselin-Labat, M.-L., Sutherland, K. D., Barker, H., Thomas, R., Shackleton, M., Forrest, N. C., Hartley, L., et al. (2007). Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation. *Nature cell biology*, *9*(2), 201–9. doi:10.1038/ncb1530

Campeau, P. M., Foulkes, W. D., & Tischkowitz, M. D. (2008). Hereditary breast cancer: new genetic developments, new therapeutic avenues. *Human genetics*, *124*(1), 31–42. doi:10.1007/s00439-008-0529-1

Carey, L. A., Perou, C. M., Livasy, C. A., Dressler, L. G., Cowan, D., Conway, K., Karaca, G., et al. (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA : the journal of the American Medical Association*, *295*(21), 2492–502. doi:10.1001/jama.295.21.2492

Carroll, J. S., Liu, X. S., Brodsky, A. S., Li, W., Meyer, C. A., Szary, A. J., Eeckhoute, J., et al. (2005). Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, *122*(1), 33–43. doi:10.1016/j.cell.2005.05.008

CDC6 cell division cycle 6 [Homo sapiens] - Gene - NCBI. (n.d.). Retrieved February 21, 2013, from http://www.ncbi.nlm.nih.gov/gene/990

Comprehensive Cancer Information - National Cancer Institute. (n.d.). Retrieved from http://www.cancer.gov/

Cowell, I. G., Okorokov, A. L., Cutts, S. A., Padget, K., Bell, M., Milner, J., & Austin, C. A. (2000). Human topoisomerase IIalpha and IIbeta interact with the C-terminal region of p53. *Experimental cell research*, *255*(1), 86–94. doi:10.1006/excr.1999.4772

Crane, R., Gadea, B., Littlepage, L., Wu, H., & Ruderman, J. V. (2004). Aurora A, meiosis and mitosis. *Biology of the cell / under the auspices of the European Cell Biology Organization*, *96*(3), 215–29. doi:10.1016/j.biolcel.2003.09.008

Creighton, C. J., Li, X., Landis, M., Dixon, J. M., Neumeister, V. M., Sjolund, A., Rimm, D. L., et al. (2009). Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(33), 13820–5. doi:10.1073/pnas.0905718106

Daniel, W. W. (2005a). *Biostatistics 8th Edition* (pp. 40–41). Wiley. Retrieved from http://www.amazon.com/Biostatistics-8th-Edition-SPSS-11-0/dp/0471778583

Daniel, W. W. (2005b). *Biostatistics 8th Edition* (p. 167).

Dedeurwaerder, S., Desmedt, C., Calonne, E., Singhal, S. K., Haibe-Kains, B., Defrance, M., Michiels, S., et al. (2011). DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO molecular medicine*, *3*(12), 726–41. doi:10.1002/emmm.201100801

Di Giammartino, D. C., Nishida, K., & Manley, J. L. (2011). Mechanisms and consequences of alternative polyadenylation. *Molecular cell*, *43*(6), 853–66. doi:10.1016/j.molcel.2011.08.017

Dixelius, J., Cross, M., Matsumoto, T., Sasaki, T., Timpl, R., & Claesson-Welsh, L. (2002). Endostatin Regulates Endothelial Cell Adhesion and Cytoskeletal Organization. *Cancer Res.*, *62*(7), 1944–1947. Retrieved from http://cancerres.aacrjournals.org/content/62/7/1944.short

Du, J., & Hannon, G. J. (2002). The centrosomal kinase Aurora-A/STK15 interacts with a putative tumor suppressor NM23-H1. *Nucleic acids research*, *30*(24), 5465–75. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=140054&tool=pmcentrez&rendertype=abstract

Edgar, R. (2002a). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210. doi:10.1093/nar/30.1.207

Edgar, R. (2002b). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210. doi:10.1093/nar/30.1.207

Goodman, S. N. (1999). Toward Evidence-Based Medical Statistics. 1: The, 995–1004.

