ALTERNATIVE POLYADENYLATION DEPENDENT 3'-UTR CHANGES IN TRIPLE NEGATIVE BREAST CANCERS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

TANER TUNCER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
BIOTECHNOLOGY

DECEMBER 2014

Approval of the Thesis:

ALTERNATIVE POLYADENYLATION DEPENDENT 3'-UTR CHANGES IN TRIPLE NEGATIVE BREAST CANCERS

Submitted by **TANER TUNCER** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Department of Biotechnology, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Filiz Bengü Dilek _____
Head of Department**, Biotechnology**

Assoc. Prof. Dr. A. Elif Erson-Bensan _____
Supervisor, **Biological Sciences Dept., METU**

Assoc. Prof. Dr. Tolga Can _____
Co-Supervisor, **Computer Engineering Dept., METU**

Examining Committee Members:

Assoc. Prof. Dr. Özlen Konu _____
Molecular Biology and Genetics Dept., Bilkent University

Assoc. Prof. Dr. A.Elif Erson-Bensan _____
Biological Sciences Dept., METU

Prof. Dr. İnci Togan _____
Biological Sciences Dept., METU

Assoc. Prof. Dr. Sreeparna Banerjee _____
Biological Sciences Dept., METU

Assoc. Prof. Dr. Mesut Muyan _____
Biological Sciences Dept., METU

Date: 24.12.2014

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: Taner Tuncer

Signature:

ABSTRACT

ALTERNATIVE POLYADENYLATION DEPENDENT 3'-UTR CHANGES IN TRIPLE NEGATIVE BREAST CANCERS

Tuncer, Taner
PhD, Department of Biotechnology
Supervisor: Assoc. Prof. Dr. A. Elif Erson-Bensan
Co-Supervisor: Assoc. Prof. Dr. Tolga Can

December 2014, 117 pages

Alternative polyadenylation (APA) may cause mRNA 3'-UTR (untranslated region) length changes in different physiological conditions and/or disease states. About half of mammalian genes use alternative polyadenylation to generate multiple mRNA isoforms differing in their 3′-UTRs. APA has been attracting attention due to roles in gene expression regulation by shortening or lengthening of 3'-UTRs upon proliferative signals. Triple Negative Breast Cancers (TNBCs) are aggressive subtypes of breast cancers with poor prognosis and high proliferation abilities. While microarray gene expression analyses provided vital information on breast cancer subtypes, conventional quantification of microarrays underestimate potential isoforms generated by alternative splicing and APA. Due to high proliferative indices; we hypothesized APA to play a role in TNBCs. We developed and used a probe based meta-analysis tool (APADetect). APADetect uses poly(A) site positions to group Affymetrix probes as proximal and distal sets. Here, 3-'UTR shortening or lengthening events were detected using APADetect in 520 TNBC and 82 normal breast samples. When compared to normal breast tissue, TNBC patients had

significantly different APA patterns. In TNBCs, 68.5% of APA events (113 of 165) were 3'-UTR shortening and 31.5% (52 of 165) were 3'-UTR lengthening. Transcripts going through APA were mostly related to metabolic and transcriptional processes, consistent with a high proliferative state. Significantly altered 3'-UTR lengths in TNBCs were then correlated with patient characteristics, and compared to ER+ patients. Significantly shortened mRNAs were predicted to harbor fewer microRNA binding sites compared to their longer versions. Such 3'-UTR shortening events correlated with relapse free survival times of TNBC patients. Expression analysis confirmed the *in silico* results and showed increased shorter isoform generation for USP9X, and YME1L1. Our results, for the first time, show a potential role of APA in TNBCs. APA pattern changes and generation of different 3' UTR isoforms may help us improve our understanding of TNBCs by discovering new cancer related genes which may have been overlooked in conventional microarray gene expression analyses.

Key words: alternative polyadenylation, 3'-UTR, APA, Triple Negative Breast Cancer.

ÖZ

TRİPLE NEGATİF MEME KANSERİNDE ALTERNATİF POLİADENİLASYONA BAĞLI 3'-UTR DEĞİŞİKLİKLERİ

Tuncer, Taner
Doktora, Biyoteknoloji Bölümü
Tez Yöneticisi: Doç. Dr. A. Elif Erson-Bensan
Ortak Tez Yöneticisi: Doç. Dr. Tolga Can

Aralık 2014, 117 sayfa

Alternative poliadenilasyon (APA) farklı fizyolojik durumlar ve/veya hastalıklarda mRNA'nın 3'-UTR'lerinin (translasyon olmayan bölge) uzunluklarında değişikliklere yol açabilir. Memeli genlerinin yaklaşık yarısı alternatif poliadenilasyon ile 3'-UTR boyları farklı mRNA izoformları oluşturabilirler. APA'nın proliferatif sinyaller sonucu 3'-UTR'lerin boyunu kısaltarak veya uzatarak gen regülasyonuna katkısı son zamanlarda dikkat çekmektedir. Triple Negatif Meme Kanserleri (TNMK) yüksek proliferatif özellikleri ve kötü prognozları olan agresif bir meme kanseri alt türüdür. Mikroçip gen ifade analizleri meme kanseri alt tipleri ile ilgili çok önemli bilgiler sağlamış olsa da, konvensiyonel mikroçip ölçümleri alternatif uçbirleştirme ve alternative poliadenilasyon ile ortaya çıkan izoformların önemini azımsamaktadır. Yüksek proliferasyon endeksleri nedeniyle, TNMK'lerinde APA'nun rol oynayabileceğini düşündük. Prob-tabanlı bir meta-analiz aracı (APADetect) geliştirerek kullandık. APADetect poli(A) sitelerinin pozisyonlarına

göre Affymetrix problarını yakın ve uzak setler olarak gruplayarak analiz etmektedir. Çalışmamızda, 520 TNMK örneği ile 82 normal meme örneğini APADetect ile karşılaştırarak 3'-UTR kısalma ve uzama durumları belirlenmiştir. TNMK hastaları normal meme dokusu ile karşılaştırıldığında belirgin APA motifleri ortaya çıkmıştır. TNMK'lerinde APA olaylarının %68.5'i (165'te 113'ü) 3'-UTR kısalmasıyken, %31.5'i (165'ten 52'si) 3'-UTR uzaması olarak bulunmuştur. Yüksek proliferasyon durumuyla tutarlı olarak, APA gözlemlenen transkriptlerin birçoğunun metabolik ve transkripsiyon süreçleriyle ilişkili genler olduğu ortaya çıkmıştır. TNMK'lerinde belirgin olarak değişen 3'-UTR durumları daha sonra hasta karakteristikleri ile korele edilmiş ve ER+ hastalarla karşılaştırılmıştır. Belirgin olarak kısalan 3'-UTR'lerin uzun versiyonlarına göre daha az mikroRNA bağlanma bölgesine sahip olduğu tahmin edilmiştir. TNMK hastalarında 3'-UTR kısalma olayları kanserin tekrarlama süresi ile korelasyon göstermiştir. Yapılan ifade analizleri *in silico* sonuçları doğrulayarak USP9X ve YME1L1 genlerinde kısa izoformun daha fazla üretildiğini göstermiştir. Bu çalışma ilk defa APA'nın TNMK'lerinde potensiyel bir rolü olduğunu göstermiştir. TNMK'lerini daha iyi anlamaya yönelik olarak APA motiflerinin değişimi ve 3'-UTR izoformlarının anlaşılması, konvensiyonel mikroçip gen ifadesi analizleri ile gözden kaçan kanserle ilgili genlerin ortaya çıkarılmasını sağlayacaktır.

Anahtar kelimeler: alternatif poliadenilasyon, 3'-UTR, APA, Triple Negatif Meme Kanseri

To my beloved Begüm

ACKNOWLEDGEMENTS

I would like to thank my supervisor Assoc. Prof. Dr. A. Elif Erson-Bensan for her encouragement and patience throughout my study.

I would like to express my sincere gratitude to my co-supervisor Assoc. Prof. Dr. Tolga Can. This thesis wouldn't be possible without his guidance and support.

I would like to thank all my thesis committee members, Prof. Dr. İnci Togan, Assoc. Prof. Dr. Mesut Muyan, Assoc. Prof. Dr. Sreeparna Banerjee, and Assoc. Prof. Dr. Özlen Konu.

I am particularly thankful to Assoc. Prof. Dr. Mesut Muyan for his invaluable suggestions and guidance.

I thank all Erson and Muyan lab members.

I wish to express my gratitude to my family.

Lastly, I would like to thank Begüm for her endless patience, support and help throughout my PhD studies.

I would like to thank TUBITAK for supporting me by National Scholarship Programme for PhD Students (2007-2012).

TABLE OF CONTENTS

ABSTRACT v

ÖZ……. vii

ACKNOWLEDGEMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xiv

LIST OF FIGURES xv

LIST OF ABBREVIATIONS xvi

CHAPTERS

1. INTRODUCTION 1

1.1 Alternative Polyadenylation 1

1.1.1 3'-end formation of pre-mRNA 1

1.1.1.1 Polyadenylation Mechanism 2

1.1.1.2 Types of Polyadenylation 3

1.1.2 Regulation of APA 5

1.1.2.1 *Cis*-sequences Controlling APA 5

1.1.2.2 Protein Regulators of APA 7

1.1.2.3 APA and Chromatin 8

1.1.2.4 Interplay between APA and Transcription 8

1.2 Affymetrix Microarray Chip Design 9

1.3 Triple Negative Breast Cancer (TNBC) 11

1.3.1 Classification of Breast Cancer 11

1.3.2 TNBC 12

1.3.3 TNBC and microRNAs....12
1.4 Aim of the Study....16
2. MATERIALS AND METHODS....17
2.1 Microarray datasets....17
2.2 APADetect....18
2.2.1 Probe level microarray data analysis....18
2.2.2 Filter Incorporation to APADetect....18
2.2.2.1 Size filter....19
2.2.2.2 Outlier probe filter....19
2.2.2.3 Outlier sample filter....20
2.2.2.4 Distal filter....20
2.2.3 APADetect User Interface....21
2.3 Significance Analysis of Microarrays (SAM)....21
2.4 Statistical Analyses....22
2.4.1 SLR plots....23
2.4.2 ROC Curve Analyses....23
2.4.3 Patient Characteristics Correlation Analyses....23
2.4.4 Relapse Free Survivals....24
2.5 microRNA predictions....24
2.5.1 Circos plots....25
2.6 Ontology Analysis....25
2.7 Expression Analysis....26
2.7.1 Cell lines....26
2.7.2 RNA Isolation....26
2.7.3 DNA contamination, Quantification and Integrity Assessment....27
2.7.4 cDNA synthesis....28

2.7.5 RT-qPCR Analysis 29
3. RESULTS AND DISCUSSION 31
3.1 APA Events in TNBCs 31
3.1.1 3'-UTR Shortening Events in TNBCs 34
3.1.1.1 Patient characteristics correlation analysis 38
3.1.1.2 Negative regulatory sites on 3'-UTRs 42
3.1.1.3 Expression Analysis 44
3.1.2 3'-UTR Lengthening Events in TNBCs 49
3.1.3 Constant 3'-UTR lengths 56
3.2 APA Patterns Comparison between TNBCs and ER+ Tumors 58
4. CONCLUSION 61
REFERENCES 67
APPENDICES
A. Patient Characteristics 79
B. APADetect Results 97
C. SAM Results 103
D. GOTERM IDs and p Values 109
E. DNA Contamination and Quantification Assessment 111
F. Primers 113
CURRICULUM VITAE 115

LIST OF TABLES

TABLES

Table 1.1. TNBC related miRNAs.....15

Table 2.1. Reaction Conditions of DNase Treatment.....28

Table 2.2. Reaction Conditions of Reverse Transcription.....29

Table 3.1. The 3'UTR shortening genes in TNBCs.....36

Table 3.2. Correlation between patient characteristics and 3'-UTR shortening.....39

Table 3.3. The 3'UTR lengthening genes in TNBCs.....51

Table 3.4. Correlation between patient characteristics and 3'-UTR lengthening.....53

Table A.1. Characteristics of TNBC dataset.....79

Table A.2 Control dataset information.....93

Table A.3 Characteristics of ER+ dataset.....94

Table B.1. 3'-UTR shortening genes in TNBC.....97

Table B.2. 3'-UTR lengthening genes in TNBC.....99

Table B.3. Constant SLR genes in TNBC.....101

Table C.1. SAM Results.....103

Table F.1. Primers used in the study.....113

LIST OF FIGURES

FIGURES

Figure 1.1. Polyadenylation machinery.....3

Figure 1.2. Polyadenylation and alternative polyadenylation events.....4

Figure 1.3. Affymetrix probe design.....10

Figure 2.1. Workflow chart…….....22

Figure 3.1. SAM analysis of SLR events in TNBCs compared to controls.....32

Figure 3.2. Ontology analysis of significant genes.....33

Figure 3.3. Scatter plots of transcripts undergoing 3'-UTR shortening in TNBCs.....35

Figure 3.4. ROC curve analysis of transcripts undergoing 3'-UTR shortening.....37

Figure 3.5. Kaplan-Meier curves.....41

Figure 3.6. Circos plot.....43

Figure 3.7. 3'-UTRs of USP9X, YME1L1, and SNX3.....46

Figure 3.8. Relative quantification of short and long USP9X isoforms.....47

Figure 3.9. Relative quantification of short and long YME1L1 isoforms.....48

Figure 3.10. Scatter plots of transcripts undergoing 3'-UTR lengthening in TNBCs.50

Figure 3.11. Kaplan-Meier curves.....55

Figure 3.12. Scatter plots of transcripts with constant 3'-UTR lengths.....57

Figure 3.13. Scatter plots of transcripts undergoing 3'-UTR shortening in TNBC and ER+ patients.....59

Figure E.1. Lack of DNA contamination in RNA samples was assessed……111

Figure E.2. RNA concentrations were determined using NanoDrop ND1000…….112

LIST OF ABBREVIATIONS

A	Adenosine
ANOVA	Analysis of variance
APA	Alternative polyadenylation
ARE	AU-rich elements
bp	Base pairs
CEL	CIMFast Event Language
cDNA	Complementary Deoxyribonucleic Acid
Cq	Quantification cycle
CPSF	Cleavage and polyadenylation-specific factor
CSTF	Cleavage stimulation factor
DNA	Deoxyribonucleic Acid
D-pA	Distal poly(A)
Dnase I	Deoxyribonuclease I
dNTP	Deoxyribonucleotide triphosphate
ER	Estrogen receptor
EST	Expressed sequence tag
FBS	Fetal bovine serum
GEO	Gene Expression Omnibus
IQCK	IQ domain containing protein K
MIQE	Minimum information for publication of quantitative real-

	time PCR experiments
miRNA	microRNA
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
PCR	Polymerase chain reaction
P-pA	Proximal poly(A)
Pri-miRNA	Primary microRNA
Pre-miRNA	Precursor microRNA
Pre-mRNA	Precursor messenger RNA
RACE	Rapid amplification of cDNA ends
RBP	RNA binding protein
RPL13	Ribosomal protein L13
RT-qPCR	Real Time Quantitative Polymerase Chain Reaction
RNA	Ribonucleic acid
Rnase	Ribonuclease
SDHA	Succinate dehydrogenase
SNX3	Sortin nexin3
USP9X	Ubiquitin specific peptidase 9, X-linked
UTR	Untranslated Region
YME1L1	YME1 like protein 1

CHAPTER 1

INTRODUCTION

1.1 Alternative Polyadenylation

1.1.1 3'-end formation of pre-mRNA

Maturation of pre-mRNAs to mRNAs takes place mostly co-transcriptionally by addition of a 7-methyl guanosine cap to the 5'-end, removal of intronic sequences by splicing, endonucleolytic cleavage at the 3'-end, and polyadenylation.

With the exception of histone replication-dependent transcripts, most pre-mRNAs acquire a polyadenylated tail after cleavage. Poly(A) tails have a limited length that can differ between species. For instance, in yeast the length of poly(A) tail is 70-80 adenines, while in humans it is 250–300 adenines on average. As mRNAs with short tails are more prone to enzymatic degradation (Guhaniyogi and Brewer 2001) or stored in a translationally dormant state (D'Ambrogio, Nagaoka et al. 2013). The length of poly(A) tails is important. Poly(A) tail has impact on mRNA characteristics, such as stability, translation, and transport (Jacobson and Peltz 1996; Wickens, Anderson et al. 1997; Garneau, Wilusz et al. 2007; Ji, Kong et al. 2011). Maturation of pre-mRNA to mRNA is a very rapid event, since polyadenylation happens almost synchronously with transcription (Proudfoot, Furger et al. 2002; Richard and Manley 2009; Licatalosi and Darnell 2010; Ji, Kong et al. 2011).

1.1.1.1 Polyadenylation Mechanism

The mature 3'-ends of mRNAs are created by a two-step reaction that involves an endonucleolytic cleavage of the pre-mRNA, followed by synthesis of a polyadenylate (polyA) tail onto the upstream cleavage product. First, polyadenylation factors cleaves the pre-mRNA at a specific site. Next, an adenosine tail is synthesized by the poly(A) Polymerase (PAP). Poly(A) binding protein (PABP) stimulates PAP to catalyze the addition of adenosine residues and regulates the length of the poly(A) tail (Kuhn, Gundel et al. 2009).

Eukaryotic polyadenylation signals generally contain a 6 nucleotide motif called poly(A) signal (PAS), which is a canonical form of AAUAAA (more than 10 variants) (Proudfoot 2011). PAS is mostly located ~10-35 nucleotides upstream of the actual cleavage site. In addition to the PAS, nearby U- or GU-rich downstream sequence elements (DSEs) and less well-defined upstream sequence elements (USEs) enhance cleavage efficiency. Recognition of these signal sequences and addition of the poly(A) tail requires numerous protein factors (Figure 1.1). The core molecular machinery responsible for 3'-end formation includes four multi-subunit protein complexes, CPSF (Cleavage and Polyadenylation Specificity Factor), CstF (Cleavage stimulation Factor), CFI (cleavage factor I) and CFII (cleavage factor II). CPSF consists of six polypeptides, named CPSF1, CPSF2, CPSF3, CPSF4, FIP1L1 (FIP1-Like 1) and WDR33 (WD repeat domain 33), where CPSF1 is the subunit that recognizes the PAS. CSTF consists of three subunits; CSTF1, CSTF2 and CSTF3.

Additional factors required for cleavage and polyadenylation process are poly(A) polymerase (PAP) (Wahle 1991; Shi, Di Giammartino et al. 2009), the scaffold protein simplekin and CFIm (cleavage factor Im) and CFIIm. CFIm is a tetramer composed of two subunits (namely, CFIm25 and CPSF6) that bind the UGUA motifs upstream of the cleavage site and two larger polypeptides (CFIm59 and CFIm72) (Yang, Coseno et al. 2011).

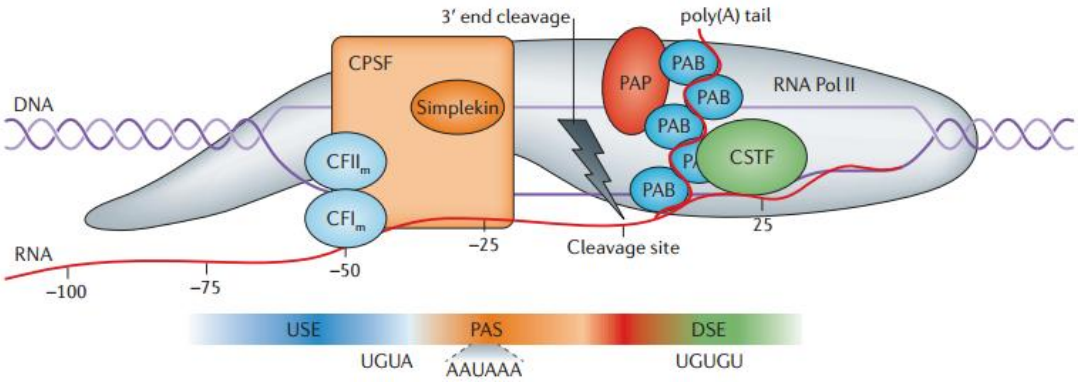

Figure 1.1. Polyadenylation machinery. Core proteins involved in cleavage and polyadenylation are depicted. Figure is taken from (Elkon, Ugalde et al. 2013).

1.1.1.2 Types of Polyadenylation

Accumulating evidence shows alternative polyadenylation (APA) to emerge as a widespread mechanism in gene expression regulation. More than 50% of human genes have multiple poly(A) signal sites, leading to multiple cleavage sites (Tian, Hu et al. 2005). Hence, alternative sites of polyadenylation may generate a 3'-UTR diversity of transcripts.

Polyadenylation events can be classified into three general groups (Figure 1.2). In the first group, there is only one PAS in the 3'-UTR, and this signal site leads to only one mRNA isoform and one type of protein product. In the second group, multiple polyadenylation sites (PASs) are present in the 3'-UTR and these signal sites give rise to different poly(A) tail positions in the mature mRNA 3'-UTR isoforms generating multiple isoforms that differ in their 3′UTR length without affecting the protein encoded by the gene. In the third and fourth groups, alternative splicing is coupled to alternative polyadenylation. In these cases, alternative PAS may exist in exons or introns. In this case, depending on the location of the stop codon different protein products may or may not be produced. Except the first group, other three

groups are considered as APA events. Through these groups, APA contributes to the complexity of the transcriptome by generating isoforms that differ either in their 3′UTRs or in their coding sequences.

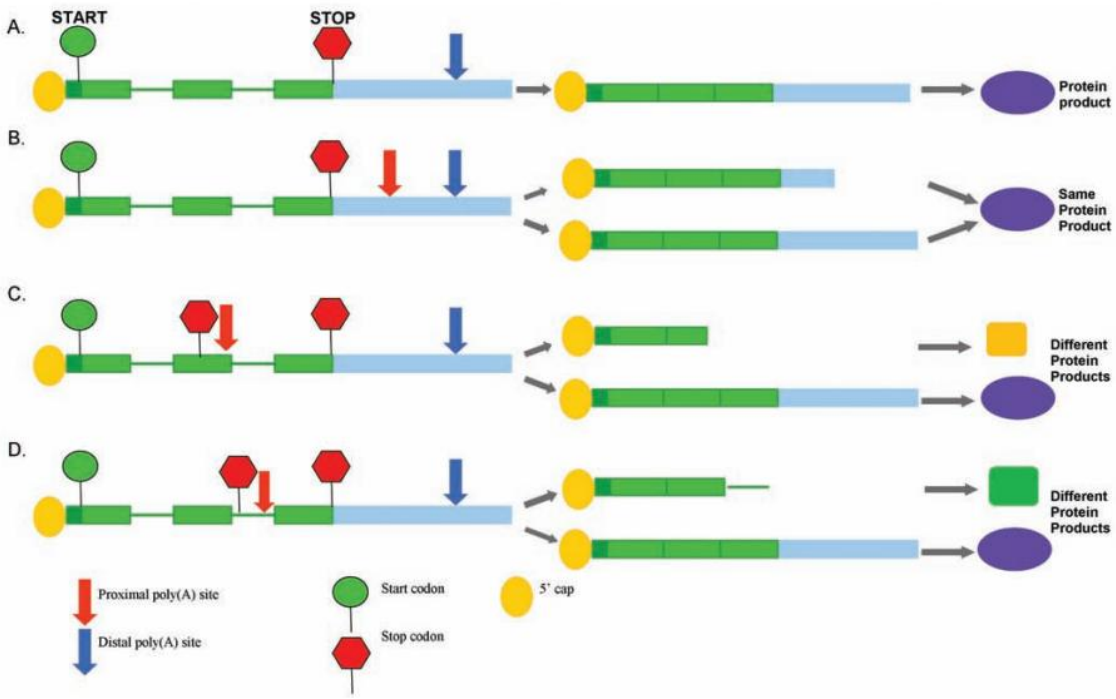

Figure 1.2. Polyadenylation and alternative polyadenylation events. (A) mRNAs with one PAS produce a single protein product. (B) mRNAs with more than one PAS in the 3'-UTR produce isoforms with different 3'-UTR lengths. Protein product of these isoforms would be the same. (C) APA coupled to alternative splicing, resulting with different mRNAs and different protein products. Alternative splicing in this case generates an isoform harboring sequences from an internal exon. (D) Alternative splicing generates an isoform from an intron, causing proximal polyadenylation, giving rise to a different protein product. Green boxes show exons, green lines show introns, light blue boxes show 3'-untranslated regions (3'-UTR's). Figure is taken from (Akman and Erson-Bensan 2014).