Hankinson, S. E., Colditz, G. A., & Willett, W. C. (2004). Review Towards an integrated model for breast cancer etiology The lifelong interplay of genes , lifestyle , and hormones, 213–218. doi:10.1186/bcr921

Hannak, E., Kirkham, M., Hyman, A. A., & Oegema, K. (2001). Aurora-A kinase is required for centrosome maturation in Caenorhabditis elegans. *The Journal of cell biology*, *155*(7), 1109–16. doi:10.1083/jcb.200108051

Herman, J. G. (1996). Methylation-specific PCR: A novel PCR assay for methylation status of CpG islands. *Proceedings of the National Academy of Sciences*, *93*(18), 9821–9826. doi:10.1073/pnas.93.18.9821

Howe, H. L., Wingo, P. A., Thun, M. J., Ries, L. A. G., Rosenberg, H. M., Feigal, E. G., & Edwards, B. K. (2001). Annual Report to the Nation on the Status of Cancer (1973 Through 1998), Featuring Cancers With Recent Increasing Trends. *JNCI Journal of the National Cancer Institute*, *93*(11), 824–842. doi:10.1093/jnci/93.11.824

Hu, Z., Fan, C., Livasy, C., He, X., Oh, D. S., Ewend, M. G., Carey, L. A., et al. (2009). A compact VEGF signature associated with distant metastases and poor outcomes. *BMC medicine*, *7*, 9. doi:10.1186/1741-7015-7-9

Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X., Qaqish, B. F., Livasy, C., et al. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC genomics*, *7*(1), 96. doi:10.1186/1471-2164-7-96

Ji, Z., Lee, J. Y., Pan, Z., Jiang, B., & Tian, B. (2009). Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(17), 7028–33. doi:10.1073/pnas.0900028106

Kliewer, S. A., Lenhard, J. M., Willson, T. M., Patel, I., Morris, D. C., & Lehmann, J. M. (1995). A prostaglandin J2 metabolite binds peroxisome proliferator-activated receptor γ and promotes adipocyte differentiation. *Cell*, *83*(5), 813–819. doi:10.1016/0092-8674(95)90194-9

Kouros-Mehr, H., Slorach, E. M., Sternlicht, M. D., & Werb, Z. (2006). GATA-3 maintains the differentiation of the luminal cell fate in the mammary gland. *Cell*, *127*(5), 1041–55. doi:10.1016/j.cell.2006.09.048

Legendre, M., Ritchie, W., Lopez, F., & Gautheret, D. (2006). Differential repression of alternative transcripts: a screen for miRNA targets. (C. Lutz, Ed.)*PLoS computational biology*, *2*(5), e43. doi:10.1371/journal.pcbi.0020043

Lembo, A., Di Cunto, F., & Provero, P. (2012). Shortening of 3'UTRs correlates with poor prognosis in breast and lung cancer. *PloS one*, *7*(2), e31129. doi:10.1371/journal.pone.0031129

Lim, E., Vaillant, F., Wu, D., Forrest, N. C., Pal, B., Hart, A. H., Asselin-Labat, M.-L., et al. (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nature medicine*, *15*(8), 907–13. doi:10.1038/nm.2000

López de Silanes, I., Zhan, M., Lal, A., Yang, X., & Gorospe, M. (2004). Identification of a target RNA motif for RNA-binding protein HuR. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(9), 2987–92. doi:10.1073/pnas.0306453101

Mayr, C., & Bartel, D. P. (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, *138*(4), 673–84. doi:10.1016/j.cell.2009.06.016

Millikan, R. C., Newman, B., Tse, C.-K., Moorman, P. G., Conway, K., Dressler, L. G., Smith, L. V, et al. (2008). Epidemiology of basal-like breast cancer. *Breast cancer research and treatment*, *109*(1), 123–39. doi:10.1007/s10549-007-9632-6

Morris, G. J., Naidu, S., Topham, A. K., Guiles, F., Xu, Y., McCue, P., Schwartz, G. F., et al. (2007). Differences in breast carcinoma characteristics in newly diagnosed African-American and Caucasian patients: a single-institution compilation compared with the National Cancer Institute's Surveillance, Epidemiology, and End Results database. *Cancer*, *110*(4), 876–84. doi:10.1002/cncr.22836

Moses, L. (1986). Think and Explain with Statistics. Retrieved from http://www.citeulike.org/group/1014/article/563026

mRNA UTR Structure Exon Intron Cap - Molecular Biology Photo Gallery. (n.d.). Retrieved January 17, 2013, from http://www.molecularstation.com/molecular-biology-images/503-rna-pictures/66-mrna-utr-structure-exon-intron-cap.html

MRP63 mitochondrial ribosomal protein 63 [Homo sapiens] - Gene - NCBI. (n.d.). Retrieved February 21, 2013, from http://www.ncbi.nlm.nih.gov/gene/78988

NSDHL NAD(P) dependent steroid dehydrogenase-like [Homo sapiens] - Gene - NCBI. (n.d.). Retrieved February 21, 2013, from http://www.ncbi.nlm.nih.gov/gene/50814

Ouchi, M., Fujiuchi, N., Sasai, K., Katayama, H., Minamishima, Y. A., Ongusaha, P. P., Deng, C., et al. (2004). BRCA1 phosphorylation by Aurora-A in the regulation of G2 to M transition. *The Journal of biological chemistry*, *279*(19), 19643–8. doi:10.1074/jbc.M311780200

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, *27*(8), 1160–7. doi:10.1200/JCO.2008.18.1370

Parkin, D. M., Pisani, P., & Ferlay, J. (1993). Estimates of the worldwide incidence of eighteen major cancers in 1985. *International Journal of Cancer*, *54*(4), 594–606. doi:10.1002/ijc.2910540413

Perou, C M, Sørlie, T., Eisen, M. B., Van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., et al. (2000). Molecular portraits of human breast tumours. *Nature*, *406*(6797), 747–52. doi:10.1038/35021093

Perou, Charles M, & Børresen-Dale, A.-L. (2011). Systems biology and genomics of breast cancer. *Cold Spring Harbor perspectives in biology*, *3*(2). doi:10.1101/cshperspect.a003293

Phelan, C. M., Lancaster, J. M., Tonin, P., Gumbs, C., Cochran, C., Carter, R., Ghadirian, P., et al. (1996). Mutation analysis of the BRCA2 gene in 49 site-specific breast cancer families. *Nature genetics*, *13*(1), 120–2. doi:10.1038/ng0596-120

Richardson, A. L., Wang, Z. C., De Nicolo, A., Lu, X., Brown, M., Miron, A., Liao, X., et al. (2006). X chromosomal abnormalities in basal-like human breast cancer. *Cancer cell*, *9*(2), 121–32. doi:10.1016/j.ccr.2006.01.013

Schneider, B. P., Winer, E. P., Foulkes, W. D., Garber, J., Perou, C. M., Richardson, A., Sledge, G. W., et al. (2008). Triple-negative breast cancer: risk factors to potential targets. *Clinical cancer research : an official journal of the American Association for Cancer Research*, *14*(24), 8010–8. doi:10.1158/1078-0432.CCR-08-1208

Senn, S., & Richardson, W. (1994). The first t-test. *Statistics in medicine*, *13*(8), 785–803. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8047737

Slamon, D., Godolphin, W., Jones, L., Holt, J., Wong, S., Keith, D., Levin, W., et al. (1989). Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science*, *244*(4905), 707–712. doi:10.1126/science.2470152

Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A., & Mcguire, W. L. (n.d.). Correlation Amplification, (21).