1.1.2 Regulation of APA

Extensive APA modulation in different physiological conditions indicates that PAS selection is under active control. The choice of using one polyadenylation site over another is determined by a combination of various features such as *cis*-acting RNA elements, trans-acting regulator proteins, kinetic coupling with transcription machinery and epigenetic modifications.

1.1.2.1 *Cis*-sequences Controlling APA

The RNA sequence elements; *cis*-sequences involved in polyadenylation can be classified as core and auxiliary elements. Core elements include a core upstream element (AAUAAA or other variants) and a core downstream element (U- or GU-rich sequences). In ~53% of human pre-mRNAs poly(A) signal is AAUAAA. However, usage of other proximal hexamer variants is common (Tian, Hu et al. 2005; Jan, Friedman et al. 2011), and are often associated with APA (Nunes, Li et al. 2010). U- or GU-rich elements serve as the binding site for CstF-64. Single nucleotide variants of U- or GU-rich elements often correlate with weaker polyadenylation signals.

Auxiliary elements often can be found further upstream and downstream of the polyadenylation cleavage site. These elements can enhance weak polyadenylation signals and may help determine the polyadenylation site chosen when multiple polyadenylation signal choices are present. The best characterized auxiliary upstream elements (AUE) are U-rich sequences called upstream sequence elements (USEs). One example is present in the COX-2 3'-UTR which has two polyadenylation signals. The proximal signal contains a noncanonical AUUAAA element and is weaker than the distal polyadenylation signal (Hall-Pogar, Zhang et al. 2005). The COX-2 proximal signal also contains three USEs. When these USEs are mutated, the usage of the weaker, proximal polyadenylation signal decreases. The USE closest to

the polyadenylation cleavage site had the most effect on polyadenylation (Hall-Pogar, Liang et al. 2007).

Another *cis*-sequence, an UGUA element has been described as a binding site for the CFIm polyadenylation factors; which is found upstream of the core elements (Venkataraman, Brown et al. 2005). UGUA recognition by the CFIm complex suggested a possible mechanism for CFIm mediated APA (Venkataraman, Brown et al. 2005; Yang, Gilmartin et al. 2011).

One of the best characterized auxiliary downstream element (ADE) is a G-rich sequence (GRS). When this 14 base long sequence deleted, 3'-end processing reduced by 75–80%. *In vitro* studies showed that GRS influenced the efficiency of 3'-end processing by stabilizing the assembly of CstF on the core downstream U-rich element. Hence, it serve as an integral part of the polyadenylation signal and can affect signal strength and regulation which will affect polyadenylation site choice eventually (Chen and Wilusz 1998).

Another novel study determined three new sequence patterns close to poly(A) sites by RNA sequencing analysis: a TTTTTTTT motif is positioned ~21 nt upstream of the poly(A) site, an AAWAAA motif (W denoting A or T) is positioned upstream of the poly(A) site, and a palindromic CCAGSCTGG sequence (S denting C or G) is found downstream of the pol(A) site (Ozsolak, Kapranov et al. 2010). These new patterns are present in intragenic and intergenic poly(A) sites, however they don't co-exist with the canonical AATAAA signal.

Lastly, secondary structures and stem-loop motifs in 3'-UTRs have been shown to regulate stability of mRNA (Erlitzki, Long et al. 2002). It is proposed that such structures could enhance or inhibit possible 3'-processing proteins to modulate APA (Di Giammartino, Nishida et al. 2011). Also, it has been shown that RNA structure has an important role in polyadenylation choice, when complementary upstream and downstream elements are present, it affects the polyadenylation choice as a result of secondary RNA structure (Darmon and Lutz 2012).

1.1.2.2 Protein Regulators of APA

Global changes in activity and/or levels of 3'-end core machinery factors have been linked to APA regulation in various studies. One of the earliest studies showed the link between CSTF2 upregulation and poly(A) site selection during B cell maturation (Takagaki and Manley 1998). Immunoglobulin M (IgM) has two APA sites; one intronic (proximal) and one in 3'-UTR (distal). Usage of the intronic poly(A) site generates an mRNA that encodes a secreted form of IgM, whereas cleavage at the distal site produces a membrane-bound IgM. During the early stages of B cell development, the canonical and stronger IgM poly(A) site was used, while in activated B cells the weaker proximal site was favored. It was revealed that this regulated poly(A) selection is controlled by CSTF2 levels which is upregulated when B cells are activated. A recent transcriptome-wide study with HeLa cells showed the increase in distal site selection when CSTF2 is depleted (Yao, Biesinger et al. 2012). Unlike CSTF2, the opposite effect was observed for CFIm, when CFIm levels are depleted proximal site usage increased in HEK293 cells (Gruber, Martin et al. 2012; Martin, Gruber et al. 2012). Overall, CSTF2 and CFIm have not only a role in 3'-end processing core machinery, but also have a regulatory role in poly(A) site selection.

Transcription factors, E2F1 and E2F2, which are involved in cell cycle regulation have been linked to poly(A) site choice by transcriptional up-regulation of 3′-end-processing genes during proliferative states (Elkon 2012).

Besides 3'-processing factors, RNA binding proteins have been found to be associated with APA control. For instance, levels of Polyadenylation binding protein nuclear 1 (PABPN1) has been found to be important for poly(A) site selection (de Klerk, Venema et al. 2012; Jenal, Elkon et al. 2012). When PABPN1 levels are decreased, selection switches towards proximal sites.

1.1.2.3 APA and Chromatin

Recent studies showed that chromatin and epigenetics modifications may have an effect on 3'-end processing events. For instance, a particular study showed that poly(A) sites are strongly depleted of nucleosomes, while a nucleosomal enrichment was observed at downstream regions of these sites (Spies, Nielsen et al. 2009). As poly(A) sites are A- and T- rich DNA stretches, nucleosomes have a low affinity for these sites. When genes with multiple poly(A) sites were examined, poly(A) sites with nucleosome depletion at the site and nucleosome enrichment at downstream from it were identified as favorable. This phenomenon suggests that nucleosome positioning might effect poly(A) site selection.

Regulation of APA by genomic imprinting has been studied in mouse, showing that transcriptional interference affected poly(A) site choice (Wood, Schulz et al. 2008).

Mcts2 and *Napl15* are retrogenes which are located within the introns of their host genes; *H13* and *Herc3*, respectively (Wood 2007; Cowley, Wood et al. 2012). Promoters of the *Mcts2* and *Napl15* are imprinted and therefore silenced on the maternal allele; yet they are unmethylated and active on the paternal allele. Host gene uses an upstream intronic poly(A) site when the retrogene is transcribed, however when retrogene is methylated, a downstream distal poly(A) site is used by the host gene. It appears the complex chromatin and epigenetic modifications at/near the polyadenylation sites are important for APA.

1.1.2.4 Interplay between APA and Transcription

The functional and physical coupling between 3′-end formation and transcription has been known for a long time. However a better understanding is required to elucidate the significance of this link on regulation of APA as the accumulating evidence indicates that transcriptional activity impacts poly(A) choice.

The effects of kinetics parameters such as the rate of RNA polymerase II (RNA pol II)-dependent transcription on the structure of mRNAs are just beginning to be elucidated. As the proximal poly(A) sites are transcribed first and are therefore encountered first by the 3′-endprocessing machinery, they have an advantage for being used over distal poly(A) sites (Shi 2012). This phenomenon has been studied in a *D. melanogaster* strain which has a low transcription elongation rate and results suggested that reduced RNA pol II kinetics leads to increased proximal poly(A) site usage (Pinto 2011).

Yet another interesting finding showed that highly transcribed genes tend to use proximal poly(A) sites and lowly transcribed genes use distal sites (Ji, Luo et al. 2011). Further evidence showed that strong transcriptional activators recruit the PAF1c component of the transcription elongation complex to the promoter, so that PAF1c recruits 3'-end processing complex and promotes proximal polyadenylation (Nagaike, Logan et al. 2011).

1.2 Affymetrix Microarray Chip Design

APADetect analyzes Affymetrix Human Genome U133A Arrays (HGU133A, NCBI GEO Accession number: GPL96) and Affymetrix Human Genome U133 Plus 2.0 Arrays (HGU133Plus2, NCBI GEO Accession number: GPL570) by using the available chip design data such as the consensus sequences of a transcript and probe recognition target sites on the transcripts (Lockhart, Dong et al. 1996).

Affymetrix HG133A and HGU133Plus2 series have been widely used in gene expression studies since their release in 2001. Microarray (also known as chip or biochip) is a collection of DNA spots (probe) attached to a solid surface. A probe is a single-stranded DNA oligonucleotide synthesized directly on the surface of the chip using photolithography and combinatorial chemistry. This 25 base oligonucleotide is designed to be complementary to a specific gene transcript. A set of probes designed to detect one transcript are called a probe set. A probe set in Affymetrix HG133A and HGU133Plus2 series consists of 11 probe pairs. An 11 probe pair set is made up

of 11 perfect match (PM) probes and 11 mismatch (MM) probes for a total of 22 probe cells. PM probes are designed to be complementary to a reference sequence while MM probes are designed to be complementary to a reference sequence except for a homomeric mismatch at the central position (e.g., 13th position of 25 base probe, T->A or C->G). Mismatch probes serve as a control for cross-hybridization. This helps to determine the background and nonspecific hybridization that contributes to the signal measured for the perfect match oligo. Probes are chosen based on information from Genebank and other nucleotide repositories. The sequences on the Affymetrix HG133A and HGU133Plus2 arrays are designed to recognize unique regions of the 3' end of the gene as shown in Figure 1.3.

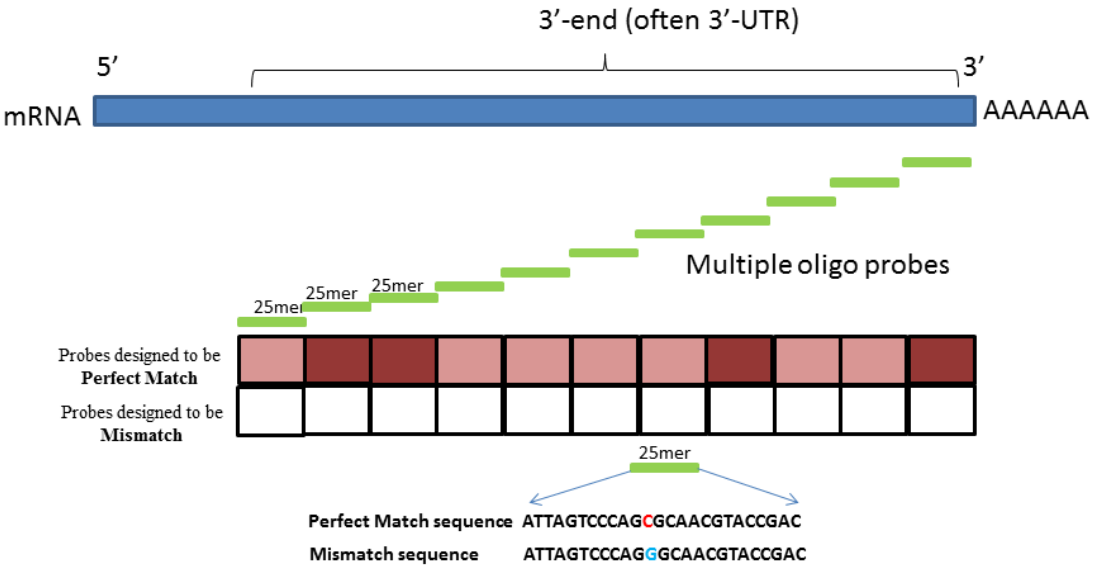

Figure 1.3. Affymetrix probe design. A probe set is made up of 11 perfect match (PM) probes and 11 mismatch (MM) probes. These probes are often designed to recognize the 3'-end of the mRNA.

Affymetrix HG133A and HGU133Plus2 series both have 11 probes in each probe set which are scattered on the chip. Signals acquired from multiple probes prevent technical errors which may result from inefficient hybridization and/or washing steps of experimental procedure. Quantification of a transcript is done by calculating the mean of these 11 probes. Despite the fact that probe sets were organized by Affymetrix, it can be re-organized into custom sub-sets. There are 2411 and 3683 unique probe sets for a total of 2066 and 3067 genes on the HGU133A and HGU133Plus2 platforms, respectively.

1.3 Triple Negative Breast Cancer (TNBC)

1.3.1 Classification of Breast Cancer

Breast cancer is a heterogeneous and complex disease, involving multiple tumor entities associated with distinguishing histological patterns and different biological features and clinical behaviors. The classical breast cancer classification systems were based on the histological appearances. Various classification systems have been developed without complete agreement between them (Weigelt and Reis-Filho 2009). The 4^{th} edition of the World Health Organization classification recognizes more than 21 subtypes of invasive breast carcinoma (Lakhani 2014; Vuong, Simpson et al. 2014). However, diagnoses based on histological criteria are rather subjective, and information on histological subtype has a limited impact on therapeutic decisions. Current classification of breast cancer subtypes based on clinicopathological parameters, including histological grade, presence of lymph node metastasis, and lympho-vascular invasion. In addition, predictive biomarkers such as expression of estrogen receptor (ER), progesterone receptor (PR) and (HER2), have proven to be clinically useful.

In recent years, heterogeneity within the subtypes of breast cancer has been of great interest. Since the increasing use of high-throughput gene expression profiling tools

such as gene expression microarrays, it is now well established that breast cancer comprises several clinically and biologically distinct subtypes.

1.3.2 TNBC

TNBC is defined by a lack of expression of the estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). TNBC represent approximately 10-17% of all breast cancers (Carey, Dees et al. 2007). TNBC patients are generally younger than the overall population of breast cancer patients with larger and higher-grade tumors. TNBC is also associated with a higher risk of distant recurrence and death, especially within 3 years of diagnosis. Owing to the lack of hormone-receptor expression, TNBC does not respond to hormonal therapy or HER2-targeted agents, leaving cytotoxic chemotherapy as the only option for systemic therapy.

There is a need for better therapeutic options for patients with TNBC, ideally in the form of targeted agents. For this purpose, better understanding of molecular mechanisms underlying the pathological outcome is vital. This can be achieved not only by genomic profiling but also by deciphering prost-transcriptional regulations affecting cancer proteome.

1.3.3 TNBC and microRNAs

APA and microRNA (miRNA) regulation are both post-transcriptional gene regulation mechanisms. APA often results in mRNA isoforms with the same coding sequence but different lengths of 3'-UTRs, while miRNAs regulate gene expression by binding to specific mRNA 3' –UTRs (Chen and Shyu 1995; Barreau, Paillard et al. 2005).

MiRNAs are ~22 nucleotides long, endogenous, small non-coding single-stranded RNAs, found in both animals and plants. MiRNAs function as negative modulators

of gene expression at post-transcriptional level. 30% of human protein-coding genes are estimated to be regulated by miRNAs (Filipowicz, Bhattacharyya et al. 2008). MiRNAs have been implicated in various biological processes including proliferation, differentiation, apoptosis, and development (Ambros 2004; Bartel 2004; Filipowicz, Bhattacharyya et al. 2008). MiRNAs modulate gene expression either by degradation and/or destabilization of mRNA.

Most miRNA genes are located in intergenic regions but they can also be found within exonic or intronic regions in either sense or antisense orientation. MiRNAs are mostly transcribed by RNA pol II from their own promoter or from promoter of the host gene in which they reside. RNA pol II synthesized large miRNA precursors called primary-miRNAs (pri-miRNAs) (Lee, Kim et al. 2004), which contain both 5′-cap structure (7MGpppG) as well as 3′-end poly(A) tail (Cai, Hagedorn et al. 2004). Clustered miRNAs might be transcribed from a single transcription unit as polycistronic primary-miRNA.

MiRNA biogenesis can be summarized in two central processing steps. First, pri-miRNAs are processed in the nucleus by RNAse III Drosha, associated to a double stranded RNA-binding protein DGCR8 (DiGeorge syndrome critical region gene 8; aka as Pasha in flies) known as the microprocessor complex, that generates ~70 nucleotides precursor miRNA products, which locally fold into stable secondary stem-loop structures (Lee, Ahn et al. 2003). Ran-GTP-dependent transporter Exportin 5 recognizes the short stem plus a ~2 nt 3′ overhang of the originated precursor molecules, and mediates the translocation to the cytoplasm (Yi, Qin et al. 2003). In the cytoplasm RNAse III enzyme Dicer associated to TRBP (TAR RNA-binding protein) or protein activator of the interferon induced protein kinase (also known as PRKRA), and Argonaute (AGO1-4) perform “dicing” which is the second cropping process. By cleaving the miRNA precursor hairpin, a transitory miRNA/miRNA* duplex is generated. This duplex contains the mature miRNA guide, generally selected according to thermodynamic properties, and the complementary passenger strand which is usually subjected to degradation.

Cancer is a disease involving multi-step changes in the genome (Hanahan and Weinberg 2011). Although current studies of the cancer genome mostly have focused

on protein-coding genes, alterations of functional non-coding sequences in cancer is another field of interest to understand the cancer genome (Calin and Croce 2006; Esquela-Kerscher and Slack 2006; Calin, Liu et al. 2007). Several studies showed that expression of miRNAs is widely deregulated in cancer in the last decade (Calin and Croce 2006; Esquela-Kerscher and Slack 2006). In human cancers, miRNAs can act as either oncogenes (He, Thomson et al. 2005; O'Donnell, Wentzel et al. 2005) or tumor suppressor genes (Calin, Dumitru et al. 2002; Johnson, Grosshans et al. 2005; Mayr, Hemann et al. 2007). While the number of miRNAs (~1000) is much smaller than that of protein-coding genes (~22000), expression signature of miRNAs reflect the developmental lineage and tissue origin of human cancer more accurately (Lu, Getz et al. 2005; Rosenfeld, Aharonov et al. 2008).

MiRNAs have been found to be involved in TNBC. In a recent study with primary TNBC and normal tissues samples, the miRNA profiling revealed 116 deregulated miRNAs, among which miR-106b, the cluster miR-17/92, miR-200 family (miR-200a, miR-200b and miR-200c), miR-21 and miR-155 were the most upregulated while let-7b, let-7c, miR-126, miR-145 and miR-205 were the most downregulated (Cascione, Gasparini et al. 2013).

For instance, miR-205 is a tumor suppressor gene which is down-regulated in TNBCs (Piovan, Palmieri et al. 2012). miR-205 is normally activated by TP53, which is a tumor suppressor gene itself and is suppressed in many cancers. When miR-205 expression is evaluated in a panel of TNBC cell lines, a critical down-regulation was observed. In fact, expression of miR-205 reduced cell proliferation and clonogenic potential in vitro.

Some other miRNAs with tumor suppressor properties include miR-203, which was suggested to act by suppressing BIRC5 and LASP1 causing reduced cell proliferation and migration, and miR-200c which was found to suppress cancer cell proliferation by targeting the expression of a protein known as X-linked inhibitor of apoptosis (XIAP) (Wang, Zheng et al. 2012).

miR-21, one of the well-known oncomiRs, play a major role in TNBC. In a study with 49 primary TNBC along with 34 matched tumor-associated normal samples, a trend between high miR-21 level and poor prognosis has been described (Radojicic,

Zaravinos et al. 2011). In the same study, a significant association for miR-221 and miR-210 also been made.

A list of miRNAs which have been shown to have validated functions either as a tumor suppressor or oncomiR in TNBCs are listed in Table 1.1.

Table 1.1. TNBC related miRNAs. List of TNBC linked miRNAs are taken from (D'Ippolito and Iorio 2013).

MicroRNA	Validated target(s)	Main biological function(s) in TNBC
Tumor suppressor		
miR-200a/b	Zeb1/Zeb2, Suz 12, EphA2	Stimulation of differentiation in undifferentiated mammary epithelial cell line
miR-200c	Zeb1/Zeb2	Inhibition of EMT
	MSN; FN1	Suppression of migration
	TrkB	Reversion of anoikisis resistance
miR-205	E2F1; LAMC1	Reduction of proliferation, cell cycle and tumor growth
miR-203	BIRC5	Reduction of proliferation
	LASP1	Inhibition of migration
miR-31	WAVE3; RhoA; Radexin	Reduction of metastatic potential
	PRKCE	Induction of apoptosis and enhancement of chemo- and radiosensitivity
miR-34a	AXL	Impairment of migration
OncomiR		
miR-181a/b	Bim	Inhibition of anoikisis
	ATM	Impairment of DNA double-strand-breaks repair
MiR-146 and miR-146b-5p	BRCA1	control of BRCA1-mediated proliferation and homologous recombination
miR-182	PFN1	Inhibition of cell proliferation and invasion Induction of apoptosis

1.4 Aim of the Study

Due to the presence of miRNA and/or other regulatory elements, 3'-UTR length changes through APA has important consequences for gene expression. TNBCs are aggressive tumors with high proliferative indices and APA is known to shift to proximal poly(A) site usage upon proliferative signals. We hypothesized, 3'-UTR length changes may enlighten the basis of some of the proto-oncogene activation cases observed in TNBCs. Therefore, in this study, we investigated APA based 3'-UTR length changes in TNBC patients compared to normal breast tissue using APADetect probe-based meta-analysis software.

CHAPTER 2

MATERIALS AND METHODS

2.1 Microarray datasets

NCBI Gene Expression Omnibus (GEO) (Barrett, Troup et al. 2011) is a functional genomics data repository. 520 TNBC sample datasets were taken from GSE31519 (Rody, Karn et al. 2011). Patient information; relapse time, tumor size, lymph node status, and tumor grade status were also collected from this study. Patient characteristics are summarized in Table A.1. A total of 82 histologically normal epithelium from cancer-free prophylactic mastectomy patients and breast cancer patients were collected from different studies; 42 from GSE20437 (Graham, de las Morenas et al. 2010), 15 from GSE9574 (Tripathi, King et al. 2008), 6 from GSE6883 (Liu, Wang et al. 2007), 6 from GSE26910 (Planche, Bacac et al. 2011), 5 from GSE21422 (Kretschmer, Sterner-Kock et al. 2011), 7 from GSE3744 (Richardson, Wang et al. 2006) and 1 from GSE2361 (Ge, Yamamoto et al. 2005) for comparison. Control samples are summarized in Table A.2. Datasets for 207 ER+ patients were gathered from GSE2034 (Wang, Klijn et al. 2005). Patient characteristics of ER+ patients are summarized in Table A.3.

All datasets used in this study were either Affymetrix Human Genome U133A Arrays (HGU133A, NCBI GEO Accession number: GPL96) or Affymetrix Human Genome U133 Plus 2.0 Arrays (HGU133Plus2, NCBI GEO Accession number: GPL570). CEL files for the corresponding experiments were downloaded to use the non-normalized raw intensities of the probes.

2.2 APADetect

APADetect was developed by Dr. Tolga Can (Department of Computer Engineering, METU, Ankara). Alpha version of APADetect was developed to detect expression levels of different 3'-UTR isoforms in E2 treated MCF7 breast cancer cell line (Akman, Can et al. 2012). In that study, all GEO series were analyzed individually and their results were combined to screen for transcripts that consistently had differently expressed split probe sets. As the sample size of that study was restricted to experimental replicates, APADetect required novel features to analyze multiple patient samples in our study. Incorporation of those novel features are discussed in the 2.2.2 section.

2.2.1 Probe level microarray data analysis

While Affymetrix microarray chips were originally designed to quantify transcript levels, here, we exploited probe position design decisions. APADetect splits probe sets into proximal and distal sub-sets based on the positions of reported (Zhang, Hu et al. 2005) polyA sites according to the UCSC Genome Browser Database (Fujita, Rhead et al. 2011). For each gene, mean signal intensities of proximal and distal probe sets are calculated and used as indicators of short and long 3'-UTR isoform abundance. Patient SLR (**S**hort/**L**ong **R**atio) values are compared to control group to detect APA based 3'-UTR isoform quantity variations.

2.2.2 Filter Incorporation to APADetect

APADetect has 4 stringency filters to increase the accuracy of SLR values. When TNBC datasets were collected, APADetect alpha was used to analyze SLRs. However, while working with hundreds of datasets, improvement of APADetect to increase accuracy of outcome was needed.

A result should be both accurate and precise in order to be valid. However, precision and accuracy are completely independent. A result can be accurate but not precise (the mean of several measurements is close to the actual value but the individual measurements vary significantly), precise but not accurate (the individual measurements are close to each other but their mean is far from the actual value) neither or both (van Bakel and Holstege 2004). Filters were incorporated to APADetect in order to increase accuracy and precision of the data produced.

2.2.2.1 Size filter

Size filter excludes transcripts with only one distal or proximal probe for accuracy concerns. HGU133A and HGU133Plus2 platforms were designed to have 11 probes per gene scattered through the chip. According to the position of the polyA site, APADetect calculates mean intensity value of proximal and distal probe sub-sets. Depending on microarray design decisions, position of polyA site could split probe subsets as 5:6, 7:4, or 10:1 etc. Since evaluation of the data obtained from a single probe isn't reliable, APADetect was developed to eliminate analysis of genes which probe subsets were split 10:1 or 1:10.

2.2.2.2 Outlier probe filter

Outlier probe filter eliminates individual outlier probes in a subset using Iglewicz and Hoaglin's median based outlier detection method (Iglewicz and Hoaglin; 1993) (Xu, Iglewicz et al. 2014). For each proximal and distal subset, the median absolute deviation (MAD) is computed as the median of absolute differences of individual probe intensities from their respective medians. The modified z-score of a probe *p* is then computed as:

$$z = \frac{0.6745\ (\text{intensity of p} - \text{median})}{\text{MAD}}$$

If the z-score of a probe is greater than 3.5, it is identified as outlier and removed from subsequent analyses. When an outlier probe is removed, samples with only one proximal/distal probe are filtered out by re-application of the "size filter" at the sample level.

2.2.2.3 Outlier sample filter

Outlier sample filter applies the same outlier detection method by Iglewicz and Hoaglin (Iglewicz and Hoaglin; 1993) and filters outlier samples with SLRs that deviate significantly from their respective patient or control group median.

2.2.2.4 Distal filter

When a transcript has alternative 3'-UTR isoforms, the long isoform will hybridize with both proximal and distal region probes, while the short isoform will hybridize only with the proximal region probes. The last filter of APADetect; distal filter eliminates probe sets with distal intensities significantly higher than proximal ones as distal intensities cannot be higher than proximal intensities. Such cases were considered as degraded/truncated RNAs, hence eliminated from the analysis. Degraded/truncated forms of RNA can be resulted from technical reasons during RNA isolation, handling, and/or storage. If the 3'-end of the RNA is intact, cDNA synthesis with oligo-dT primer will lead to a short cDNA which can hybridize to distal probes on the microarray.

2.2.3 APADetect User Interface

APADetect is accessible online at http://www.ceng.metu.edu.tr/~tcan/APADetect/ and is provided as a command line Java program with source code.

The GSM numbers of TNBC patients, control samples and ER+ patients obtained from GEO were listed in Table A.1 A2 and A3, respectively. These GSM numbers were used in APADetect user interface (UI).

2.3 Significance Analysis of Microarrays (SAM)

TIGR Multiexperiment Viewer (MeV) software (Saeed, Sharov et al. 2003) was used for further analysis of APADetect output file. MeV is a flexible microarray data analysis tool for clustering, visualization, classification, statistical analysis and biological theme discovery (http://www.tm4.org/).

SAM was used to test SLRs identified by APADetect for statistical significance (Tusher, Tibshirani et al. 2001). SAM is a statistical technique for finding significant genes in a set of microarray experiments. SAM assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of different measurements to estimate the percentage of gene identified by chance, the false positive rate (FPR). By analyzing random fluctuations in the data, the signal-to-noise ratio can be decreased. In our case, SAM uses repeated permutations of the data to determine if the SLR values of any genes are significantly different than control samples.

A work flow chart is presented in Figure 2.1 to summarize the steps from data collection to obtaining SLR values.

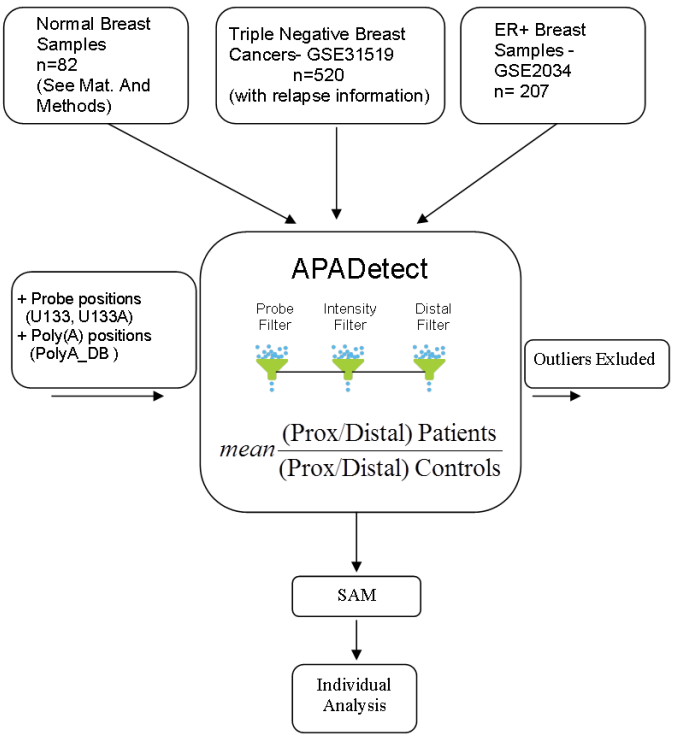

Figure 2.1. Workflow chart. The steps from data collection to SLR value determination are summarized.

2.4 Statistical Analyses

SLR values obtained by APADetect were log transformed and used for further statistical analyses. APADetect analysis presented log SLR values for each gene that can fit the criteria of the tool. SAM approach determined the significant genes. Afterwards, genes showing 3'-UTR shortening/lengthening events were individually further analyzed.

2.4.1 SLR plots

Most significant individual log SLR values were demonstrated as scatter plots by GraphPad software. The data is displayed as a collection of points, each representing an SLR of a single patient on the graphic. Significances of the data (TNBCs and controls) for each gene were further analyzed by unpaired t-test with Welch`s Correction. Welch's t-test is an adaptation of Student's t-test, and is intended for use when the two groups have possibly unequal variances.

2.4.2 ROC Curve Analyses

ROC (Receiver Operating Characteristics) curve analysis was performed for sensitivity and specificity of each significant gene (Metz 1978; Wray, Yang et al. 2010). ROC curves are beneficial for assessing the accuracy of predictions. Y-axis represents sensitivity, it shows true positive rate; while x-axis denotes specificity, showing false positive rate. The area under the curve (AUC) is used as a global indicator of diagnostic performance, ranging from 0 to 1 (Berrar and Flach 2012). ROC curve analyses were performed with GraphPad, which also reports AUC values. AUC quantifies the overall ability of the SLRs to discriminate between TNBCs and controls. A perfect test with zero false positives and zero false negatives has an area of 1.00, while a truly useless test has an area of 0.5. GraphPad also reports the standard error of the AUC with 95% confidence interval. ROC curve evaluation is completed by P value determination. If the p value is small, that means the SLR of a gene of interest discriminate between TNBCs and controls.

2.4.3 Patient Characteristics Correlation Analyses

To investigate possible correlations between 3’-UTR shortening events and patient characteristics, individual increased SLR values were compared to tumor grade (G1-G2 vs G3), tumor size (T1 vs T2) or lymph node status (LN 0 vs 1). For correlation

analysis of SLRs and patient characteristics, unpaired t-test is used to examine the differences between the means of two groups. To evaluate if these two groups have differences in variance f-test is used. For both tests, it is assumed that the samples are from a normal distribution.

2.4.4 Relapse Free Survivals

Kaplan-Meier plots were employed to visualize the association between the significant SLR values and survival. The Kaplan-Meier survival curve is defined as the probability of surviving in a given length of time while considering time in many small intervals (Goel, Khanna et al. 2010).

To explore the role of high vs short SLRs within the patient samples, we grouped patients with top 25% (High) and bottom 25% (Low) SLR values for the significant genes and compared the two group's relapse free survival times up to 10 years. Statistical significance and p values were determined by log-rank test using GraphPad.

2.5 microRNA predictions

Shortening of the 3'-UTR confers a selective advantage by allowing escape from microRNA-mediated repression. To test this, short and long isoforms of significant genes were analyzed by miRNA-mRNA interaction prediction tools. 3 different tools were used; TargetScan (http://www.targetscan.org/), PITA (http://genie.weizmann.ac.il/), and FindTar3 (http://bio.sz.tsinghua.edu.cn/) to increase the accuracy of the predictions.

TargetScan predicts targets of miRNAs by investigating the presence of conserved 8-mer and 7-mer sites that match the seed region of each miRNA (Lewis, Burge et al. 2005). In mammals, predictions are ranked based on the predicted efficacy of targeting as calculated using the context plus scores of the sites (Grimson, Farh et al.

2007; Garcia, Baek et al. 2011). PITA is another algorithm for miRNA target identification based on a thermodynamic model (Kertesz, Iovino et al. 2007). PITA scores miRNA-target interactions by an energy score, ΔΔG, equal to the difference between the energy gained by binding of the microRNA to the target, ΔGduplex, and the energy required to make the target region accessible for microRNA binding, Δ Gopen. Lastly, FindTar3 predicts miRNA-mRNA interaction based on seed-pairing, free energy of miRNA:mRNA duplex, and proper dynamic score. Moreover, FindTar3 concerns about central loop score which is the partial complementarity of miRNA: mRNA duplex to avoid miRNA target induced instability of miRNAs (Ye, Lv et al. 2008).

2.5.1 Circos plots

The intersection of all 3 prediction sets and their relevant binding sites were visualized by a Circos plot (Krzywinski, Schein et al. 2009). Circos is an ideal diagram to visualize miRNA:3'-UTR interactions. After listing the predicted miRNAs and their target sequences on 3'-UTR with coordinates, Dr. Can plotted the Circos for selected genes.

In the Circos plot, short 3'-UTR designated as S (red) and long 3'-UTRs designated as L (blue) and corresponding predicted miRNAs are visualized. Total 3'-UTR sizes and miRNA binding prediction positions are drawn to scale. Proximal 3'-UTR targeting miRNA links are shown in red and distal 3'-UTR targeting miRNA links are shown in blue.

2.6 Ontology Analysis

The Database for Annotation, Visualization and Integrated Discovery (DAVID) was used for ontology analysis (Huang da, Sherman et al. 2009; Huang da, Sherman et al. 2009). DAVID was designed to classify genes (and their proteins) in order to facilitate high-throughput analysis. DAVID bioinformatics resources contain an

integrated biological knowledge base and analytic tools aimed at systematically mining biological meaning from large gene/protein lists. Genes can be classified according to molecular function, biological process, family/subfamily, or pathway. Here, genes showing significant 3'-UTR changes were analyzed for their molecular functions.

2.7 Expression Analysis

MIQE Guidelines (Bustin, Benes et al. 2009) and checklist were followed throughout the RNA isolation, quantification, cDNA synthesis and RT-qPCR analysis.

2.7.1 Cell lines

BT-20, HCC1143, HCC1937, MDA-MB-157, MDA-MB-231, and MDA-MB-468 cells were grown in DMEM with Earle's salts, 10% FBS, 1% pen/strep and 0.1% non-essential amino acids. MCF10A cells were grown in DMEM/F12 medium with 5% horse serum, 100 mg/mL Epidermal Growth Factor (EGF), 1 mg/mL cholera toxin, 1 mg/mL hydrocortisone, and 10 mg/mL insulin. Cell lines were grown as monolayers and were incubated at 37^{o}C with 95% humidified air and 5% CO_2.

2.7.2 RNA Isolation

Total RNA was isolated using High Pure RNA Isolation Kit (Roche Applied Science). RNA isolations were performed according to the manufacturer's instructions.

2.7.3 DNA contamination, Quantification and Integrity Assessment

For DNase treatment, total RNA was incubated with DNase I (Roche Applied Science) at 37^{o}C for 1 hour followed by phenol-chloroform extraction (reaction conditions given in Table 2.1).

After DNase treatment, lack of DNA contamination was confirmed by PCR using *GAPDH* (Glyceraldehyde 3-phosphate dehydrogenase) specific primers. Following conditions were used for the PCR reactions: incubation at 94^{0}C for 10 minutes, 40 cycles of 94^{0}C for 30 seconds, 56^{0}C for 30 seconds, and 72^{0}C for 30 seconds, and final extension at 72^{0}C for 5 minutes. MDA-MB-231 cDNA was used as a positive control.

After confirming absence of DNA, quantification of RNA samples was done using NanoDrop ND1000 (Thermo Scientific) spectrophotometrically (Appendix E). Purity of the RNA samples was determined by A260/A280 and A260/A230 ratios. Higher or lower A260/A280 ratios indicates whether the sample is contaminated by proteins or a reagent such as phenol, likewise A260/A230 ratio could be an indicator of a problem with the extraction procedure. An examplificatory result was shown in Appendix E showing a proper RNA sample, with A260/A280 ratio 1,95 and A260/A230 ratio 2,15. All RNA samples used in this study were fitting the criteria in terms of purity.

Table 2.1. Reaction Conditions of DNase Treatment

Components	**Amount**
Total RNA	10 µg
10X Incubation Buffer (400 mM Tris-HCl, 100 mM NaCl, 60 mM $MgCl_2$, 10 mM $CaCl_2$, pH 7.9)	10 µL
DNase I recombinant, RNase-free (10 units/µl) Roche Applied Sciences Cat. No. 04 716 728 001	1 µL
RiboLock™ RNase Inhibitor (20u/µL) Fermentas #EO0381 *One unit of RiboLock™ RNase Inhibitor inhibits the activity of 5 ng RNase A by 50%	1 µL
RNase-free Water	To 100 µL
DNase treatment was done at 37º C for 60 min. Reaction was terminated by phenol/chloroform extraction.	

2.7.4 cDNA synthesis

cDNAs (20 µl) were synthesized using the RevertAid First Strand cDNA Synthesis Kit (Fermentas) using 2 µg total RNA and oligo dT primers (reaction conditions given in Table 2.2).

Table 2.2. Reaction Conditions of Reverse Transcription

Components	**Amount**
Total RNA	2 µg
Oligo$(dT)_{18}$ primer (100 µM 0.5 µg/µl (15 A260 u/ml))	1 µL
5X Reaction Buffer (250 mM Tris-HCl (pH 8.3), 250 mM KCl, 20 mM $MgCl_2$, 50 mM DTT)	4 µL
10 mM dNTP mix	2 µL
RiboLock™ RNase Inhibitor (20u/µL) *One unit of RiboLock™ RNase Inhibitor inhibits the activity of 5 ng RNase A by 50%	1 µL
RevertAid™ M-MuLV Reverse Transcriptase (200u/µL) *One unit of RevertAid™ M-MuLV RT incorporates 1 nmol of dTMP into a polynucleotide fraction (adsorbed on DE-81) in 10 min at 37°C	1 µL
Nuclease-free water	to 20 µL
cDNA was synthesized at 42º C for 60 minutes. Reaction was terminated by keeping at 70º C for 5 minutes. cDNA samples were stored at -20º C.	

2.7.5 RT-qPCR Analysis

For RT-qPCR, SYBR® Green Mastermix (Roche Applied Science) was used using the Rotor Gene 6000 (Corbett, Qiagen) cycler. 20 µL reactions were performed with 300 nM of specific primer pairs. Primers used in this study are listed in Appendix F. The fold change for the isoforms was normalized against the reference gene; *SDHA* (NM_004168) (Gur-Dedeoglu, Konu et al. 2009). Reference gene, *SDHA*, expression is consistent across breast cancer cell lines (Gur-Dedeoglu, Konu et al. 2009).

For the relative quantification, the reaction efficiency incorporated ΔΔCq formula was used (Fleige, Walf et al. 2006). Three independent biological replicates with 3 technical replicas per experiment were used for each PCR. One-way ANOVA with

Tukey’s multiple comparison post-test was performed using GraphPad Prism (California, USA).

CHAPTER 3

RESULTS AND DISCUSSION

3.1 APA Events in TNBCs

Growing evidence suggests 3'-UTR length changes in different physiological conditions such as differentiation, development, transformation, and proliferation (Ghosh, Soni et al. 2008; Ji, Lee et al. 2009; Ji and Tian 2009; Mayr and Bartel 2009). TNBCs are highly proliferative cells with metastatic properties. As the relationship between 3'-UTR shortening and proliferation has been described in several studies (Sandberg, Neilson et al. 2008; Akman, Can et al. 2012), we investigated APA patterns in TNBCs to better understand the molecular etiology of this highly proliferative and aggressive subtype of breast cancer.

In this study, we compared 520 TNBC microarray datasets with 82 histologically normal epithelium from breast cancer patients and cancer-free prophylactic mastectomy microarray datasets by APADetect, a probe-based microarray analysis tool. APADetect analyzed 2066 genes on HGU133A and 3067 genes on HGU133Plus2.0 platforms. 3'-UTR shortening or lengthening events were represented as SLR (Short/Long Ratio) values; which are proximal/distal probe intensity means. APADetect listed a total of 2126 APA cases with either shortening, lengthening or constant SLRs. 50 from each are listed in Appendix B.

APADetect results of SLRs in normal breast tissue and TNBC patients were then analyzed by SAM (Tusher, Tibshirani et al. 2001). According to APADetect analysis followed by SAM 165 genes' SLRs were found significantly different in TNBCs

compared to controls. SAM output of these events is shown in Figure 3.1. Out of 165 significant APA events, 68.5% (113 of 165) of them were 3'-UTR shortening and 31.5% (52 of 165) were 3'-UTR lengthening events. A complete list of 165 genes and their significance rankings are listed in Appendix C.

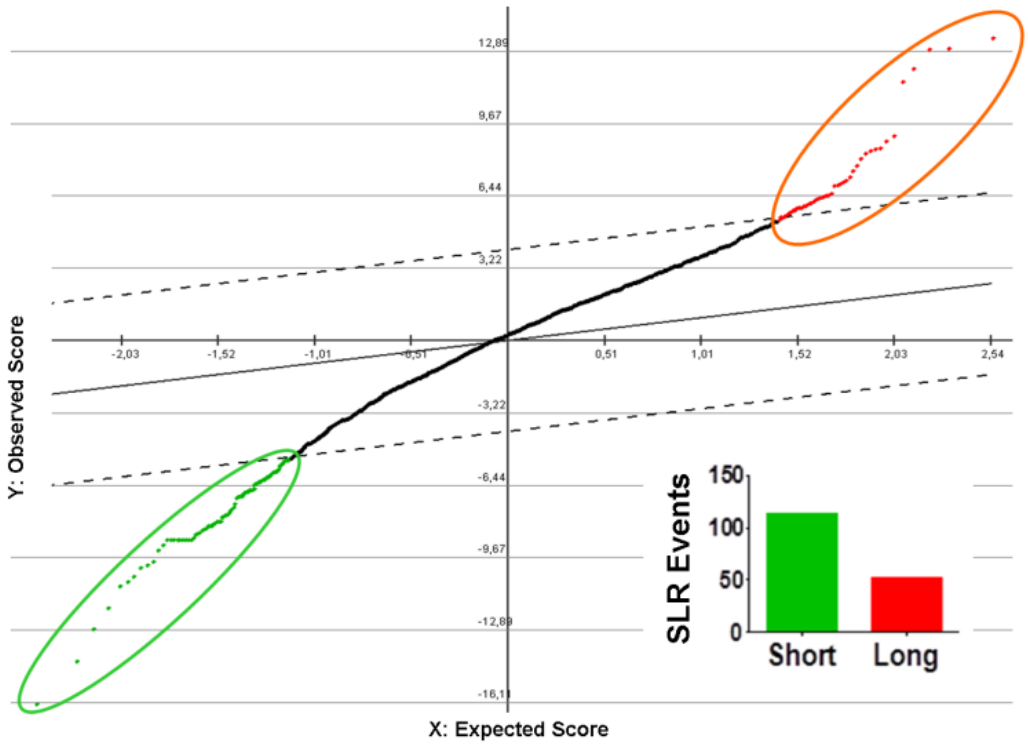

Figure 3.1. SAM analysis of SLR events in TNBCs compared to controls. SAM analysis of SLR events in 520 TNBCs compared to 82 normal breast tissues. A total of 2182 genes that passed stringency filters of APADetect were analyzed by SAM (Tusher, Tibshirani et al. 2001) to evaluate statistical significance of SLR events. Out of 165 significant APA events, 68.5% (113 of 165) of them were 3'-UTR shortening (green bar), and 31.5% (52 of 165) were 3'-UTR lengthening (red bar) events.

Then, an ontology analysis on both shortening and lengthening APA events was performed by DAVID (Huang da, Sherman et al. 2009; Huang da, Sherman et al. 2009). When genes are classified in terms of biological processes they belong, metabolic processes such as protein metabolism and macromolecule biosynthesis were found to be enriched as shown in Figure 3.2. GOTERM IDs of annotations along with the p values are listed in Appendix D. Enrichment in protein metabolism and macromolecule biosynthesis are consistent with proliferative nature of TNBCs.

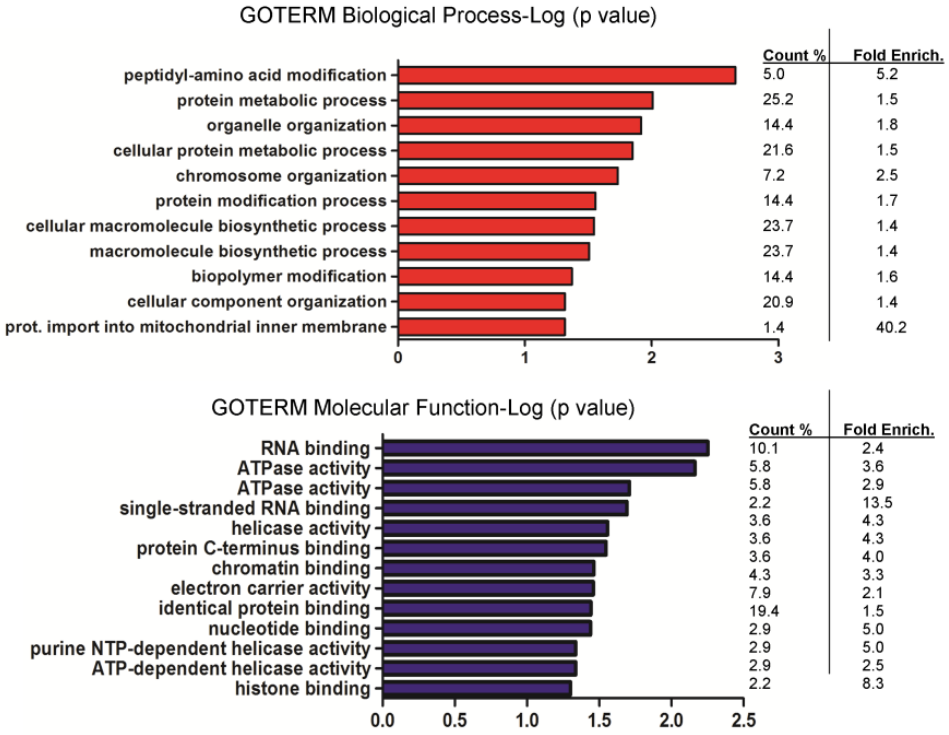

Figure 3.2. Ontology analysis of significant genes. DAVID ontology analysis of significantly changed SLRs of 165 genes based on biological processes and molecular functions. Count % refers to the percent of genes in the dataset associated with each GO term. Fold enrichment value shows how that particular activity is abundant in the data set compared to genome.

Ontology analysis also revealed that the set of genes showing APA were highly enriched for RNA binding related molecular functional categories. These observations can be attributed to the fact that RNA-binding proteins constitute a significant fraction of cellular proteins and have important roles in cells. Their functions include control of transcription and translation, DNA repair, splicing, apoptosis and proliferation.

3.1.1 3'-UTR Shortening Events in TNBCs

Previous studies showed the tendency of proliferative and cancer cells to produce mRNAs with shorter 3'-UTRs, evading miRNA and RBPs regulation (Sandberg, Neilson et al. 2008; Mayr and Bartel 2009; Akman, Can et al. 2012). Since SAM results demonstrated 3'-UTR shortening was prominent to lengthening events, genes with most significant SLR values were further analyzed.

Six high SLR genes; IQCK (IQ Domain-Containing Protein K), USP9X (Ubiquitin Specific Peptidase 9 X-Linked), RPL13 (Ribosomal Protein L13), SNX3 (Sorting Nexin-3), TOP2A (Topoisomerase DNA II alpha), and YME1L1 (YME1-like 1 ATPase) had significant differences in the means and variances of SLRs in 520 TNBC patient samples compared to controls. Scatter plots of these genes are depicted in Figure 3.3. Although the input number of control or patient samples for individual genes was as given in the relevant data set, due to outliers detected by stringency filters, numbers of controls and patients varied for each gene. Sample numbers that passed the filters are indicated in the figures and legends.