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(14), 8418–23. doi:10.1073/pnas.0932692100

Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., et al. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(18), 10393–8. doi:10.1073/pnas.1732912100

Srikantan, S., Abdelmohsen, K., Lee, E. K., Tominaga, K., Subaran, S. S., Kuwano, Y., Kulshrestha, R., et al. (2011). Translational control of TOP2A influences doxorubicin efficacy. *Molecular and cellular biology*, *31*(18), 3790–801. doi:10.1128/MCB.05639-11

TCF3 transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47) [Homo sapiens] - Gene - NCBI. (n.d.). Retrieved February 21, 2013, from http://www.ncbi.nlm.nih.gov/gene/6929

TOP2A topoisomerase (DNA) II alpha 170kDa [Homo sapiens] - Gene - NCBI. (n.d.). Retrieved February 21, 2013, from http://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=ShowDetailView&TermToSearch=7153

Usary, J., Llaca, V., Karaca, G., Presswala, S., Karaca, M., He, X., Langerød, A., et al. (2004). Mutation of GATA3 in human breast tumors. *Oncogene*, *23*(46), 7669–78. doi:10.1038/sj.onc.1207966

Van den Brandt, P. A. (2000). Pooled Analysis of Prospective Cohort Studies on Height, Weight, and Breast Cancer Risk. *American Journal of Epidemiology*, *152*(6), 514–527. doi:10.1093/aje/152.6.514

Wang, L., Xiang, J., Yan, M., Zhang, Y., Zhao, Y., Yue, C., Xu, J., et al. (2010). The mitotic kinase Aurora-A induces mammary cell migration and breast cancer metastasis by activating the Cofilin-F-actin pathway. *Cancer research*, *70*(22), 9118–28. doi:10.1158/0008-5472.CAN-10-1246

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, *34*(1/2), 28–35.

WHO | World Health Organization. (n.d.). Retrieved from http://www.who.int/en/

Wu, Z.-S., Wu, Q., Wang, C.-Q., Wang, X.-N., Wang, Y., Zhao, J.-J., Mao, S.-S., et al. (2010). MiR-339-5p inhibits breast cancer cell migration and invasion in vitro and may be a potential biomarker for breast cancer prognosis. *BMC cancer*, *10*(1), 542. doi:10.1186/1471-2407-10-542

Zhang, H., Hu, J., Recce, M., & Tian, B. (2005). PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic acids research*, *33*(Database issue), D116–20. doi:10.1093/nar/gki055

Ziegler, R. G., Hoover, R. N., Pike, M. C., Hildesheim, A., Nomura, A. M. Y., West, D. W., Wu-Williams, A. H., et al. (1993). Migration Patterns and Breast Cancer Risk in Asian-American Women. *JNCI Journal of the National Cancer Institute*, *85*(22), 1819–1827. doi:10.1093/jnci/85.22.1819

Zlotorynski, E., & Agami, R. (2008). A PASport to cellular proliferation. *Cell*, *134*(2), 208–10. doi:10.1016/j.cell.2008.07.003

TEZ FOTOKOPİ İZİN FORMU

ENSTİTÜ

Fen Bilimleri Enstitüsü ☐

Sosyal Bilimler Enstitüsü ☐

Uygulamalı Matematik Enstitüsü ☐

Enformatik Enstitüsü ☐

Deniz Bilimleri Enstitüsü ☐

YAZARIN

Soyadı : .....
Adı : .....
Bölümü : .....

TEZİN ADI (İngilizce) : .....
.....
.....
.....
.....

TEZİN TÜRÜ : Yüksek Lisans ☐ Doktora ☐

1. Tezimin tamamı dünya çapında erişime açılsın ve kaynak gösterilmek şartıyla tezimin bir kısmı veya tamamının fotokopisi alınsın. ☐
2. Tezimin tamamı yalnızca Orta Doğu Teknik Üniversitesi kullancılarının erişimine açılsın. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.) ☐
3. Tezim bir (1) yıl süreyle erişime kapalı olsun. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.) ☐

Yazarın imzası ..... Tarih .....