Known functions of these genes and their relevance to cancer pathways are listed in Table 3.1.

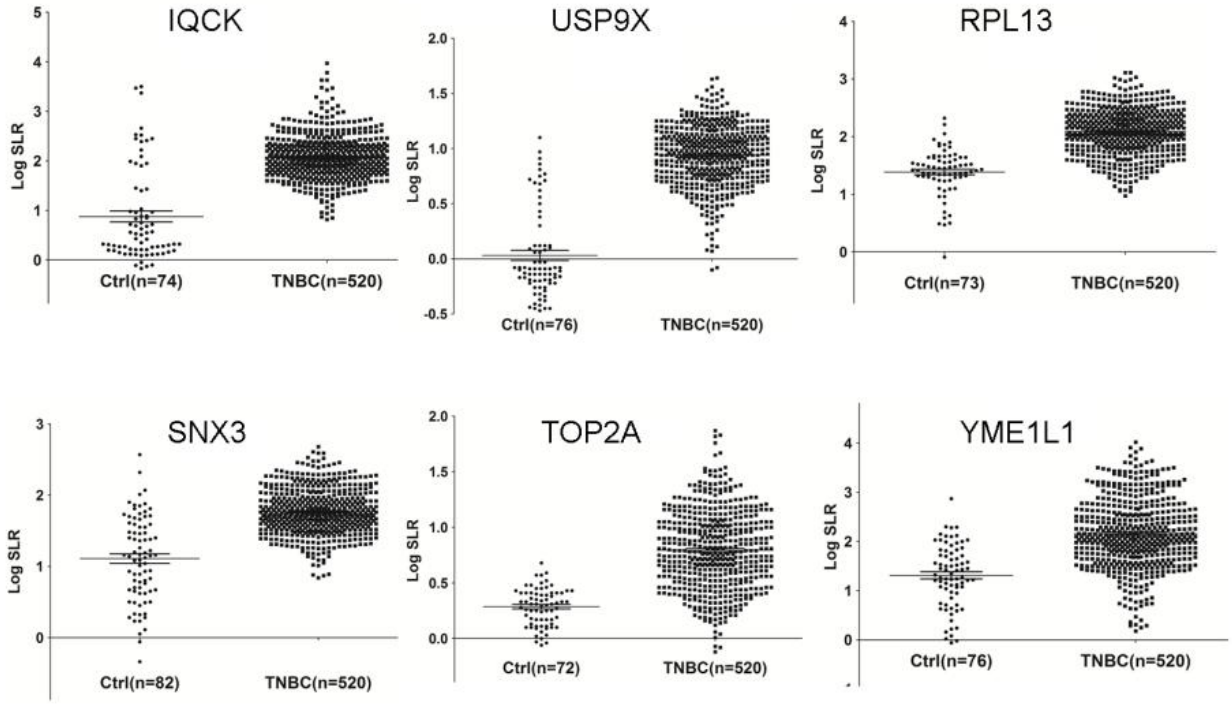

Figure 3.3. Scatter plots of transcripts undergoing 3'-UTR shortening in TNBCs. Individual SLR values of *IQCK, USP9X, RPL13, SNX3, TOP2A* and *YME1L1* as the most significantly shortened 3'-UTRs in TNBCs compared to normal breast tissue (Control group, Ctrl). Scatter plots were drawn using GraphPad software. Significance of group means for TNBCs and controls were calculated by t-test (Unpaired t test with Welch`s Correction). All p values were < 0.0001. The control numbers vary due to outliers excluded by APADetect stringency filters.

Table 3.1. The 3'UTR shortening genes in TNBCs.

Gene	Function/Relevance to Cancer	References
IQCK (IQ domain containing protein K)	- IQ domain known to bind calmodulin in a Ca2+-independent manner. - Its function remains unknown.	
USP9X (Ubiquitin specific peptidase 9,X-linked)	-Depletion of USP9X led to down-regulation of SMURF1 and significantly impaired cellular migration. -USP9X downregulation reduces ER+ breast cancer cells resistant to tamoxifen.	(Xie, Avello et al. 2013; Oosterkamp, Hijmans et al. 2014)
RPL13 (Ribosomal protein L13)	-It is an essential component of 60S ribosomal subunit. -Upregulation of RPL13 expression levels was detected in colorectal, gastric and liver cancers.	(Kobayashi, Sasaki et al. 2006)
SNX3 (Sortin nexin 3)	-It is known to control Wingless/Wnt secretion through regulating retromer-dependent recycling.	(Zhang, Wu et al. 2011)
TOP2A (Topoisomerase DNA II alpha)	-It codes for a topoisomerase and is targeted by anthracyclines used during treatment of TNBC patients as well as other cancers.	(Romero, Caldés et al. 2012; Joerger and Thürlimann 2013)
YME1L1 (YME1-like 1 ATPase)	-It is required for apoptotic resistance, cristae morphogenesis, and cell proliferation. -It is implicated in cancer development and progression and has also been identified as a MYC-responsive gene.	(Stiburek, Cesnekova et al. 2012) (Guo, Malek et al. 2000; Wan, Gong et al. 2004)

ROC (Receiver Operating Characteristics) curve analysis was performed for 6 genes with significantly high SLRs to evaluate sensitivity and specificity of each increased SLR event compared to controls (Metz 1978; Wray, Yang et al. 2010). As can be seen from Figure 3.4, AUC (area under the curve) values for all six genes were higher than 0.8 ($p<0.0001$) where maximum value for AUC is 1. Thus, ROC curves further confirmed the specificity and sensitivity of the individual SLR values of these genes in their abilities to discriminate TNBC from controls.

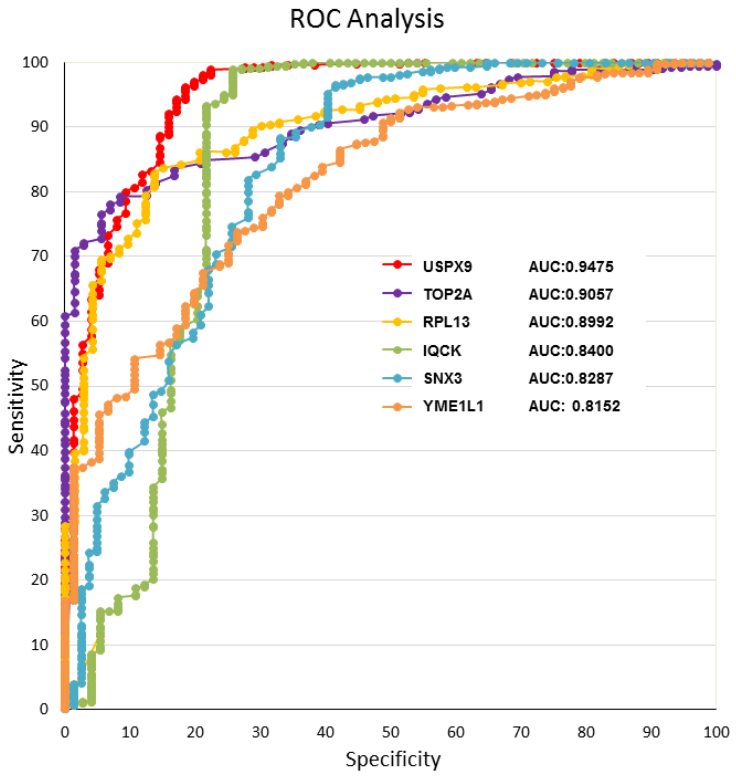

Figure 3.4. ROC curve analysis of transcripts undergoing 3'-UTR shortening. ROC (Receiver Operator Characteristic) curves for the most significantly shortened 3'UTRs in 520 TNBCs compared to controls. Area under curve values for each gene's individual SLR is given in the key.

3.1.1.1 Patient characteristics correlation analysis

Given the significant 3'-UTR shortening events implicated by high SLRs compared to controls, further investigation was performed to assess possible correlations between patient characteristics and increased SLR genes. For this purpose, the means of tumor grade (G1-G2 vs G3), tumor size (T1 vs T2) or lymph node status (LN 0 vs 1) of patients were compared by unpaired t test with Welch's correction and F-Test. Group means' and variances of patient characteristics are listed in Table 3.2.

Table 3.2. Correlation between patient characteristics and 3'-UTR shortening. 3'-UTR shortening events in 520 TNBC patients and their correlation with (Grade1-2 vs Grade 3), Tumor Size (T1 vs T2) and Lymph Node metastasis (0 vs 1) status. Unpaired t test with Welch's correction was used for statistical significance of group means and variances.

	Grade				Tumor				Lymph Node			
Genes	**Mean G1 - G2**	**Mean G3**	**P (t-test)**	**P (F- test)**	**Mean T1**	**Mean T2**	**P (t-test)**	**P (F- test)**	**Mean LN 0**	**Mean LN 1**	**P (t-test)**	**P (F- test)**
YME1L1	2.10 ± 0.06	2.20 ± 0.04	0,1605	0.9507	2.02 ± 0.06	2.14 ± 0.04	0,1156	0.2185	2.01 ± 0.03	1.85 ± 0.07	0,0412*	< 0.0001
IQCK	2.08 ± 0.04	2.06 ± 0.02	0,7119	0.7798	2.04 ± 0.04	2.01 ± 0.02	0,5177	0.1716	2.04 ± 0.02	1.93 ± 0.04	0,0152*	**0.0037****
USP9X	0.87 ± 0.03	0.94 ± 0.01	0,0463*	**0.0016****	0.87 ± 0.03	0.95 ± 0.01	0,0439*	0.2378	0.92 ± 0.02	0.93 ± 0.03	0,5626	**0.009****
RPL13	2.09 ± 0.04	2.09 ± 0.02	0,8875	0.2694	2.05 ± 0.04	2.12 ± 0.02	0,1449	0.3381	2.17 ± 0.02	2.18 ± 0.04	0,8325	0.0186*
SNX3	1.67 ± 0.03	1.77 ± 0.17	**0,0054****	0.1249	1.71 ± 0.03	1.75 ± 0.02	0,2825	0.5457	1.70 ± 0.02	1.65 ± 0.03	0,12	0,1007
TOP2A	0.71 ± 0.03	0.84 ± 0.02	**0.0002*****	0.2612	0.75 ± 0.03	0.78 ± 0.02	0.5503	0.3867	0.76 ± 0.02	0.75 ± 0.03	0.9074	0.9584

The means of Grade 1-2 vs Grade 3 tumor SLRs for *TOP2A* and *SNX3* were significantly different, however; variance within groups was also high. No other very significant correlation was observed. Considering the striking SLR differences in patients and controls, observing a shared SLR profile for specific genes within the patient group in terms of tumor grade, size and lymph node status suggested such APA changes to possibly occur early in tumorigenesis.

Next, to further dissect the role of high vs short SLRs within the patient samples, we grouped patients with top 25% (High) and bottom 25% (Low) SLR values for these 6 genes and plot Kaplan-Meier curves to compare the two group's relapse free survival times up to 10 years. Interestingly, as depicted in Figure 3.5 A high and low *YME1L1*, *SNX3* and *USP9X* SLR values correlated with the poor and good prognosis of patients with hazard ratios of 4.75 ($p<0.001$), 3.362 ($p<0.0001$) and 1.99 ($p<0.005$) (95%CI), respectively. No significant relapse free survival rate difference was detected for the highest 25% or lowest 25% SLR patient groups for the *IQCK*, *RPL13* or *TOP2A* genes (Figure 3.5 B).

A.

B.

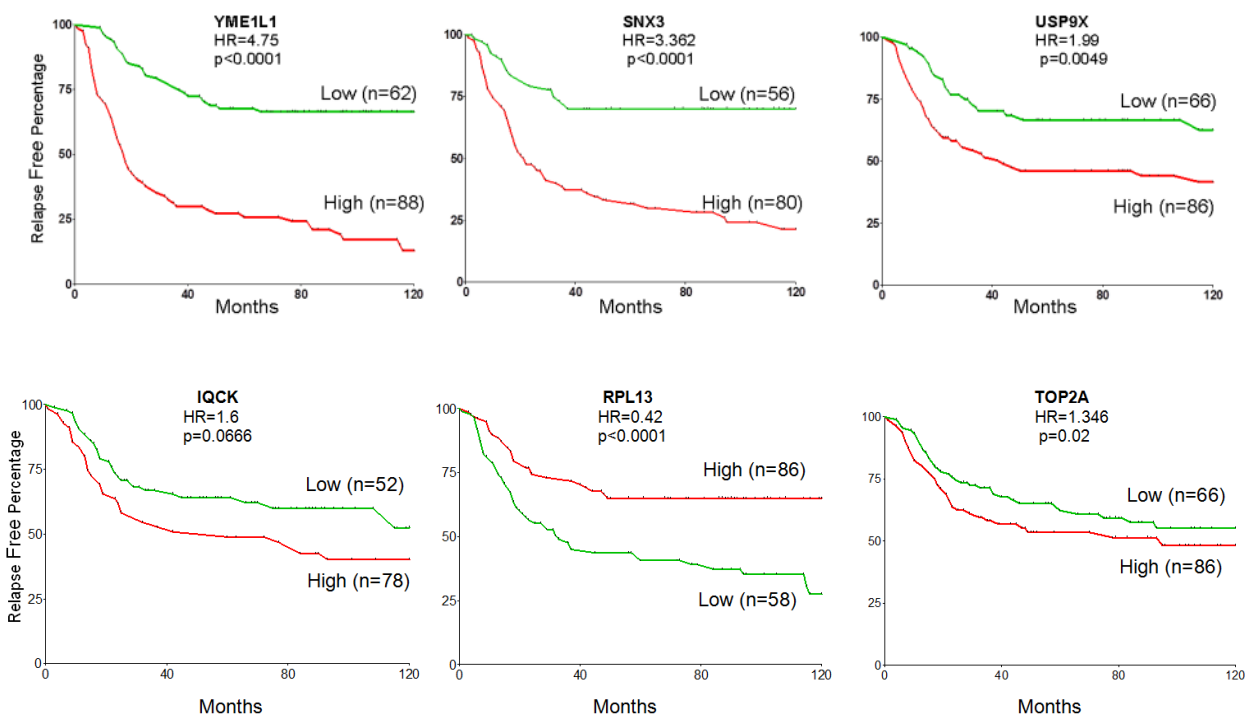

Figure 3.5. Kaplan-Meier curves. Relapse free survival rates of TNBC patients with high (highest 25%) (red line) or low (lowest 25%) (green line) A. SLRs for YME1L1, SNX3 and USP9X for a period of 120 months (10 years). Hazard ratios (95% CI) were 4.75 ($p<0.001$), 3.362 ($p<0.0001$) and 1.99 ($p<0.005$), respectively B. SLRs for IQCK, RPL13 or TOP2A for a period of 120 months (10 years). Hazard ratios (95% CI) were 4.75 ($p<0.001$), 3.362 ($p<0.0001$) and 1.99 ($p<0.005$) respectively.

TNBCs are known to be an aggressive group of breast cancers with higher rates of relapse compared to ER/PR positive and HER2 positive breast cancers. Significant correlation between YME1L1, SNX3 and USP9X SLR values and relapse might be noteworthy to further analyze functions and post-transcriptional regulations of these genes to delineate molecular etiology of TNBCs.

3.1.1.2 Negative regulatory sites on 3'-UTRs

To investigate whether 3'-UTR shortening was correlated with possible miRNA binding site losses, 3 different miRNA binding prediction tools (PITA, FindTar3 and TargetScan) were used to search for common miRNA predictions for the short vs long 3' UTRs of *USP9X, RPL3, SNX3* and *YME1L1*.

As can be seen from Figure 3.6 few miRNA predictions for all short isoforms whereas a noticeable number of miRNA binding predictions were made for the longer isoforms. *IQCK* and *TOP2A* transcripts' APA is coupled to alternative splicing; proximal poly(A) sites reside within their terminal exons. Hence, *IQCK* and *TOP2A* were not included in the miRNA binding prediction analysis, since experimentally confirmed 3'-UTR boundaries for these potential alternatively spliced isoforms were not known.

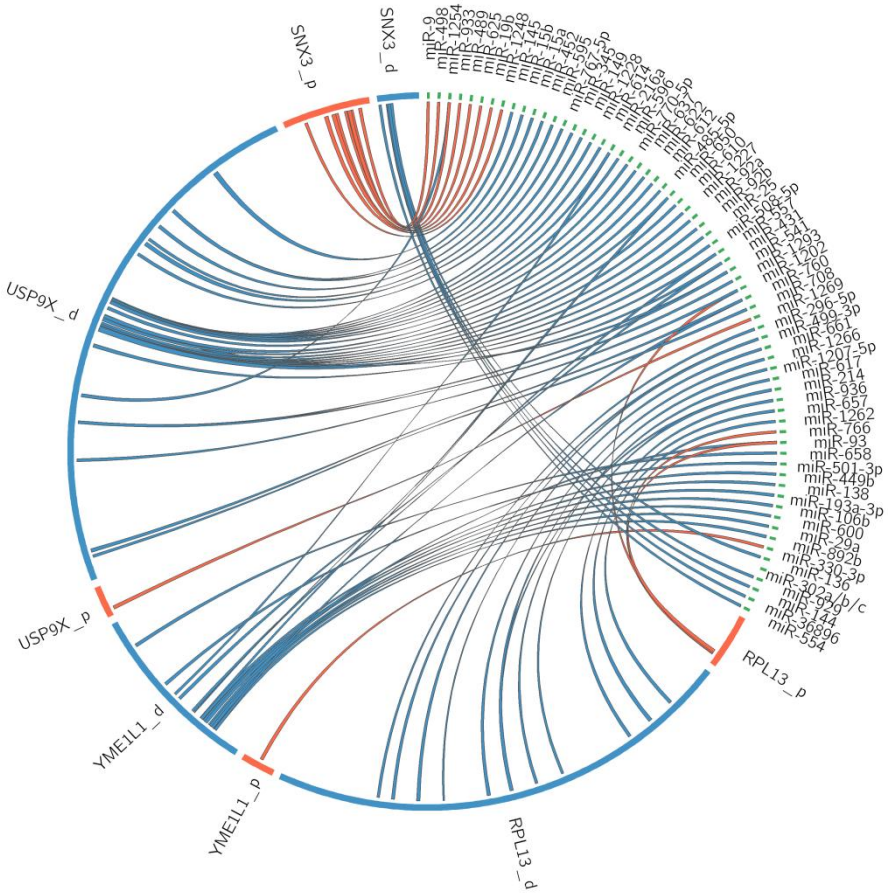

Figure 3.6. Circos plot. Short (shown as "S" in red) and long (shown as "L" in blue) 3'-UTRs of USP9X, RPL13, SNX3 and YME1L1 and corresponding miRNAs predicted by bioinformatics tools are illustrated by Circos. Total 3'-UTR sizes and miRNA binding prediction positions are drawn to scale. Proximal 3'-UTR targeting miRNA links are shown in red and distal 3' UTR targeting miRNA links are shown in blue.

miRNA binding predictions for these genes pointed out that long 3'-UTR isoforms could be more susceptible to post-transcriptional miRNA dependent repression than the shorter counterparts. These predictions are in accordance with the earlier studies showing that oncogenic transformation tends to favor 3'-UTR shortening and these

shorter isoforms are more stable transcripts than longer isoforms (Mayr and Bartel 2009).

As depicted in Circos, miR-216a, miR-650, miR-557, miR-431 and miR-760 are predicted to be targeting distal 3'-UTRs of both USP9X and YME1L1. Although there is not any study related to breast cancer, miR-216a has been implicated as a metastatic inhibitor in a mouse model (Faraji, Hu et al. 2014). This is consistent with the fact that USP9X and YME1L1 may be evading miR-216a by shortening their 3'-UTRs in metastatic TNBCs.

Likewise, expression of miR-431 has been associated with cell proliferation inhibition (Tanaka, Sugaya et al. 2012). In this study with RSa human cell line, miR-431 shown to target IGFR1 and IRS2, consequently suppressing MAPK pathway.

miR-760 has been found to promote cellular senescence through p53-p21 pathway by down-regulating CKII both in colorectal cancer (HCT116 cells) and lung fibroblast cells (Kim, Lee et al. 2012; Lee, Kim et al. 2014). Interestingly, miR-760 was identified as one of the 10 miRNAs related to chemoresistance in breast cancer and suggested as a prognostic biomarker (Lv, Xia et al. 2014). Together with let-7e, miR-125b, and miR-145, its expression is up-regulated in chemoresistant breast cancer tissues.

It seems like above mentioned miRNAs are targeting multiple genes to regulate metastasis, senescence, and proliferation. During tumorigenesis, to avoid epigenetic regulation of miRNAs, (1) cells can suppress expression of tumor suppressor miRNAs and/or (2) genes may evade miRNA-mediated regulation by shortening their 3'-UTRs. Based on our prediction analysis results, it seems like USP9X, YME1L1, and RPL3 favor 3'-UTR shortening to avoid miRNA regulation.

3.1.1.3 Expression Analysis

Short and long isoform specific primers were designed to verify the in silico results for YME1L1, and USP9X due to their potential role in TNBC development based on survival analysis. RT-qPCR primers were designed to amplify short and long

isoforms, same forward primer and two different reverse primers were chosen. Positions of primers together with APA sites are depicted in Figure 3.7.

Since short isoform primers also anneal to the long isoform differentiation of expression of short isoform from long isoform is challenging. Therefore, relative expression of short product (short+long transcript) to long product (long isoform) was detected by RT-qPCR. During RT-qPCR experiments, reaction efficiencies of each experiment were evaluated and incorporated into the ΔΔCq formula to diminish inter-experimental variations (Fleige, Walf et al. 2006).

Normal breast tissue, MCF10A (non-tumorigenic, immortalized breast cell line) (Soule, Maloney et al. 1990) and 7 TNBC breast cancer cell lines were used for expression analysis. As illustrated in Figure 3.8 and Figure 3.9 SLRs were significantly different in breast cancer cell lines compared to normal breast tissue for both YME1L1, and USP9X. While MCF10A is an immortalized, non-tumorigenic cell line, it is known for their high proliferative rates (doubling time less than 24 hours). It is also identified to be ER-, PR-, HER2- and based on genome profiling analysis it is considered non-transformed counterparts of the TNBC cancer cell lines (Charafe-Jauffret, Ginestier et al. 2006; Chavez, Garimella et al. 2010). These might be the causes of high SLR values of MCF10A compared to normal breast sample.

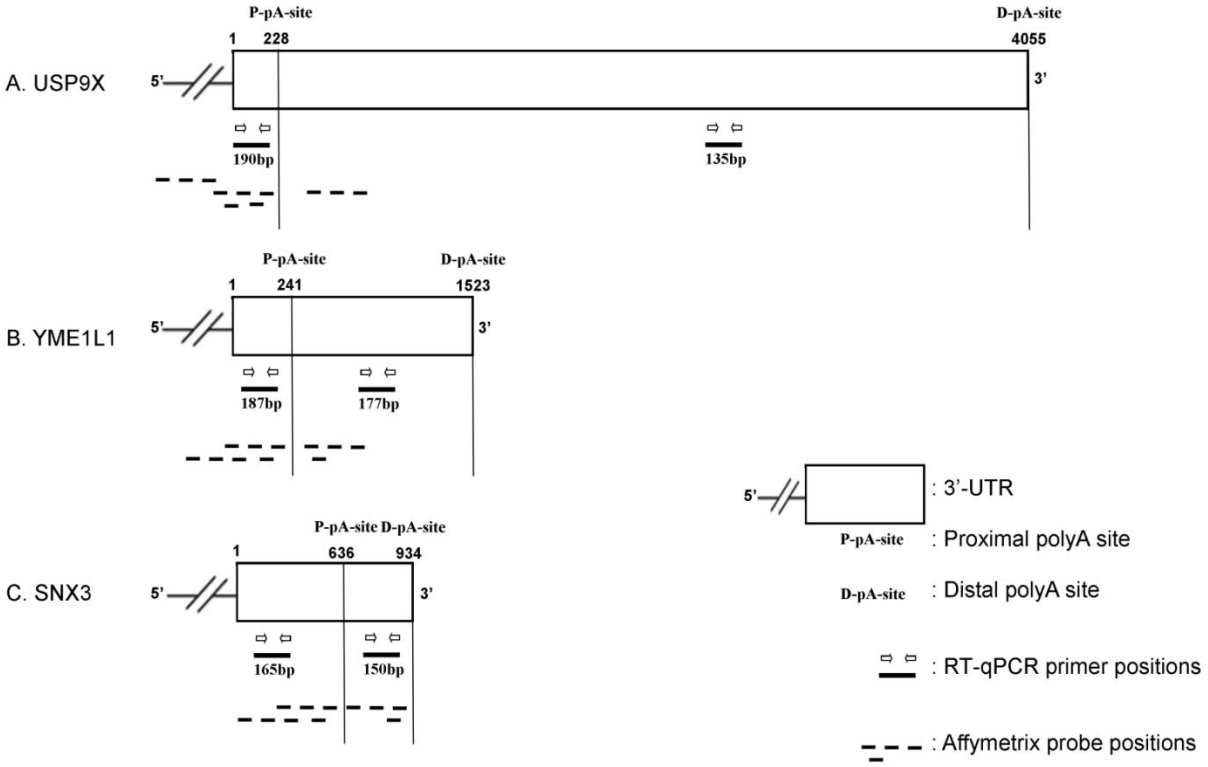

Figure 3.7. 3'-UTRs of USP9X, YME1L1, and SNX3. 3'-UTR's with proximal (P-pA-site) and distal (D-pA-site) polyA sites are illustrated. Positions of 11 Affymetrix probes with respect to polyA sites and RT-qPCR primer positions and product sizes are shown.

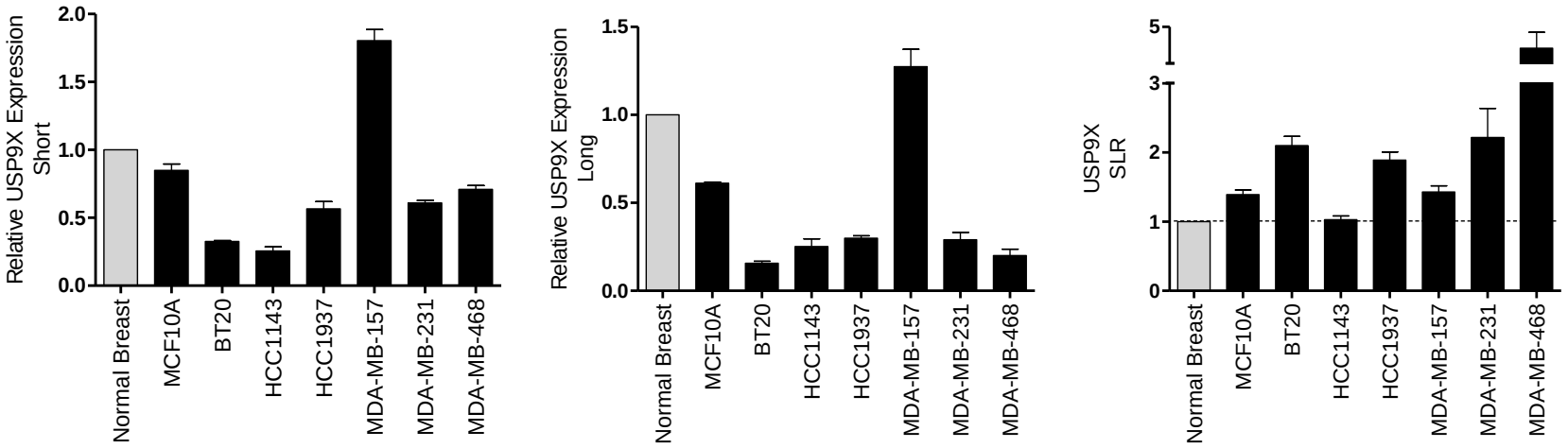

Figure 3.8. Relative quantification of short and long *USP9X* isoforms. Quantification of short and long isoforms of USP9X was performed separately. The fold change for the isoforms was normalized against the reference gene; *SDHA*. Quantification was done using the reaction efficiency correction and ΔΔCq method (Fleige, Walf et al. 2006). The baseline for the short and long isoforms in normal breast sample was set to 1. SLR values were calculated by dividing short values to the longs. Experiments were repeated 3 times with 3 technical replicates.

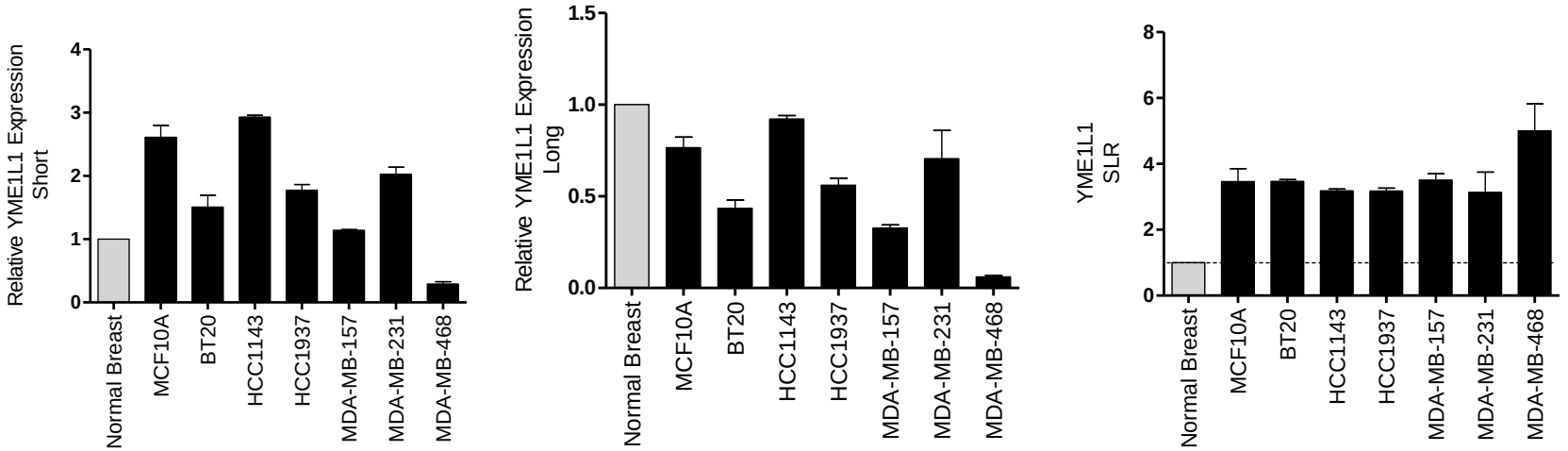

Figure 3.9. Relative quantification of short and long *YME1L1* isoforms. Quantification of short and long isoforms of YME1L1 was performed separately. The fold change for the isoforms was normalized against the reference gene; *SDHA*. Quantification was done using the reaction efficiency correction and ΔΔCq method (Fleige, Walf et al. 2006). The baseline for the short and long isoforms in normal breast sample was set to 1. SLR values were calculated by dividing short values to the longs. Experiments were repeated 3 times with 3 technical replicates.

In conformity with TNBC patient SLRs, TNBC cell line results showed high SLRs compared to normal breast for YME1L1 and USP9X, indicating short isoform of these transcripts are more prominent than longer isoforms in TNBC cell lines.

3.1.2 3'-UTR Lengthening Events in TNBCs

In order to investigate 3'-UTR lengthening in TNBC, 6 representative genes (SLC46A3, MUT, OGT, ERG, TNPO1 and TMBIM4) with low SLRs were selected based on APADetect and SAM scores. Scatter plot analysis of these genes showed significant mean and variance differences between TNBC and controls (Fig 3.10).

Current literature suggests that 3'-UTR lengthening increase the regulation on gene expression through gain of regulatory elements such as RBP recognition sites and miRNA sites. It is also known that APA switches to distal polyA sites during development (Ji, Lee et al. 2009) and differentiation (Ji and Tian 2009) as well as in neuronal cells (Ghosh, Soni et al. 2008).

Although there is a tendency for proximal polyadenylation in cancer cells, APADetect and SAM results showed some significant lengthening cases. Functions of these genes are listed in Table 3.3.

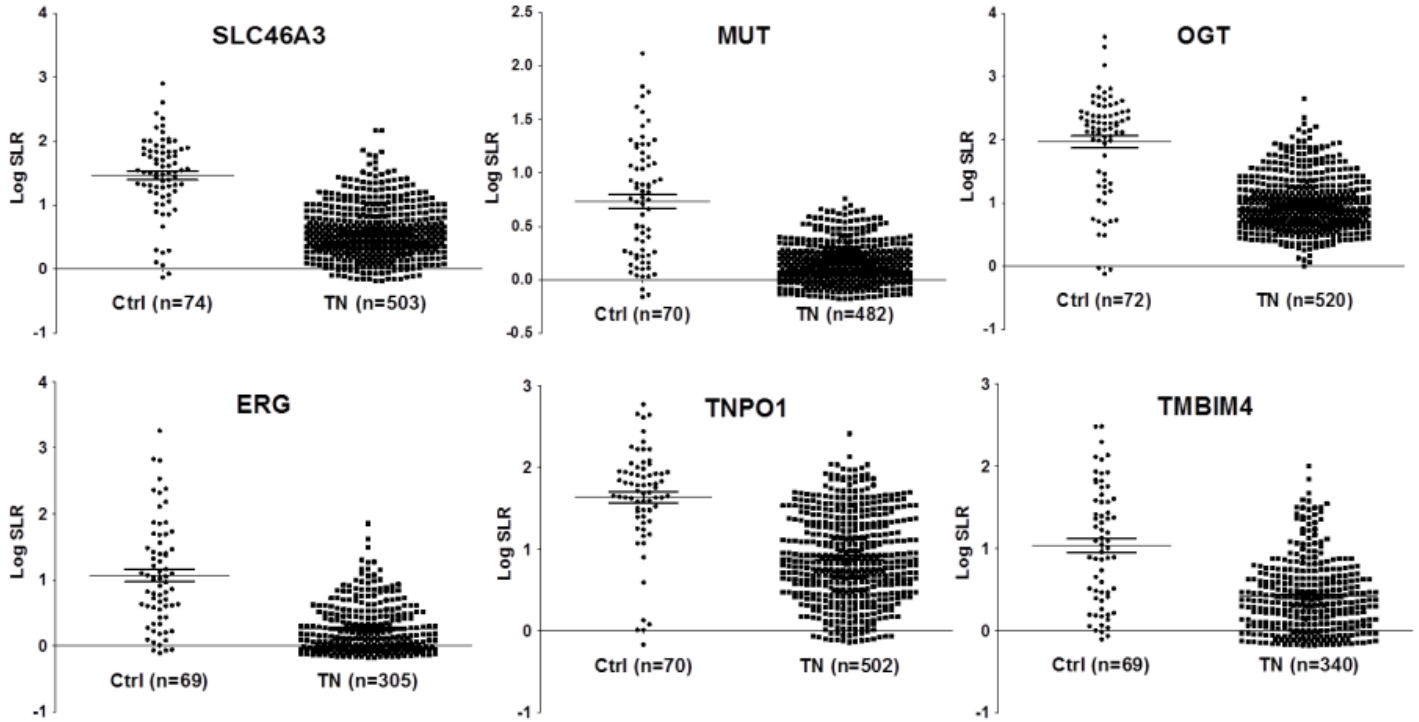

Figure 3.10. Scatter plots of transcripts undergoing 3'-UTR lengthening in TNBCs. Individual SLR values of SLC46A3, MUT, OGT, ERG, TNPO1 and TMBIM4 as the most significantly lengthened 3'-UTRs in TNBCs compared to normal breast tissue (Control group, Ctrl). Scatter plots were drawn using GraphPad software. Significance of group means and variances for TNBCs and controls were calculated by t-test (Unpaired t test with Welch`s Correction). All p values were < 0.0001. The control numbers vary due to outliers excluded by APADetect stringency filters.

Table 3.3. The 3'UTR lengthening genes in TNBCs.

Gene	Function/Relevance to Cancer	References
SLC46A3 (solute carrier family 46, member 3)	- A member of the major facilitator superfamily and SLC46A family. - Its function remains unknown.	
MUT (methylmalonyl CoA mutase) aka MCM	-Catalyzes an essential step in the degradation of several branch-chain amino acids and odd-chain fatty acids.	(Ledley, Lumetta et al. 1988; Kambo, Sharma et al. 2005)
OGT (O-linked N-acetylglucosamine)	-Adds of a sugar moiety to serine/threonine residues of cytosolic or nuclear proteins. -regulates cancer cell metabolic reprograming and survival stress signaling via regulation of HIF-1α	(Caldwell, Jackson et al. 2010; Ferrer, Lynch et al. 2014)
ERG (ETS-related gene) aka p55	-Transcriptional regulator of embryonic development, cell proliferation, differentiation, angiogenesis, inflammation, and apoptosis. -Fusion gene products, such as TMPSSR2-ERG and NDRG1-ERG in prostate cancer, EWS-ERG in Ewing's sarcoma and FUS-ERG in acute myeloid leukemia.	(Mochmann, Neumann et al. 2014)
TNPO1 (Transportin 1) aka MIP1	-Mediates the nuclear import of the RNA-binding protein hnRNPA1, RNA editing protein ADAR1, and PABPN1.	(Fritz, Strehblow et al. 2009; Mallet and Bachand 2013; Twyffels, Gueydan et al. 2014)
TMBIM4 (Transmembrane BAX Inhibitor Motif Containing 4)	-Implicated in death receptor regulation, and modulation of endoplasmic reticulum (ER) calcium homeostasis.	(Reimers, Choi et al. 2008)

Interestingly, TNPO1, Transportin-1 has been identified as the import factor for the PABPN1 from cytoplasm to nucleus (Mallet and Bachand 2013). PABPN1 (poly(A)-binding protein, nuclear 1) is required for progressive and efficient polymerization of poly(A) tails and controls the size of the poly(A) tail. Moreover, it has been shown that PABPN1 suppresses proximal APA by selectively binding to distal poly(A) sites in oculopharyngeal muscular dystrophy (OPMD) (Jenal, Elkon et al. 2012) (de Klerk 2012). In our case, increasing regulation on TNPO1 by increased 3'-UTR length may eventually cause decreased activation of PABPN1.

O-GlcNAcylation is a dynamic post-translational modification involved in a wide range of biological processes and diseases such as cancer (Forma, Jozwiak et al. 2014; Ma and Vosseller 2014). O-GlcNAcylation can increase and decrease the activity of enzymes as well as interfere with protein stability and interaction, which makes OGT an important regulator of metabolism and intracellular signaling of tumor cells. OGT links the cellular metabolic state and the epigenetic status of cancer cells by interacting with and modifying many epigenetic factors, such as HCF-1, TET, mSin3A, HDAC, and BAP1 (Chen, Chen et al. 2013; Cox and Marsh 2013; Lazarus, Jiang et al. 2013). Although it has been implicated as an oncogene in a few studies (Huang, Pan et al. 2013), its role in gene regulation still remains to be elucidated in breast cancer. OGT seems as an important epigenetic regulator itself, affects of 3'-UTR lengthening of OGT on TNBC proteome could be diverse.

Based on up-to-date literature, *SLC46A3* and *MUT* are not implicated in tumorigenesis. *SLC46A* is a member of solute carrier family proteins, which are membranous transporters. MUT is an enzyme residing in mitochondria, it is mainly implicated in metabolic disorders such as methlymalonic acidemia (Zsengeller, Aljinovic et al. 2014).

Patient characteristics were investigated to reveal possible correlations with lengthened 3'-UTRs. However, no significant correlations between SLRs of these genes and tumor grade (G1-G2 vs G3), tumor size (T1 vs T2) or lymph node status (LN 0 vs 1) was found (Table 3.4). Likewise, no correlation between low SLRs of these genes and relapse free survival time was detected when Kaplan-Meier curves plotted.

Table 3.4. Correlation between patient characteristics and 3'-UTR lengthening.

	Grade				Tumor				Lymph Node			
Genes	**Mean G1- G2**	**Mean G3**	**P (t-test)**	**P (F- test)**	**Mean T1**	**Mean T2**	**P (t-test)**	**P (F- test)**	**Mean LN 0**	**Mean LN 1**	**P (t-test)**	**P (F- test)**
SLC46A3	0.66 ± 0.04	0.55 ± 0.02	0.0179*	0,1834	0.59 ± 0.04	0.54 ± 0.02	0,2655	0,6974	0.50± 0.02	0.49 ± 0.04	0,8717	0,5646
MUT	0.19 ± 0.02	0.15 ± 0.01	0,1167	0,2377	0.16 ± 0.02	0.15 ± 0.01	0,4861	0,1656	0.14 ± 0.01	0.18 ± 0.02	0,3262	0,2304
OGT	1.0 ± 0.04	1.0 ± 0.02	0,8566	0,5851	1.04 ± 0.04	0.95 ± 0.01	0,1874	0,3704	1.00 ± 0.03	1.02 ± 0.04	0,7249	0,7984
ERG	0.25 ± 0.04	0.25 ± 0.03	0,9747	0,3213	0.23 ± 0.04	0.26 ± 0.03	0,5663	0,5296	0.15 ± 0.02	0.24 ± 0.05	0,0779	0,3694
TNPO1	0.89 ± 0.05	0.92 ± 0.03	0,6146	0,1706	0.86 ± 0.06	0.82 ± 0.03	0,484	0,4428	0.79 ± 0.03	0.57 ± 0.04	**< 0.0001*****	0,4007
TMBIM4	0.41 ± 0.06	0.41 ± 0.03	0,951	0,0707	0.43 ± 0.06	0.45± 0.06	0,901	0,3657	0.39 ± 0.06	0.35 ± 0.06	0,8894	0,8564

SLR correlation with Grade (Grade1-2 vs Grade3), Tumor Size (T1 vs T2) and Lymph Node metastasis (0 vs 1) in 520 TNBC patients. Unpaired t test with Welch's correction was used for statistical significance of group means and variances.

Next, Kaplan-Meier curves were plotted for lengthened 3'-UTRs. To evaluate the role of high vs short SLRs within the patient samples, we grouped patients with top 25% (High) and bottom 25% (Low) SLR values for these 6 genes and plot Kaplan-Meier curves to compare the two group's relapse free survival times up to 10 years. As depicted in Figure 3.11 high and low TNPO1, and SLC46A3 SLR values correlated with the poor and good prognosis of patients with hazard ratios of 2.6 ($p<0.001$), and 1.82 ($p=0.0064$) (95%CI), respectively. No significant relapse free survival rate difference was detected for the highest 25% or lowest 25% SLR patient groups for the ERG, OGT, MUT, and TMBIM4 genes.

Here, survival analysis of TNPO1 indicated that; when 3'-UTR of TNPO1 lengthened patients have poor prognosis. TNPO1 has been shown to import PABPN1 from cytoplasm to nucleus (Mallet and Bachand 2013). As PABPN1 is known to suppresses proximal APA usage (Jenal, Elkon et al. 2012) (de Klerk 2012), decreased import of PABPN1 to nucleus may lead to increased usage of proximal APA usage. Overall, lengthening of TNPO1 3'-UTR may affect global APA in favor of proximal poly(A) site usage. This outcome may affect the progression of cancer in concordance with previous studies (Mayr and Bartel 2009).

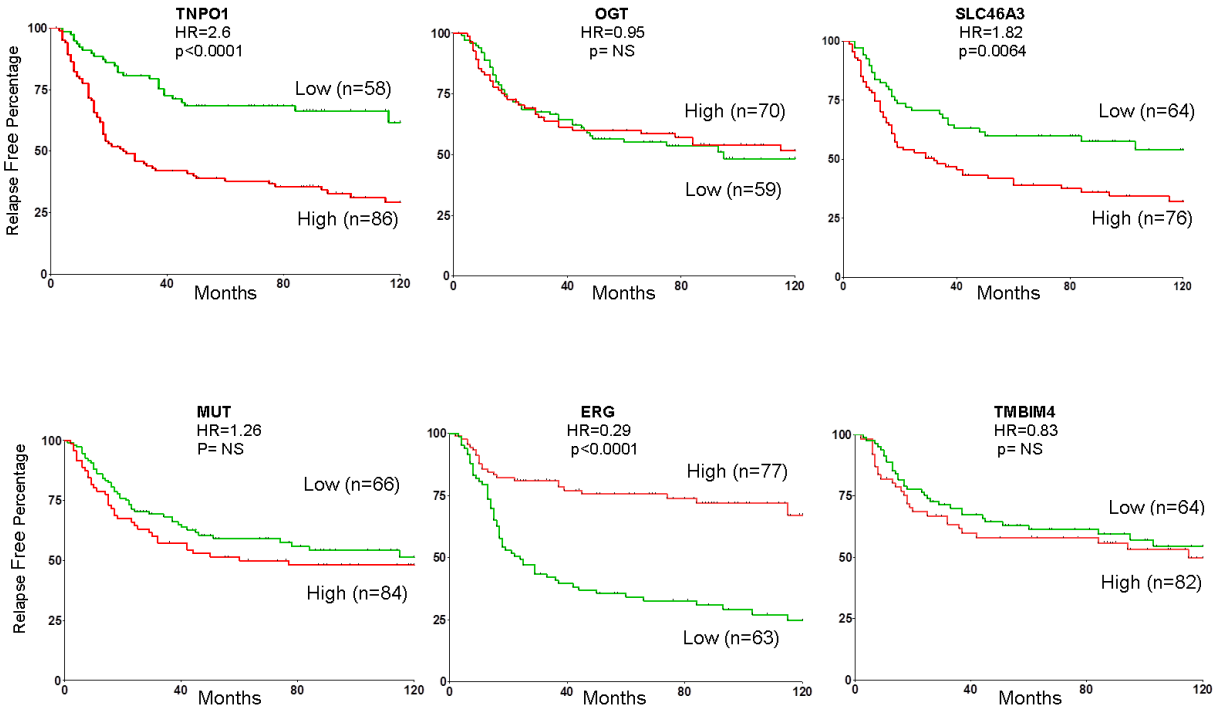

Figure 3.11. Kaplan-Meier curves. Relapse free survival rates of TNBC patients with high (highest 25%) (red line) or low (lowest 25%) (green line). SLRs for TNPO1, OGT, SLC46A3, MUT, ERG, and TMBIM4 for a period of 120 months (10 years). Hazard ratios (95% CI) were 2.6 (p<0.001), 0.95 (p=NS)), 1.82 (p=0.0064), 1.26 (p=NS), 0.29 (p<0.001) and 0.83 (p=NS), respectively.

3.1.3 Constant 3'-UTR lengths

In order to validate our APA based 3'-UTR length change detection approach, six random genes were selected from the group of genes that did not have a significantly different SLR in TNBCs compared to controls. Individual scatter plots of these genes showed no significant change in the means and variances of SLRs in the TNBC and controls (Figure 3.12), suggesting no 3'-UTR length change.

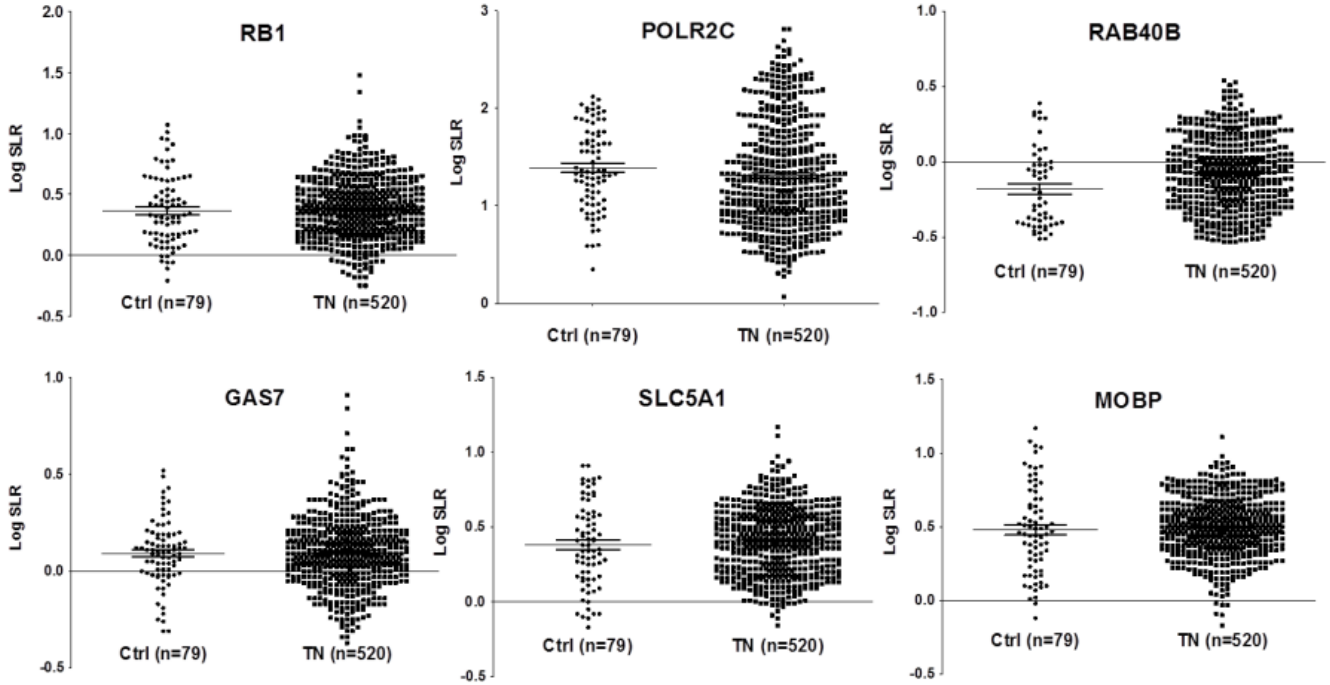

Figure 3.12. Scatter plots of transcripts with constant 3'-UTR lengths. Individual SLR values of *RB1*, *POLR2C*, *RAB40B*, *GAS7*, *SLC5A1*, *MOBP*. 3'UTR lengths of these genes did not change in TNBCs compared to normal breast tissue (Control group, Ctrl). Scatter plots were drawn using GraphPad software. Significance of group means and variances for TNBCs and controls were calculated by t-test (Unpaired t test with Welch`s Correction). No significance was detected.

Overall, out of 2126 genes analyzed, 1003 did not have a detectable APA pattern change. Threshold was set to ±10%; SLR values ranging from 0.9 to 1.1 were considered as constant SLRs.

3.2 APA Patterns Comparison between TNBCs and ER+ Tumors

To investigate whether 3'-UTR shortening events are specific to TNBC, a group of ER+ breast cancer patients (n= 207) (Wang, Klijn et al. 2005) were compared to TNBC patient samples (Figure 3.13). Here, a significant difference between was found between TNBCs compared to ER+ breast cancers for *USP9X, SNX3, TOP2A and YME1L1*. SLR for IQCK was not as striking as other genes. Besides, no significant difference was detected for RPL13.

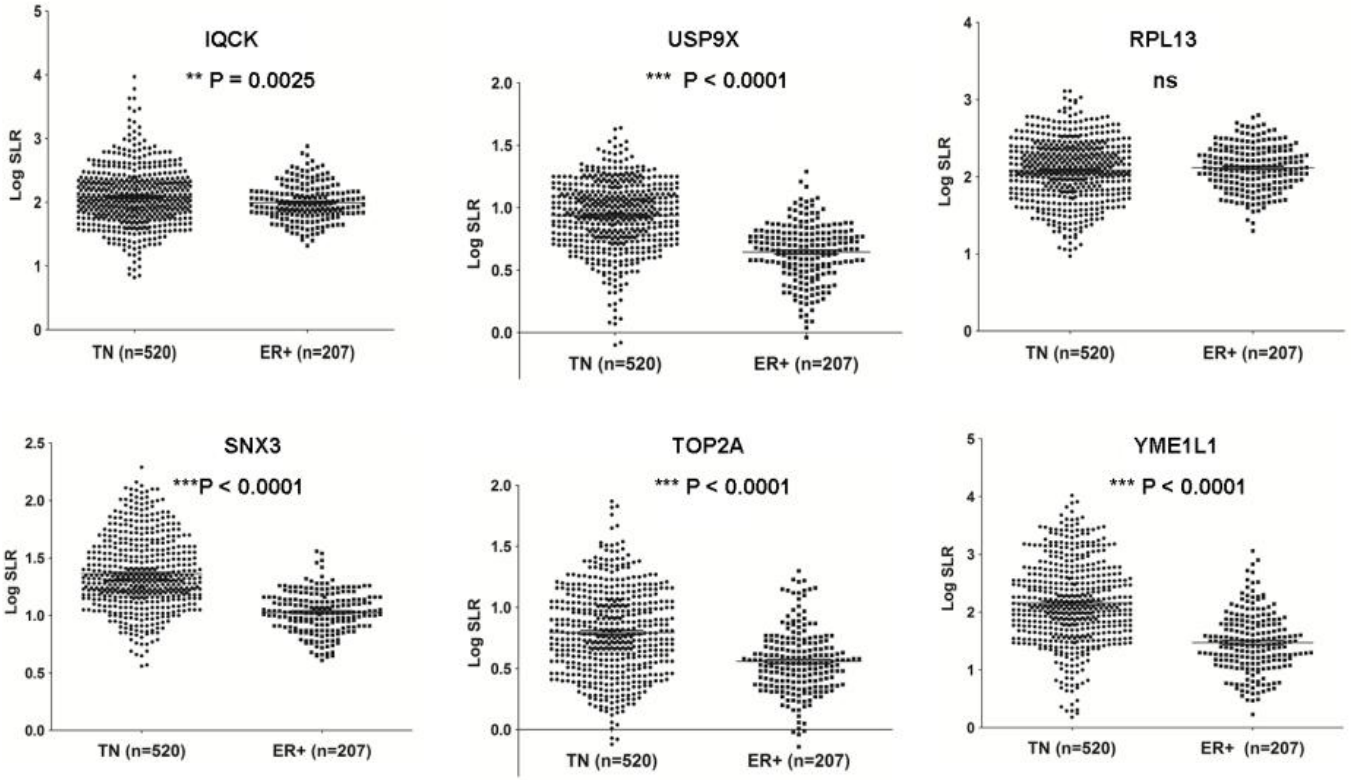

Figure 3.13. Scatter plots of transcripts undergoing 3'-UTR shortening in TNBC and ER$^+$ patients. APA based 3'UTR shortening of *IQCK*, *USP9X*, *RPL13*, *SNX3*, *TOP2A* and *YME1L1* in TNBCs (n=520) compared to ER+ patients (n=207). Scatter plots were drawn using GraphPad software. Significance of group means and variances for TNBCs and ER+ patients were calculated by t-test (Unpaired t test with Welch`s Correction).

Although a general increase in selected genes' SLR is observed in both TNBC and ER+ patients, significantly higher SLRs of 4 out of 6 genes in TNBC patients compared to ER+ patients maybe in agreement with the high proliferative rate of TNBCs compared to ER+ tumors.

On the other hand, the fact that similar genes were altered in terms of 3'-UTR lengths, we may consider the activation of same APA machinery during breast specific neoplastic transformation.

CHAPTER 4

CONCLUSION

There is growing evidence on the consequences of APA switches in different contexts. To investigate whether APA has a role in TNBC, we used a probe based method of analyzing APA dependent changes in large datasets. Alpha version of APADetect was developed to detect expression levels of different 3'-UTR isoforms in E2 treated MCF7 breast cancer cell line (Akman, Can et al. 2012). Here, we improved this validated approach for large dataset use and made it publically available for APA and 3'UTR length change determination.

DNA microarrays have been widely used to examine gene expression. The Affymetrix platforms are the most frequently used gene chips, employing DNA probes of 25 nucleotides designed to hybridize to different regions of target mRNA. The targeted region is often predisposed on the 3'-UTR, which can lead to biases in detection. While a significant number of mammalian genes undergo APA under different cellular conditions, this can lead to variable transcripts with different hybridization properties on Affymetrix platforms. This could lead to inaccurate interpretation of microarray data when the changes of expression levels actually result from alternative use of polyadenylation sites.

In this study, we provide initial evidence on deregulated APA in TNBC patients (n=520) compared to controls (n=82) using APADetect and SAM analyses. APA selection have been shown to be tissue specific (Wang, Sandberg et al. 2008), by analyzing normal breast samples with TNBC samples, we aimed to eliminate breast tissue specific changes. Thus, the genes we found to undergo APA regulation are likely to be due to deregulated molecular mechanisms in breast cancer cells.

SAM pointed out to 165 genes with significant APA changes. Out of 165 significant APA events, 68.5% (113 of 165) of them were 3'-UTR shortening and 31.5% (52 of 165) were 3'-UTR lengthening events. An ontology analysis on APA events showed molecular function of these genes was mostly either nucleic acid/protein binding or catabolically active proteins. The altered APA events detected in these group of proteins may have further amplified effects in downstream pathways such as altered transcription and metabolism regulation.

We further investigated a group of top ranking genes with 3'-UTR shortening events in detail. These genes; IQCK (IQ Domain-Containing Protein K), USP9X (Ubiquitin Specific Peptidase 9 X-Linked), RPL13 (Ribosomal Protein L13), SNX3 (Sorting Nexin-3), TOP2A (Topoisomerase DNA II alpha), and YME1L1 (YME1-like 1 ATPase) showed significant differences in the means and variances of increased SLRs in patients compared to controls. ROC analysis further confirmed the specificity and sensitivity of the individual SLR values of these six genes in their abilities to separate TNBC from controls.

To confirm the possible affects of these 3'-UTR shortening events, miRNA prediction tools were used. Results showed few microRNA predictions for short

isoforms whereas a prominent number of microRNA binding predictions were made for the longer isoforms. These findings were in agreement with the earlier observations that oncogenic transformation tends to favor 3'-UTR shortening and these shorter isoforms are more stable transcripts than longer isoforms (Mayr and Bartel 2009).

When we explored the functions and cancer relevance of these 6 high SLR genes, interesting studies were found. For instance, SNX3 is known to control Wingless/Wnt secretion through regulating retromer-dependent recycling (Zhang, Wu et al. 2011). Wnt signaling has recently been associated with metastasis in triple negative breast cancers (Dey, Barwick et al. 2013).

Ubiquitin specific proteases (USPs) have already been linked to breast cancer by various groups including ours (Akhavantabasi, Akman et al. 2010; Sacco, Coulson et al. 2010). Within this line, overexpression of USP9X was reported in a range of different cancers including breast and was associated with beta-catenin stability, TGF-beta signaling through deubiquitination of SMAD4 as well as regulation of the AMP-activated protein kinase family (Taya, Yamamoto et al. 1999; Murray, Jolly et al. 2004; Deng, Zhou et al. 2007; Al-Hakim, Zagorska et al. 2008; Dupont, Mamidi et al. 2009).

TOP2A, codes for a topoisomerase and is targeted by anthracyclines used during treatment of TNBC patients as well as other cancers (Romero, Caldés et al. 2012; Joerger and Thürlimann 2013). YME1L1 is required for apoptotic resistance, cristae morphogenesis, and cell proliferation (Stiburek, Cesnekova et al. 2012) and has been implicated in cancer development and progression and has also been identified as a

MYC-responsive gene (Guo, Malek et al. 2000; Wan, Gong et al. 2004). RPL13 is an essential component of 60S ribosomal subunit. Ribosome biogenesis is known to increase during proliferation (Pogue-Geile, Geiser et al. 1991). Accordingly, upregulation of RPL13 expression levels was detected in colorectal, gastric and liver cancers (Kobayashi, Sasaki et al. 2006). IQCK, a novel protein with IQ motif remains to be characterized in terms of its possible protein-protein interactions and its role in cancer cells.

According to relapse-free survival analysis, we found *USP9X*, *SNX3*, and *YME1L1* SLRs to correlate with higher relapse rates in TNBC patients. By taking into account the functions and relapse correlations, SLR values of USP9X and SNX3 were determined by expression analysis. A significant difference was found in TNBC cell lines compared to normal breast, supporting *in silico* analysis.

Yet another interesting finding was within the 3'-UTR lengthening events. Survival analysis of TNPO1 showed lengthening of its 3'-UTR correlates with poor prognosis in TNBC patients. Considering its function to import PABN1 to nucleus, possible negative regulations upon TNPO1 3'-UTR may lead to a global APA selection favoring proximal sites.

Additionally, when TNBC patients compared to ER+ patients, SLR values of YME1L1, SNX3, USP9X and TOP2A were higher in TNBCs. While a cancer specific APA pattern may be evident in ER+ and TNBC patients, higher SLRs of selected genes in TNBC could be an indication of higher proliferation rates and/or aggressiveness. Therefore, functional studies will be of interest to delineate a TNBC enhanced role of these genes during breast tumorigenesis. It was also of interest to

observe that although TNBC is a heterogeneous population within itself and the SLRs were significantly different from controls, there was no progressive increase of SLRs in higher grade tumors, larger tumor sizes or lymph node metastasis. These findings may be an indication of APA deregulation to be an early event in tumorigenenesis.

Based on previous and our findings, it appears APA has important regulatory roles not only in normal cells but also in cancer cells. Considering the hallmarks of cancer, increased proliferation rates and activated signaling pathways, APA may appear as a novel regulation mechanism to explain certain oncogene activation and/or tumor suppressor inactivation cases where no other explanation exists. Considering the additional complexity of alternative splicing mechanisms as well, APA is likely to help us improve the understanding of deregulated genome and its reflection on proteome in cancers. Functional studies are underway to delineate the significance of the APA events described here for TNBCs.

REFERENCES

Akhavantabasi, S., H. B. Akman, et al. (2010) USP32 is an active, membrane-bound ubiquitin protease overexpressed in breast cancers. Mamm Genome 21(7-8): 388-97.

Akman, B. H., T. Can, et al. (2012) Estrogen-induced upregulation and 3'-UTR shortening of CDC6. In: *Nucleic Acids Res*, England. 10679-88.

Akman, H. B. and A. E. Erson-Bensan (2014) Alternative polyadenylation and its impact on cellular processes. Microrna 3(1): 2-9.

Al-Hakim, A. K., A. Zagorska, et al. (2008) Control of AMPK-related kinases by USP9X and atypical Lys(29)/Lys(33)-linked polyubiquitin chains. Biochem J 411(2): 249-60.

Ambros, V. (2004) The functions of animal microRNAs. In: *Nature*, England. 350-5.

Barreau, C., L. Paillard, et al. (2005) AU-rich elements and associated factors: are there unifying principles? In: *Nucleic Acids Res*, England. 7138-50.

Barrett, T., D. B. Troup, et al. (2011) NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res 39(Database issue): D1005-10.

Bartel, D. P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. In: *Cell*, United States. 281-97.

Berrar, D. and P. Flach (2012) Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). Brief Bioinform 13(1): 83-97.

Bustin, S. A., V. Benes, et al. (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem 55(4): 611-22.

Cai, X., C. H. Hagedorn, et al. (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. In: *RNA*, United States. 1957-66.

Caldwell, S. A., S. R. Jackson, et al. (2010) Nutrient sensor O-GlcNAc transferase regulates breast cancer tumorigenesis through targeting of the oncogenic transcription factor FoxM1. Oncogene 29(19): 2831-42.

Calin, G. A. and C. M. Croce (2006) MicroRNA signatures in human cancers. In: *Nat Rev Cancer*, England. 857-66.

Calin, G. A., C. D. Dumitru, et al. (2002) Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. In: *Proc Natl Acad Sci U S A*, United States. 15524-9.

Calin, G. A., C. G. Liu, et al. (2007) Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. In: *Cancer Cell*, United States. 215-29.

Carey, L. A., E. C. Dees, et al. (2007) The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. Clin Cancer Res 13(8): 2329-34.

Cascione, L., P. Gasparini, et al. (2013) Integrated MicroRNA and mRNA Signatures Associated with Survival in Triple Negative Breast Cancer. In: *PLoS One*, United States. e55910.

Charafe-Jauffret, E., C. Ginestier, et al. (2006) Gene expression profiling of breast cell lines identifies potential new basal markers. Oncogene 25(15): 2273-84.

Chavez, K. J., S. V. Garimella, et al. (2010) Triple negative breast cancer cell lines: one tool in the search for better treatment of triple negative breast cancer. In: *Breast Dis*, Netherlands. 35-48.

Chen, C. Y. and A. B. Shyu (1995) AU-rich elements: characterization and importance in mRNA degradation. In: *Trends Biochem Sci*, England. 465-70.

Chen, F. and J. Wilusz (1998) Auxiliary downstream elements are required for efficient polyadenylation of mammalian pre-mRNAs. Nucleic Acids Res 26(12): 2891-8.

Chen, Q., Y. Chen, et al. (2013) TET2 promotes histone O-GlcNAcylation during gene transcription. Nature 493(7433): 561-4.

Cowley, M., A. J. Wood, et al. (2012) Epigenetic control of alternative mRNA processing at the imprinted Herc3/Nap1l5 locus. Nucleic Acids Res. 40: 8917-26.

Cox, E. J. and S. A. Marsh (2013) Exercise and diabetes have opposite effects on the assembly and O-GlcNAc modification of the mSin3A/HDAC1/2 complex in the heart. Cardiovasc Diabetol 12: 101.

D'Ambrogio, A., K. Nagaoka, et al. (2013) Translational control of cell growth and malignancy by the CPEBs. Nat Rev Cancer 13(4): 283-90.

D'Ippolito, E. and M. V. Iorio (2013) MicroRNAs and triple negative breast cancer. Int J Mol Sci 14(11): 22202-20.

Darmon, S. K. and C. S. Lutz (2012) Novel upstream and downstream sequence elements contribute to polyadenylation efficiency. RNA Biol 9(10): 1255-65.

de Klerk, E. (2012) Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation. Nucleic Acids Res. 40: 9089-101.

de Klerk, E., A. Venema, et al. (2012) Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation. Nucleic Acids Res 40(18): 9089-101.

Deng, S., H. Zhou, et al. (2007) Over-expression of genes and proteins of ubiquitin specific peptidases (USPs) and proteasome subunits (PSs) in breast cancer tissue observed by the methods of RFDD-PCR and proteomics. Breast Cancer Res Treat 104(1): 21-30.

Dey, N., B. G. Barwick, et al. (2013) Wnt signaling in triple negative breast cancer is associated with metastasis. BMC Cancer 13(1): 537.

Di Giammartino, D. C., K. Nishida, et al. (2011) Mechanisms and consequences of alternative polyadenylation. In: *Mol Cell*. 2011 Elsevier Inc, United States. 853-66.

Dupont, S., A. Mamidi, et al. (2009) FAM/USP9x, a deubiquitinating enzyme essential for TGFbeta signaling, controls Smad4 monoubiquitination. Cell 136(1): 123-35.

Elkon, R. (2012) E2F mediates enhanced alternative polyadenylation in proliferation. Genome Biol. 13: R59.

Elkon, R., A. P. Ugalde, et al. (2013) Alternative cleavage and polyadenylation: extent, regulation and function. Nat Rev Genet 14(7): 496-506.

Erlitzki, R., J. C. Long, et al. (2002) Multiple, conserved iron-responsive elements in the 3'-untranslated region of transferrin receptor mRNA enhance binding of iron regulatory protein 2. In: *J Biol Chem*, United States. 42579-87.

Esquela-Kerscher, A. and F. J. Slack (2006) Oncomirs - microRNAs with a role in cancer. In: *Nat Rev Cancer*, England. 259-69.

Faraji, F., Y. Hu, et al. (2014) An integrated systems genetics screen reveals the transcriptional structure of inherited predisposition to metastatic disease. Genome Res 24(2): 227-40.

Ferrer, C. M., T. P. Lynch, et al. (2014) O-GlcNAcylation regulates cancer metabolism and survival stress signaling via regulation of the HIF-1 pathway. Mol Cell 54(5): 820-31.

Filipowicz, W., S. N. Bhattacharyya, et al. (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? In: *Nat Rev Genet*, England. 102-14.

Fleige, S., V. Walf, et al. (2006) Comparison of relative mRNA quantification models and the impact of RNA integrity in quantitative real-time RT-PCR. Biotechnol Lett 28(19): 1601-13.

Forma, E., P. Jozwiak, et al. (2014) The potential role of O-GlcNAc modification in cancer epigenetics. Cell Mol Biol Lett 19(3): 438-60.

Fritz, J., A. Strehblow, et al. (2009) RNA-regulated interaction of transportin-1 and exportin-5 with the double-stranded RNA-binding domain regulates nucleocytoplasmic shuttling of ADAR1. Mol Cell Biol 29(6): 1487-97.

Fujita, P. A., B. Rhead, et al. (2011) The UCSC Genome Browser database: update 2011. Nucleic Acids Res 39(Database issue): D876-82.

Garcia, D. M., D. Baek, et al. (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. In: *Nat Struct Mol Biol*, United States. 1139-46.

Garneau, N. L., J. Wilusz, et al. (2007) The highways and byways of mRNA decay. In: *Nat Rev Mol Cell Biol*, England. 113-26.

Ge, X., S. Yamamoto, et al. (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. Genomics 86(2): 127-41.

Ghosh, T., K. Soni, et al. (2008) MicroRNA-mediated up-regulation of an alternatively polyadenylated variant of the mouse cytoplasmic {beta}-actin gene. Nucleic Acids Res 36(19): 6318-32.

Goel, M. K., P. Khanna, et al. (2010) Understanding survival analysis: Kaplan-Meier estimate. Int J Ayurveda Res 1(4): 274-8.

Graham, K., A. de las Morenas, et al. (2010) Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. Br J Cancer 102(8): 1284-93.

Grimson, A., K. K. Farh, et al. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. In: *Mol Cell*, United States. 91-105.

Gruber, A. R., G. Martin, et al. (2012) Cleavage factor Im is a key regulator of 3' UTR length. RNA Biol 9(12): 1405-12.

Guhaniyogi, J. and G. Brewer (2001) Regulation of mRNA stability in mammalian cells. Gene 265(1-2): 11-23.

Guo, Q. M., R. L. Malek, et al. (2000) Identification of c-myc responsive genes using rat cDNA microarray. Cancer Res 60(21): 5922-8.

Gur-Dedeoglu, B., O. Konu, et al. (2009) Identification of endogenous reference genes for qRT-PCR analysis in normal matched breast tumor tissues. Oncol Res 17(8): 353-65.

Hall-Pogar, T., S. Liang, et al. (2007) Specific trans-acting proteins interact with auxiliary RNA polyadenylation elements in the COX-2 3'-UTR. Rna 13(7): 1103-15.

Hall-Pogar, T., H. Zhang, et al. (2005) Alternative polyadenylation of cyclooxygenase-2. Nucleic Acids Res 33(8): 2565-79.

Hanahan, D. and R. A. Weinberg (2011) Hallmarks of cancer: the next generation. In: *Cell*. 2011 Elsevier Inc, United States. 646-74.

He, L., J. M. Thomson, et al. (2005) A microRNA polycistron as a potential human oncogene. In: *Nature*, England. 828-33.

Huang da, W., B. T. Sherman, et al. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37(1): 1-13.

Huang da, W., B. T. Sherman, et al. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1): 44-57.

Huang, X., Q. Pan, et al. (2013) O-GlcNAcylation of cofilin promotes breast cancer cell invasion. J Biol Chem 288(51): 36418-25.

Jacobson, A. and S. W. Peltz (1996) Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. Annu Rev Biochem 65: 693-739.

Jan, C. H., R. C. Friedman, et al. (2011) Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. In: *Nature*, England. 97-101.

Jenal, M., R. Elkon, et al. (2012) The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. Cell 149(3): 538-53.

Ji, X., J. Kong, et al. (2011) An RNA-protein complex links enhanced nuclear 3' processing with cytoplasmic mRNA stabilization. In: *EMBO J*, England. 2622-33.

Ji, Z., J. Y. Lee, et al. (2009) Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc Natl Acad Sci U S A 106(17): 7028-33.

Ji, Z., W. Luo, et al. (2011) Transcriptional activity regulates alternative cleavage and polyadenylation. Mol Syst Biol 7: 534.

Ji, Z. and B. Tian (2009) Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. PLoS One 4(12): e8419.

Joerger, M. and B. Thürlimann (2013) Chemotherapy regimens in early breast cancer: major controversies and future outlook. Expert Rev Anticancer Ther 13(2): 165-78.

Johnson, S. M., H. Grosshans, et al. (2005) RAS is regulated by the let-7 microRNA family. In: *Cell*, United States. 635-47.

Kambo, A., V. S. Sharma, et al. (2005) Nitric oxide inhibits mammalian methylmalonyl-CoA mutase. J Biol Chem 280(11): 10073-82.

Kertesz, M., N. Iovino, et al. (2007) The role of site accessibility in microRNA target recognition. Nat Genet 39(10): 1278-84.

Kim, S. Y., Y. H. Lee, et al. (2012) MiR-186, miR-216b, miR-337-3p, and miR-760 cooperatively induce cellular senescence by targeting alpha subunit of protein kinase CKII in human colorectal cancer cells. Biochem Biophys Res Commun 429(3-4): 173-9.

Kobayashi, T., Y. Sasaki, et al. (2006) Activation of the ribosomal protein L13 gene in human gastrointestinal cancer. Int J Mol Med 18(1): 161-70.

Kretschmer, C., A. Sterner-Kock, et al. (2011) Identification of early molecular markers for breast cancer. Mol Cancer 10(1): 15.

Krzywinski, M., J. Schein, et al. (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19(9): 1639-45.

Kuhn, U., M. Gundel, et al. (2009) Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. In: *J Biol Chem*, United States. 22803-14.

Lakhani, S. R. (2014) Keynote Lecture: KN02 UPDATE ON WHO BREAST CLASSIFICATION - 2 YEARS ON. Pathology 46 Suppl 2: S1.

Lazarus, M. B., J. Jiang, et al. (2013) HCF-1 is cleaved in the active site of O-GlcNAc transferase. Science 342(6163): 1235-9.

Ledley, F. D., M. R. Lumetta, et al. (1988) Mapping of human methylmalonyl CoA mutase (MUT) locus on chromosome 6. Am J Hum Genet 42(6): 839-46.

Lee, Y., C. Ahn, et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. In: *Nature*, England. 415-9.

Lee, Y., M. Kim, et al. (2004) MicroRNA genes are transcribed by RNA polymerase II. In: *EMBO J*, England. 4051-60.

Lee, Y. H., S. Y. Kim, et al. (2014) Upregulation of miR-760 and miR-186 is associated with replicative senescence in human lung fibroblast cells. Mol Cells 37(8): 620-7.

Lewis, B. P., C. B. Burge, et al. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. In: *Cell*, United States. 15-20.

Licatalosi, D. D. and R. B. Darnell (2010) RNA processing and its regulation: global insights into biological networks. In: *Nat Rev Genet*, England. 75-87.

Liu, R., X. Wang, et al. (2007) The prognostic role of a gene signature from tumorigenic breast-cancer cells. N Engl J Med 356(3): 217-26.

Lockhart, D. J., H. Dong, et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14(13): 1675-80.

Lu, J., G. Getz, et al. (2005) MicroRNA expression profiles classify human cancers. In: *Nature*, England. 834-8.

Lv, J., K. Xia, et al. (2014) miRNA expression patterns in chemoresistant breast cancer tissues. Biomed Pharmacother.

Ma, Z. and K. Vosseller (2014) Cancer Metabolism and Elevated O-GlcNAc in Oncogenic Signaling. J Biol Chem.

Mallet, P. L. and F. Bachand (2013) A proline-tyrosine nuclear localization signal (PY-NLS) is required for the nuclear import of fission yeast PAB2, but not of human PABPN1. Traffic 14(3): 282-94.

Martin, G., A. R. Gruber, et al. (2012) Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. Cell Rep 1(6): 753-63.

Mayr, C. and D. P. Bartel (2009) Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 138(4): 673-84.

Mayr, C. and D. P. Bartel (2009) Widespread shortening of 3[prime]UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 138: 673-84.

Mayr, C. and D. P. Bartel (2009) Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 138: 673-84.

Mayr, C., M. T. Hemann, et al. (2007) Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation. Science 315(5818): 1576-9.

Metz, C. E. (1978) Basic principles of ROC analysis. Semin Nucl Med 8(4): 283-98.

Mochmann, L. H., M. Neumann, et al. (2014) ERG induces a mesenchymal-like state associated with chemoresistance in leukemia cells. Oncotarget 5(2): 351-62.

Murray, R. Z., L. A. Jolly, et al. (2004) The FAM deubiquitylating enzyme localizes to multiple points of protein trafficking in epithelia, where it associates with E-cadherin and beta-catenin. Mol Biol Cell 15(4): 1591-9.

Nagaike, T., C. Logan, et al. (2011) Transcriptional activators enhance polyadenylation of mRNA precursors. Mol Cell 41(4): 409-18.

Nunes, N. M., W. Li, et al. (2010) A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence. Embo j 29(9): 1523-36.

O'Donnell, K. A., E. A. Wentzel, et al. (2005) c-Myc-regulated microRNAs modulate E2F1 expression. In: *Nature*, England. 839-43.

Oosterkamp, H. M., E. M. Hijmans, et al. (2014) USP9X downregulation renders breast cancer cells resistant to tamoxifen. Cancer Res 74(14): 3810-20.

Ozsolak, F., P. Kapranov, et al. (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. In: *Cell*. 2010 Elsevier Inc, United States. 1018-29.

Pinto, P. A. (2011) RNA polymerase II kinetics in polo polyadenylation signal selection. EMBO J. 30: 2431-44.

Piovan, C., D. Palmieri, et al. (2012) Oncosuppressive role of p53-induced miR-205 in triple negative breast cancer. Mol Oncol 6(4): 458-72.

Planche, A., M. Bacac, et al. (2011) Identification of prognostic molecular features in the reactive stroma of human breast and prostate cancer. PLoS One 6(5): e18640.

Pogue-Geile, K., J. R. Geiser, et al. (1991) Ribosomal protein genes are overexpressed in colorectal cancer: isolation of a cDNA clone encoding the human S3 ribosomal protein. Mol Cell Biol 11(8): 3842-9.

Proudfoot, N. J. (2011) Ending the message: poly(A) signals then and now. In: *Genes Dev*, United States. 1770-82.

Proudfoot, N. J., A. Furger, et al. (2002) Integrating mRNA processing with transcription. In: *Cell*, United States. 501-12.

Radojicic, J., A. Zaravinos, et al. (2011) MicroRNA expression analysis in triple-negative (ER, PR and Her2/neu) breast cancer. Cell Cycle 10(3): 507-17.

Reimers, K., C. Y. Choi, et al. (2008) The Bax Inhibitor-1 (BI-1) family in apoptosis and tumorigenesis. Curr Mol Med 8(2): 148-56.

Richard, P. and J. L. Manley (2009) Transcription termination by nuclear RNA polymerases. In: *Genes Dev*, United States. 1247-69.

Richardson, A. L., Z. C. Wang, et al. (2006) X chromosomal abnormalities in basal-like human breast cancer. Cancer Cell 9(2): 121-32.

Rody, A., T. Karn, et al. (2011) A clinically relevant gene signature in triple negative and basal-like breast cancer. Breast Cancer Res 13(5): R97.

Rojas-Rivera, D. and C. Hetz (2014) TMBIM protein family: ancestral regulators of cell death. Oncogene.

Romero, A., T. Caldés, et al. (2012) Topoisomerase 2 alpha: a real predictor of anthracycline efficacy? Clin Transl Oncol 14(3): 163-8.

Rosenfeld, N., R. Aharonov, et al. (2008) MicroRNAs accurately identify cancer tissue origin. In: *Nat Biotechnol*, United States. 462-9.

Sacco, J. J., J. M. Coulson, et al. (2010) Emerging roles of deubiquitinases in cancer-associated pathways. IUBMB Life 62(2): 140-57.

Saeed, A. I., V. Sharov, et al. (2003) TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34(2): 374-8.

Sandberg, R., J. R. Neilson, et al. (2008) Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science 320(5883): 1643-7.

Saraiva, N., D. L. Prole, et al. (2013) hGAAP promotes cell adhesion and migration via the stimulation of store-operated Ca2+ entry and calpain 2. J Cell Biol 202(4): 699-713.

Shi, Y. (2012) Alternative polyadenylation: New insights from global analyses. In: *RNA*.

Shi, Y., D. C. Di Giammartino, et al. (2009) Molecular architecture of the human pre-mRNA 3' processing complex. In: *Mol Cell*, United States. 365-76.

Soule, H. D., T. M. Maloney, et al. (1990) Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. Cancer Res 50(18): 6075-86.

Spies, N., C. B. Nielsen, et al. (2009) Biased chromatin signatures around polyadenylation sites and exons. Mol Cell 36(2): 245-54.

Stiburek, L., J. Cesnekova, et al. (2012) YME1L controls the accumulation of respiratory chain subunits and is required for apoptotic resistance, cristae morphogenesis, and cell proliferation. Mol Biol Cell 23(6): 1010-23.

Takagaki, Y. and J. L. Manley (1998) Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation. In: *Mol Cell*, United States. 761-71.

Tanaka, T., S. Sugaya, et al. (2012) Inhibition of cell viability by human IFN-beta is mediated by microRNA-431. Int J Oncol 40(5): 1470-6.

Taya, S., T. Yamamoto, et al. (1999) The deubiquitinating enzyme Fam interacts with and stabilizes beta-catenin. Genes Cells 4(12): 757-67.

Tian, B., J. Hu, et al. (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. In: *Nucleic Acids Res*, England. 201-12.

Tripathi, A., C. King, et al. (2008) Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients. Int J Cancer 122(7): 1557-66.

Tusher, V. G., R. Tibshirani, et al. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98(9): 5116-21.

Twyffels, L., C. Gueydan, et al. (2014) Transportin-1 and Transportin-2: protein nuclear import and beyond. FEBS Lett 588(10): 1857-68.

van Bakel, H. and F. C. Holstege (2004) In control: systematic assessment of microarray performance. EMBO Rep 5(10): 964-9.

Venkataraman, K., K. M. Brown, et al. (2005) Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. Genes Dev 19(11): 1315-27.

Vuong, D., P. T. Simpson, et al. (2014) Molecular classification of breast cancer. Virchows Arch 465(1): 1-14.

Wahle, E. (1991) A novel poly(A)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation. In: *Cell*, United States. 759-68.

Wan, D., Y. Gong, et al. (2004) Large-scale cDNA transfection screening for genes related to cancer development and progression. Proc Natl Acad Sci U S A 101(44): 15724-9.

Wang, C., X. Zheng, et al. (2012) MicroRNA-203 suppresses cell proliferation and migration by targeting BIRC5 and LASP1 in human triple-negative breast cancer cells. J Exp Clin Cancer Res 31: 58.

Wang, E. T., R. Sandberg, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456(7221): 470-6.

Wang, Y., J. G. Klijn, et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365(9460): 671-9.

Weigelt, B. and J. S. Reis-Filho (2009) Histological and molecular types of breast cancer: is there a unifying taxonomy? Nat Rev Clin Oncol 6(12): 718-30.

Wickens, M., P. Anderson, et al. (1997) Life and death in the cytoplasm: messages from the 3' end. In: *Curr Opin Genet Dev*, England. 220-32.

Wood, A. J. (2007) A screen for retrotransposed imprinted genes reveals an association between X chromosome homology and maternal germ-line methylation. PLoS Genet. 3: e20.

Wood, A. J., R. Schulz, et al. (2008) Regulation of alternative polyadenylation by genomic imprinting. Genes Dev 22(9): 1141-6.

Wray, N. R., J. Yang, et al. (2010) The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet 6(2): e1000864.

Xie, Y., M. Avello, et al. (2013) Deubiquitinase FAM/USP9X interacts with the E3 ubiquitin ligase SMURF1 protein and protects it from ligase activity-dependent self-degradation. J Biol Chem 288(5): 2976-85.

Xu, Y., B. Iglewicz, et al. (2014) Robust Estimation of the Parameters of Distributions, with Applications to Outlier Detection. Comput Stat Data Anal 75: 66-80.

Yang, Q., M. Coseno, et al. (2011) Crystal structure of a human cleavage factor CFI(m)25/CFI(m)68/RNA complex provides an insight into poly(A) site recognition and RNA looping. Structure 19: 368-77.

Yang, Q., G. M. Gilmartin, et al. (2011) The structure of human cleavage factor I(m) hints at functions beyond UGUA-specific RNA binding: a role in alternative polyadenylation and a potential link to 5' capping and splicing. RNA Biol 8(5): 748-53.

Yao, C., J. Biesinger, et al. (2012) Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. Proc Natl Acad Sci U S A 109(46): 18773-8.

Ye, W., Q. Lv, et al. (2008) The effect of central loops in miRNA:MRE duplexes on the efficiency of miRNA-mediated gene regulation. PLoS One 3(3): e1719.

Yi, R., Y. Qin, et al. (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. In: *Genes Dev*, United States. 3011-6.

Zhang, H., J. Hu, et al. (2005) PolyA_DB: a database for mammalian mRNA polyadenylation. Nucleic Acids Res 33(Database issue): D116-20.

Zhang, P., Y. Wu, et al. (2011) SNX3 controls Wingless/Wnt secretion through regulating retromer-dependent recycling of Wntless. Cell Res 21(12): 1677-90.

Zsengeller, Z. K., N. Aljinovic, et al. (2014) Methylmalonic acidemia: a megamitochondrial disorder affecting the kidney. Pediatr Nephrol 29(11): 2139-46.

APPENDIX A

PATIENT CHARACTERISTICS

Table A.1. Characteristics of TNBC dataset.

Nodal status: 0: negative, 1: positive
Size of the tumor: 1: up to 1 cm, 2: >1cm
Grade of the tumor: 12: G1 or G2, 3: G3
Month: event-free interval
Event: 1: yes, 0: no

GSM_accession	Nodal status	Age	T1vs2	G12vs3	Month	Event	Chemotherapy
GSM782523	0	59	2	3	93	0	NO
GSM782524	0	65	2	12	37	0	NO
GSM782525	1	66	2	12	31	0	NO
GSM782526	0						NO
GSM782527	1						NO
GSM782528	0	60	2	12	120	0	NO
GSM782529	0	57	2	3	25	1	NO
GSM782530	1	46	1	12	23	1	YES
GSM782531	0	55	2	3	21	1	YES
GSM782532	0	61	2	3	45	0	YES
GSM782533	0	29	1	3	45	1	YES
GSM782534		69	1	3	52	0	YES
GSM782535	1	47	2	3	16	1	YES
GSM782536	0	51	1	12	70	0	YES
GSM782537	0	51	2	12	77	0	YES
GSM782538	0	57	2	3	47	0	YES
GSM782539	1	80	2	3	11	1	YES
GSM782540	0	48	2	3	76	0	YES
GSM782541	0	53	2	3	69	0	YES
GSM782542	0	53	1	3	28	0	YES
GSM782543	0	67	2	12	34	0	YES

Table A.1. Characteristics of TNBC dataset continued

GSM_accession	Nodal status	Age	T1vs2	G12vs3	Month	Event	Chemotherapy
GSM782544	0	56	2	3	10	1	YES
GSM782545	1	53	2	3	29	0	YES
GSM782546	0	63	1	3	36	0	YES
GSM782547	0	47	2	3	31	0	YES
GSM782548	0	45	1	3	36	0	YES
GSM782549	0	42	2	3	6	1	YES
GSM782550	0	51	1	3	33	0	YES
GSM782551	0	54	2	3	10	1	YES
GSM782552	0	40	1	3	27	0	YES
GSM782553	0	57	1	3	45	0	YES
GSM79115	0	40	1	3	120	0	NO
GSM79117	1	52	2	3	120	0	
GSM79122	0	74	1	12	120	0	NO
GSM79145	0	40	2	12	120	0	NO
GSM79147	1	34	2	3	8	0	
GSM79165	0	62	2	12	120	0	NO
GSM79191	0	58	2	12	120	0	
GSM79195	1	68	1	12	120	0	
GSM79196	1	77	2	3	43	0	NO
GSM79225	0	70	1	12	95	0	NO
GSM79231	1	77	2	12	6	1	NO
GSM79251		88	2	3	2	0	NO
GSM79253	0	75	1	3	30	1	NO
GSM79255	0	77	2	3	37	1	NO
GSM79270	0	32	1	12	120	0	NO
GSM79271	1	46	2	3	0	0	NO
GSM79280	0	50	1	3	120	0	NO
GSM79287	1	37	1	3	120	0	
GSM79292	0	84	2	3	11	0	
GSM79299	0	37	2	3	119	0	NO
GSM79303	0	44	1	12	118	0	NO
GSM79306	0	38	2	3	2	1	NO
GSM79322	1	61	2	12	115	1	NO
GSM79329	0	68	1	3	7	1	NO
GSM79336	0	77	2	3	3	1	NO
GSM79344		87	1	3	120	0	NO
GSM79356	0	61	2	12	120	0	NO
GSM782554	1	52	2	3	6	1	YES
GSM782555	0	41	2	3	67	0	YES
GSM782556	1	41	2	3	11	1	YES

Table A.1. Characteristics of TNBC dataset continued

GSM_accession	Nodal status	Age	T1vs2	G12vs3	Month	Event	Chemotherapy
GSM782557	0	30	2	3	51	0	YES
GSM782558	0	59	1	12	48	0	YES
GSM782559	1	71	2	3	58	0	YES
GSM782560	0	43	2	3	3	1	YES
GSM782561	0	60	1	3	48	0	YES
GSM782562	0	31	1	3	10	1	YES
GSM782563	0	64	2	3	120	0	YES
GSM782564	0	33	1	3	74	1	YES
GSM782565	0	57	2		104	0	YES
GSM782566	0	32	2	3	22	1	YES
GSM782567	0	45	1	12	53	0	YES
GSM782568	0	56	2	3	37	1	YES
GSM65839	0	73	2	3	8	1	NO
GSM65841	0	64	1	12	120	0	NO
GSM65842	0	47	2	3	120	0	NO
GSM65843	0	39	2	3	84	1	NO
GSM65845	0	32	2	3	120	0	NO
GSM65852	0	53	2		42	1	NO
GSM65861	0	71	1	12	32	1	NO
GSM65865	0	43	2	3	120	0	NO
GSM65876	0	52	1	12	63	0	NO
GSM65878	0	38	1	12	9	1	NO
GSM65879	0	47	1		42	0	NO
GSM150795	0	51	1				NO
GSM150797	0	29	2	3			NO
GSM107076				3	15	1	NO
GSM107084				3	13	1	
GSM107094					49	1	NO
GSM107114				3	11	1	NO
GSM107117				3	76	0	NO
GSM107120				3	90	0	NO
GSM107124				3	98	0	NO
GSM107128				12	92	0	NO
GSM107131				3	76	0	NO
GSM107143				12	89	0	NO
GSM107144				3	79	0	NO
GSM107148				12	73	0	NO
GSM107150				12	75	1	NO
GSM107153				3	87	0	NO
GSM107157				12	100	0	NO

Table A.1. Characteristics of TNBC dataset continued

GSM_accession	Nodal status	Age	T1vs2	G12vs3	Month	Event	Chemotherapy
GSM107167				3	98	0	NO
GSM107181				3	13	1	
GSM107201				12	95	0	NO
GSM107204				3	82	0	NO
GSM107206				3	17	1	NO
GSM107210				3	77	1	NO
GSM107211				3	86	0	
GSM107217					18	1	NO
GSM107224				3	72	0	NO
GSM107225				3	19	1	NO
GSM38092		55					
GSM46933		65					
GSM46938		65					
GSM46952		45					
GSM46968	0	45	2	3			
GSM53031		75					
GSM53033		65					
GSM53172	1	55	2	3			
GSM53187		65					
GSM89023	1	45	2	3			
GSM102506	1	35	1	3			
GSM102525	0	55	2	3			
GSM102526	0	45	1	12			
GSM102545	1	65	1	12			
GSM102564	0	45	2	3			
GSM117581	1	45	1	3			
GSM117661	1	55	2	3			
GSM117692	0	65	2	3			
GSM138011		85					
GSM152698	1	75	2	3			
GSM179854							
GSM179885							
GSM50035	1	61	2		22	1	YES
GSM50040	0	57	2		91	0	YES
GSM50041	0	73	2				
GSM50045	0	58	2		120	0	YES
GSM50046	0	32	1		96	0	YES
GSM50048	0	49	1		73	0	YES
GSM50050	0	63	1				
GSM50052	0	47	2		55	0	YES

Table A.1. Characteristics of TNBC dataset continued

GSM_accession	Nodal status	Age	T1vs2	G12vs3	Month	Event	Chemotherapy
GSM50055	0	33	2				
GSM50057	0	46	2				
GSM50058	0	51	2				
GSM50060	0	52	2		42	1	YES
GSM50061	0	61	2		48	1	YES
GSM50063	0	49	2		37	1	YES
GSM50064	0	70	2		11	1	YES
GSM50066	0	80	2		15	1	YES
GSM50067	0	78	2		9	1	
GSM50069	0	73	2		46	1	YES
GSM50070	1	58	2		40	1	YES
GSM50071	1	78	2		70	0	NO
GSM50073	1	50	2		87	0	YES
GSM50075	1	49	2		88	0	YES
GSM50079	1	40	2		78	0	YES
GSM50089	1	30	2		105	0	
GSM50092	1	50	2		77	0	YES
GSM50093	1	64	2		94	0	YES
GSM50094	1	57	1		39	1	YES
GSM50096	1	32	2		17	1	YES
GSM50102	1	48	2		8	1	YES
GSM50106	1	60	2		19	1	YES
GSM50107	1	52	2		75	0	YES
GSM50112	1	43	2		14	1	YES
GSM50119	1	54	2		50	0	YES
GSM50123	1	59	2		47	0	YES
GSM50125	1	59	2				
GSM151295	1	53	2		6	1	NO
GSM151309	1	50	2	3	120	0	NO
GSM36788	0	40	2	12	79	0	NO
GSM36793	0	36	2	12	101	0	NO
GSM36795	0	71	2	12	88	0	NO
GSM36797	0	44	1	3	9	1	NO
GSM36798	0	41	1	3	106	0	NO
GSM36809	0	69	1	3	56	0	NO
GSM36822	0	73	2	3	50	0	NO
GSM36824	0	70	2	3	86	0	NO
GSM36828	0	51	2	3	76	0	NO
GSM36835	0	68	2	3	14	1	NO
GSM36846	0	36	1	3	103	0	NO

Table A.1. Characteristics of TNBC dataset continued

GSM_accession	Nodal status	Age	T1vs2	G12vs3	Month	Event	Chemotherapy
GSM36855	0	34	2	3	120	0	NO
GSM36862	0	42	2	3	25	1	NO
GSM36876	0	42	2	3	120	0	NO
GSM36883	0	73	2	12	88	0	NO
GSM36889	0	44	2	3	108	0	NO
GSM36890	0	64	2	12	116	0	NO
GSM36891	0	65	1	3	87	0	NO
GSM36901	0	29	2	3	101	0	NO
GSM36905	0	47	1	12	17	1	NO
GSM36906	0	47	1	3	101	0	NO
GSM36909	0	78	2	12	95	0	NO
GSM36912	0	53	2	3	97	0	NO
GSM36923	0	46	2	3	23	1	NO
GSM36927	0	50	2	3	8	1	NO
GSM36931	0	49	2	12	37	1	NO
GSM36935	0	50	1	3	114	0	NO
GSM36949	0	53	2	3	6	1	NO
GSM36952	0	46	1	3	13	1	NO
GSM36959	0	75	2	3	120	0	NO
GSM36960	0	48	1	12	51	1	NO
GSM36961	0	68	1	12	120	0	NO
GSM36966	0	70	2	3	90	0	NO
GSM36969	0	41	1	3	32	1	NO
GSM36977	0	72	2	3	120	0	NO
GSM36981	0	60	2	3	87	0	NO
GSM36991	0	66	1	3	114	0	NO
GSM37002	0	53	2	3	16	1	NO
GSM37017	0	41	2	3	108	0	NO
GSM37021	0	58	2	12	109	0	NO
GSM37022	0	58	1	12	30	1	NO
GSM37044	0	74	2	3	112	0	NO
GSM37045	0	43	1	3	120	0	NO
GSM37047	0	44	1	3	108	0	NO
GSM37048	0	43	1	3	87	0	NO
GSM37050	0	69	1	12	18	1	NO
GSM37051	0	43	2	3	18	1	NO
GSM120649	0	53	1	12	120	0	NO
GSM120651	0	34	2	3	88	0	NO
GSM120655	0	43	2	3	10	1	NO
GSM120656	0	38	2	3	22	1	NO

Table A.1. Characteristics of TNBC dataset continued

GSM_accession	Nodal status	Age	T1vs2	G12vs3	Month	Event	Chemotherapy
GSM120657	0	69	2	3	61	0	NO
GSM120658	0	42	1	3	120	0	NO
GSM120660	0	38	1	3	116	0	NO
GSM120663	0	49	2	3	29	1	NO
GSM120666	0	34	2	12	87	0	NO
GSM120668	0	35	1	3	7	1	NO
GSM120671	0	32	2	12	24	1	NO
GSM120672	0	47	2	12	120	0	NO
GSM120674	0	63	2	3	86	0	NO
GSM120675	0	44	1	3	119	0	NO
GSM120677	0	82	2	3	68	0	NO
GSM120679	0	74	2	3	42	1	NO
GSM120682	0	58	1	3	76	0	NO
GSM120683	0	63	1	3	120	0	NO
GSM120684	0	44	2	12	120	0	NO
GSM120685	0	57	1	3	88	0	NO
GSM120686	0	66	1	3	81	0	NO
GSM120687	0	41	2	12	9	1	NO
GSM120688	0	44	2	3	120	0	NO
GSM120690	0	58	2	3	88	0	NO
GSM120691	0	61	2	12	103	0	NO
GSM120692	0	39	2	3	89	0	NO
GSM120695	0	38	2	12	90	0	NO
GSM120696	0	33	2	3	120	0	NO
GSM120698	0	57	2	12	61	0	NO
GSM120699	0	40	1	3	80	0	NO
GSM120701	0	50	2	3	69	0	NO
GSM120702	0	39	2	3	4	1	NO
GSM120703	0	59	2	12	91	0	NO
GSM120705	0	73	2	3	83	0	NO
GSM120706	0	46	1	3	102	0	NO
GSM782569	0	65	2	12	11	1	
GSM782570	1	32	2	3	39	1	
GSM782571	0	56	2	3	45	0	
GSM782572	0	55	2	12	45	0	
GSM782573	1	44	2	12	23	1	
GSM782574	0	67	2	12	34	1	
GSM782575	1	46	2	3	15	1	
GSM782576	0	60	2	3	51	0	
GSM782577	1	63	2	12	44	0	

Table A.1. Characteristics of TNBC dataset continued

GSM_accession	Nodal status	Age	T1vs2	G12vs3	Month	Event	Chemotherapy
GSM782578	1	60	2	3	30	0	
GSM782579	1	44	2	12	27	0	
GSM782580	1	69	2	12	26	0	
GSM782581	1	50	2	3	25	0	
GSM782582	1	44	2	3	10	1	
GSM782583							
GSM782584	0	39	2	3	23	0	
GSM782585	0	41	2	3	23	0	
GSM782586	1	44	2	3	23	0	
GSM782587	1	42	2	12	24	0	
GSM505351	1	50	2	3			
GSM505331	0	75	2	3			
GSM505460	0	52	2	3			
GSM505345	0	65	1	3			
GSM505369	0	59	2	3			
GSM505343	1	66	2	3			
GSM505372	0	38	2	3			
GSM505358	0	38	1	3			
GSM505336	1	61	1	3			
GSM505333	1	29	2	3			
GSM505350	1	67	2	12			
GSM505370	1	35	1	3			
GSM505359	0	48	2	3			
GSM505360	1	52	2	3			
GSM505357	1	44	2	3			
GSM505421	1	60	2	3			
GSM505417	1	51	2	3			
GSM505423	1	48	2	3			
GSM505411	1	46	2	3			
GSM505413	1	51	2	12			
GSM505435	1	71	1	12			
GSM505487	1	63	1	3			
GSM505459	1	62	2	3			
GSM505467	1	46	2	12			
GSM505428	1	38	2	3			
GSM505449	1	41	2	3			
GSM505473	1	62	2	3			
GSM505462	1	50	2	12			
GSM505450	1	58	2	3			
GSM232199	0	73	1	3	120	0	NO

Table A.1. Characteristics of TNBC dataset continued

GSM_accession	Nodal status	Age	T1vs2	G12vs3	Month	Event	Chemotherapy
GSM232204	0	82	1		120	0	NO
GSM26871	0		2	3			
GSM26878	1		2	12			
GSM26880	0		2	3			
GSM26882	1		2	3			
GSM26883	0		2				
GSM26884	1		2	12			
GSM26886	1		2	12			
GSM26887	0		2	12			
GSM26888	0			3			
GSM26889	1		2	12			
GSM26892	1		2	3			
GSM26893	1		2	3			
GSM26895	0		2	3			
GSM26898	1		2	3			
GSM26902	1		2	3			
GSM26903	0		2				
GSM26905	1		2	12			
GSM26906	1		2	3			
GSM26908	0		2	3			
GSM26910	1		2	3			
GSM26912	0		2	3			
GSM282385	0	57	2	12	89	0	NO
GSM282398	0	50	2	3	15	1	NO
GSM282413	0	68	1	3	17	1	NO
GSM282427	0	56	2	3	120	0	NO
GSM282435	0	40	2	3	115	0	NO
GSM282440	0	52	2	3	106	0	NO
GSM282446	0	67	1	12	114	0	NO
GSM282454	0	43	1	3	8	1	NO
GSM282457	0	54	1	3	120	0	NO
GSM282464	0	49	2	3	120	0	NO
GSM282465	0	55	2	3	90	0	NO
GSM282474	0	45	2	3	120	0	NO
GSM282482	0	58	2	12	116	1	NO
GSM282493	0	64	2	12	120	0	NO
GSM282497	0	65	2	3	93	0	NO
GSM282511	0	58	2	3	120	0	NO
GSM282528	0	69	1	12	120	0	NO
GSM282535	0	34	1	3	37	1	NO

Table A.1. Characteristics of TNBC dataset continued

GSM_accession	Nodal status	Age	T1vs2	G12vs3	Month	Event	Chemotherapy
GSM282551	0	58	1	12	8	1	NO
GSM282565	0	55	2	3	78	1	NO
GSM282569	0	50	2	3	10	1	NO
GSM125130							NO
GSM505513	1	42	2	3			
GSM505514	1	44	2	3			
GSM505516	1	39	2	3			
GSM505517	1	59	2	12			
GSM505526	1	68	2	12			
GSM505527	1	60	2	12			
GSM505528	1	52	2	3			
GSM505531	1	46	2	3			
GSM505533	0	50	2	12			
GSM505534	0	66	2	12			
GSM505541	1	62	2	3			
GSM505543	1	68	2	3			
GSM505545	1	32	2	3			
GSM505548	1	51	2	12			
GSM505552	1	40	2	3			
GSM505553	1	50	2	3			
GSM505554	1	52	2	3			
GSM505561	0	48	1	12			
GSM505562	1	61	2	3			
GSM505563	1	51	2	3			
GSM505564	1	49	2	3			
GSM505566	1	57	2	3			
GSM505568	1	50	2	3			
GSM505570	1	43	2	12			
GSM505572	1	58	2	3			
GSM505574	0	72	2	3			
GSM505575	0	33	2	3			
GSM505580	1	46	2	12			
GSM505583	1	53	2	3			
GSM505584	1	51	2	3			
GSM505586	1	64	2	3			
GSM505588	0	36	2	3			
GSM505592	1	50	2	3			
GSM85474				3			
GSM85476				3			
GSM85477				3			

Table A.1. Characteristics of TNBC dataset continued

GSM_accession	Nodal status	Age	T1vs2	G12vs3	Month	Event	Chemotherapy
GSM85478				3			
GSM85479				3			
GSM85480				3			
GSM85481				3			
GSM85482				3			
GSM85483				3			
GSM85484				3			
GSM85485				3			
GSM85486				3			
GSM85487				3			
GSM85489				3			
GSM85490				3			
GSM85492				3			
GSM782588	0	59	1	12	56	0	NO
GSM782589	1	62	2	3	66	1	NO
GSM305166	0				101	0	NO
GSM320216							
GSM320228							
GSM320230							
GSM320234							
GSM320237							
GSM152339		56	1	12			
GSM152343		85	2	3			
GSM152344		29	2	12			
GSM152349		65	2	3			
GSM152352		65	2	3			
GSM152357		41	2	3			
GSM152361		51	1	3			
GSM346882				3			
GSM346883				3			
GSM346884				3			
GSM346885				3			
GSM346886				3			
GSM346889				3			
GSM346890				3			
GSM346892				3			
GSM346900				3			
GSM346904				3			
GSM272166							
GSM272167							

Table A.1. Characteristics of TNBC dataset continued

GSM_accession	Nodal status	Age	T1vs2	G12vs3	Month	Event	Chemotherapy
GSM272221							
GSM272242							
GSM272287							
GSM177925	0	52	2	3	120	0	NO
GSM177935	0	37	2	3	14	1	NO
GSM177937	0	53	2	3	103	1	NO
GSM177943	0	39	1	12	108	0	NO
GSM177944	0	52	2	12	18	1	NO
GSM177954	0	37	2	12	26	1	NO
GSM177956	0	40	2	3	17	1	NO
GSM177963	0	45	2	3	57	0	NO
GSM177966	0	40	2	3	51	0	NO
GSM177970	0	47	1	3	108	0	NO
GSM177885	0	57	2	3	24	1	NO
GSM177897	0	46	2	3	120	0	NO
GSM177898	0	57	2	12	93	1	NO
GSM177899	0	33	2	3	15	1	NO
GSM177903	0	57	2	12	60	1	NO
GSM177909	0	58	2	3	4	1	NO
GSM177913	0	47	2	3	120	0	NO
GSM177988	0	43	2	3	88	0	NO
GSM177999	0	43	2	12	13	1	NO
GSM178003	0	43	2	3	95	1	NO
GSM178009	0	53	2	3	9	1	NO
GSM178012	0	38	2	3	120	0	NO
GSM178021	0	33	2	3	120	0	NO
GSM178023	0	42	2	3	120	0	NO
GSM178058	0	46	2	3	120	0	NO
GSM178065	0	47	2	3	120	0	NO
GSM178075	0	48	2	3	95	0	NO
GSM178078	0	39	2	3	120	0	NO
GSM178079	0	46	2	3	13	1	NO
GSM178082	0	39	2	3	44	1	NO
GSM177978	0	47	2	3	120	0	NO
GSM177979	0	39	2	3	84	1	NO
GSM177980	0	32	2	3	120	0	NO
GSM177985	0	45	2	3	120	0	NO
GSM177993	0	38	1	12	9	1	NO
GSM177994	0	47	1	12	42	0	NO
GSM178013	0	40	1	3	120	0	NO

Table A.1. Characteristics of TNBC dataset continued

GSM_accession	Nodal status	Age	T1vs2	G12vs3	Month	Event	Chemotherapy
GSM178016	0	40	2	12	120	0	NO
GSM178034	0	32	1	12	120	0	NO
GSM178035	0	50	1	3	120	0	NO
GSM308259		61	2	3	13	1	NO
GSM308261		55	2	3	18	1	NO
GSM308271		50	1	12	29	1	NO
GSM308272		56	2	3	14	1	NO
GSM308273		55	1	12	19	1	NO
GSM308275		39	2	3	7	1	NO
GSM308280		51	2	12	16	1	NO
GSM308281		66	1	3	5	1	NO
GSM308283		55	1	3	13	1	NO
GSM308285		50	2	3	7	1	NO
GSM308290		62	2	3	6	1	NO
GSM308295		63	2	3	11	1	NO
GSM308301		42	1	12	15	1	YES
GSM308307		36	2	3	29	1	NO
GSM308309		42	2	12	20	1	NO
GSM308311		57	2	12	36	1	NO
GSM308312		44	2	12	115	1	NO
GSM308313					4	1	
GSM308314		58	2	3	8	1	NO
GSM308316		43	1	3	0	1	NO
GSM308324		57	2	3	14	1	NO
GSM308328		52	2	3	16	1	NO
GSM308329		40	2	3	1	1	
GSM308330		31	2	3	0	1	
GSM308333		30	1	3	18	1	NO
GSM308335		38	2	12	50	1	NO
GSM308336		28	2	12	8	1	NO
GSM308338					10	1	
GSM308339		49	2	3	3	1	NO
GSM308340		46	2	12	60	1	NO
GSM308341		57	2	12	17	1	NO
GSM308344		79	2	12	6	1	NO
GSM308346		69	2	3	13	1	NO
GSM308348		39	2	3	25	1	YES
GSM308349		51	2	3	4	1	NO
GSM308350		65		3	15	1	NO
GSM308352		44	2	3	0	1	

Table A.1. Characteristics of TNBC dataset continued

GSM_accession	Nodal status	Age	T1vs2	G12vs3	Month	Event	Chemotherapy
GSM308354		35	1	3	21	1	NO
GSM308356		39	2	3	18	1	YES
GSM308364		44	1	12	29	1	NO
GSM308374		51	2	3	6	1	NO
GSM308376		56	2	3	6	1	NO
GSM308377		28	2	3	19	1	YES
GSM308378		41	2	3	15	1	NO
GSM308380		56	2	3	7	1	NO
GSM308388		37	1	3	17	1	YES
GSM308391					7	1	
GSM308392					6	1	
GSM308393		40	2	3	33	1	NO
GSM308394		34	2	3	22	1	YES
GSM308404		50	2	3	8	1	NO
GSM308405					24	1	
GSM308407		40	1	3	94	1	NO
GSM308408		45	2	12	4	1	NO
GSM308438		63	2	12	35	1	NO
GSM308449		56	2	3	47	1	NO

Table A.2 Control dataset information.

	Total	Histologically normal breast epithelium from BC patient	Histologically normal breast epithelium from prophylactic mastectomy patient	Reduction mammoplasty breast epithelium (cancer free samples)
GSE20437	42	18	6	18
GSE9574	15	0	0	15
GSE6883	6	0	0	6
GSE26910	6	6	0	0
GSE21422	5	0	0	5
GSE3744	7	0	0	7
GSE2361	1	0	0	1
	82	38	6	52

Table A.3 Characteristics of ER+ dataset.

Sample	Disease state	Specimen	Genotype	Other
GSM549274	stage 4	BMN grade 3	ER (IHC): +	postchemo: residual invasive cancer
GSM750738	stage 4	BMN grade 3	ER (IHC): +	postchemo: no invasive cancer
GSM549279	stage 4	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM549294	stage 4	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM549304	stage 4	na	ER (IHC): +	postchemo: residual invasive cancer
GSM549305	stage 4	na	ER (IHC): +	postchemo: residual invasive cancer
GSM549307	stage 4	na	ER (IHC): +	postchemo: residual invasive cancer
GSM549306	stage 4	na	ER (IHC): +	postchemo: no invasive cancer
GSM549308	stage 4	na	ER (IHC): +	postchemo: no invasive cancer
GSM549271	stage 3	BMN grade 3	ER (IHC): +	postchemo: residual invasive cancer
GSM549232	stage 3	BMN grade 3	ER (IHC): +	postchemo: no invasive cancer
GSM549259	stage 3	BMN grade 2	ER (IHC): +	postchemo: no invasive cancer
GSM549269	stage 3	BMN grade 1	ER (IHC): +	postchemo: residual invasive cancer
GSM549302	stage 3	na	ER (IHC): +	postchemo: residual invasive cancer
GSM549239	stage 2	BMN grade 3	ER (IHC): +	postchemo: residual invasive cancer
GSM549244	stage 2	BMN grade 3	ER (IHC): +	postchemo: residual invasive cancer
GSM549249	stage 2	BMN grade 3	ER (IHC): +	postchemo: residual invasive cancer
GSM549260	stage 2	BMN grade 3	ER (IHC): +	postchemo: residual invasive cancer
GSM549266	stage 2	BMN grade 3	ER (IHC): +	postchemo: residual invasive cancer
GSM549293	stage 2	BMN grade 3	ER (IHC): +	postchemo: residual invasive cancer
GSM549236	stage 2	BMN grade 3	ER (IHC): +	postchemo: no invasive cancer
GSM549238	stage 2	BMN grade 3	ER (IHC): +	postchemo: no invasive cancer
GSM549251	stage 2	BMN grade 3	ER (IHC): +	postchemo: no invasive cancer
GSM549258	stage 2	BMN grade 3	ER (IHC): +	postchemo: no invasive cancer
GSM549264	stage 2	BMN grade 3	ER (IHC): +	postchemo: no invasive cancer
GSM549241	stage 2	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM549247	stage 2	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM549261	stage 2	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM549270	stage 2	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM549277	stage 2	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM549280	stage 2	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM549281	stage 2	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM549285	stage 2	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM549288	stage 2	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM549292	stage 2	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM549295	stage 2	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM549297	stage 2	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM750743	stage 2	BMN grade 2	ER (IHC): +	postchemo: residual invasive cancer
GSM549272	stage 2	na	ER (IHC): +	postchemo: residual invasive cancer

Table A.3 Characteristics of ER+ dataset. continued

Sample	Disease state	Specimen	Genotype	Other
GSM549276	stage 1	BMN grade 1	ER (IHC): +	postchemo: residual invasive cancer
GSM549275	stage 1	na	ER (IHC): +	postchemo: residual invasive cancer
GSM549284	stage 0	BMN grade 3	ER (IHC): +	postchemo: residual invasive cancer
GSM590859	stage 2	grade 1	ER +	postchemo: residual invasive cancer
GSM590844	stage 2	grade 2	ER +	postchemo: no invasive cancer
GSM590878	stage 2	grade 2	ER +	postchemo: no invasive cancer
GSM590841	stage 2	grade 2	ER +	postchemo: residual invasive cancer
GSM590843	stage 2	grade 2	ER +	postchemo: residual invasive cancer
GSM590895	stage 2	grade 2	ER +	postchemo: residual invasive cancer
GSM590897	stage 2	grade 2	ER +	postchemo: residual invasive cancer
GSM590882	stage 2	grade 3	ER +	postchemo: no invasive cancer
GSM590849	stage 2	grade 3	ER +	postchemo: residual invasive cancer
GSM590892	stage 2	grade 3	ER +	postchemo: residual invasive cancer
GSM590900	stage 2	grade 3	ER +	postchemo: residual invasive cancer
GSM590847	stage 3	grade 2	ER +	postchemo: residual invasive cancer
GSM590857	stage 3	grade 2	ER +	postchemo: residual invasive cancer
GSM590865	stage 3	grade 2	ER +	postchemo: residual invasive cancer
GSM590872	stage 3	grade 2	ER +	postchemo: residual invasive cancer
GSM590883	stage 3	grade 2	ER +	postchemo: residual invasive cancer
GSM590887	stage 3	grade 2	ER +	postchemo: residual invasive cancer
GSM590888	stage 3	grade 2	ER +	postchemo: residual invasive cancer
GSM590891	stage 3	grade 2	ER +	postchemo: residual invasive cancer
GSM590899	stage 3	grade 2	ER +	postchemo: residual invasive cancer
GSM590845	stage 3	grade 3	ER +	postchemo: no invasive cancer
GSM590846	stage 3	grade 3	ER +	postchemo: no invasive cancer
GSM590875	stage 3	grade 3	ER +	postchemo: no invasive cancer
GSM590881	stage 3	grade 3	ER +	postchemo: no invasive cancer
GSM590854	stage 3	grade 3	ER +	postchemo: residual invasive cancer
GSM590856	stage 3	grade 3	ER +	postchemo: residual invasive cancer
GSM590861	stage 3	grade 3	ER +	postchemo: residual invasive cancer
GSM590863	stage 3	grade 3	ER +	postchemo: residual invasive cancer
GSM590866	stage 3	grade 3	ER +	postchemo: residual invasive cancer
GSM590876	stage 3	grade 3	ER +	postchemo: residual invasive cancer
GSM590893	stage 3	grade 3	ER +	postchemo: residual invasive cancer
GSM590868	stage 3	NA	ER +	postchemo: residual invasive cancer

APPENDIX B

APADetect Results

Table B.1. 3'-UTR shortening genes in TNBC

Gene Name	Affymetrix Probeset ID	Poly(A) Site ID	Proximal Probes#	Distal Probes#	SLR (TNBC compared to controls)
BGN	201262_s_at	Hs.821.1.15	8	3	2.62
YME1L1	216304_x_at	Hs.499145.1.10	7	4	2.38
MMP7	204259_at	Hs.2256.1.7	8	3	2.24
MMP7	204259_at	Hs.2256.1.3	9	2	2.2
HNRNPA1	214280_x_at	Hs.546261.1.20	5	6	2.18
HNRNPA1	214280_x_at	Hs.546261.1.23	5	6	2.18
HNRNPA1	214280_x_at	Hs.546261.1.24	5	6	2.18
IQCK	215131_at	Hs.460217.1.5	4	7	2.18
SLC16A3	217691_x_at	Hs.500761.1.20	6	5	2.14
AURKA	204092_s_at	Hs.250822.1.2	5	5	2.09
USP9X	201100_s_at	Hs.77578.1.84	8	3	2.03
RPL13	212933_x_at	Hs.410817.1.11	5	5	1.96
WARS	200628_s_at	Hs.497599.1.4	8	2	1.91
METTL9	217868_s_at	Hs.279583.1.15	8	3	1.91
IGL@	215379_x_at	Hs.449585.1.66	4	7	1.9
MED14	202610_s_at	Hs.407604.1.7	8	3	1.9
METTL9	217868_s_at	Hs.279583.1.14	6	4	1.87
RPL19	200029_at	Hs.381061.1.6	8	3	1.83
VASP	202205_at	Hs.515469.1.28	8	3	1.82
PNRC1	209034_at	Hs.75969.1.7	4	7	1.78
GLRX3	214205_x_at	Hs.42644.1.14	3	7	1.77
GLRX3	214205_x_at	Hs.42644.1.16	3	7	1.77
GLRX3	214205_x_at	Hs.42644.1.17	3	7	1.77
PAOX	50400_at	Hs.532469.1.20	8	6	1.77
YARS2	218470_at	Hs.505231.1.6	2	7	1.76

Table B.1. 3'-UTR shortening genes in TNBC continued

Gene Name	Affymetrix Probeset ID	Poly(A) Site ID	Proximal Probes#	Distal Probes#	SLR (TNBC compared to controls)
TCF3	213730_x_at	Hs.371282.1.10	5	6	1.76
TOP2A	201291_s_at	Hs.156346.1.29	7	3	1.71
SNX3	213545_x_at	Hs.12102.1.3	4	6	1.71
CNBP	206158_s_at	Hs.518249.1.4	8	3	1.7
TBX2	40560_at	Hs.531085.1.21	12	4	1.68
LMAN2	200805_at	Hs.75864.1.4	6	5	1.67
SNX3	200067_x_at	Hs.12102.1.3	7	4	1.67
VEGFA	210512_s_at	Hs.73793.1.28	4	5	1.66
YARS2	218470_at	Hs.505231.1.4	6	5	1.64
HMOX2	218121_at	Hs.284279.1.14	2	8	1.57
ADSL	210250_x_at	Hs.75527.1.19	9	2	1.57
QPCT	205174_s_at	Hs.79033.1.18	7	2	1.55
WBP5	217975_at	Hs.533287.1.3	8	3	1.55
WDTC1	40829_at	Hs.469154.1.38	11	5	1.54
NPAT	209798_at	Hs.171061.1.3	7	3	1.52
WIZ	52005_at	Hs.442138.1.3	12	4	1.52
SLC16A1	209900_s_at	Hs.75231.1.3	6	5	1.51
C1QB	202953_at	Hs.8986.1.7	9	2	1.5
STX6	214441_at	Hs.518417.1.1	8	3	1.5
STX6	214441_at	Hs.518417.1.2	8	3	1.5
STX6	214441_at	Hs.518417.1.3	8	3	1.5
STX6	214441_at	Hs.518417.1.4	8	3	1.5
STX6	214441_at	Hs.518417.1.5	8	3	1.5
STX6	214441_at	Hs.518417.1.6	8	3	1.5
STX6	214441_at	Hs.518417.1.7	8	3	1.5
STX6	214441_at	Hs.518417.1.8	8	3	1.5
MGAT2	203102_s_at	Hs.93338.1.10	9	2	1.5

Table B.2. 3'-UTR lengthening genes in TNBC

Gene Name	Affymetrix Probeset ID	Poly(A) Site ID	Proximal Probes#	Distal Probes#	SLR (TNBC compared to controls)
ACADSB	205355_at	Hs.81934.1.20	6	3	0.44
BANF1	210125_s_at	Hs.433759.1.3	8	2	0.44
TCF3	210776_x_at	Hs.371282.1.11	4	7	0.44
MYO1B	212364_at	Hs.439620.1.63	8	3	0.44
AK2	212175_s_at	Hs.470907.1.4	7	3	0.43
TNPO1	209225_x_at	Hs.482497.1.46	6	5	0.43
KIAA0090	212395_s_at	Hs.439200.1.4	9	2	0.42
SERPINE2	212190_at	Hs.38449.1.4	8	2	0.42
GHITM	209249_s_at	Hs.352656.1.19	8	2	0.41
NOX4	219773_at	Hs.371036.1.3	6	4	0.4
UBE2N	201524_x_at	Hs.524630.1.8	2	8	0.4
KEAP1	202417_at	Hs.465870.1.6	3	8	0.4
COPA	208684_at	Hs.162121.1.9	6	3	0.39
SLC7A1	212290_at	Hs.14846.1.4	2	9	0.39
EVL	217838_s_at	Hs.125867.1.33	7	2	0.39
LPP	202822_at	Hs.444362.1.35	6	4	0.39
DIABLO	219350_s_at	Hs.169611.1.6	8	3	0.38
TTC9	213174_at	Hs.79170.1.10	4	7	0.38
NFIC	206929_s_at	Hs.170131.1.21	7	3	0.38
ILF3	217805_at	Hs.465885.1.62	8	3	0.38
TBC1D16	222116_s_at	Hs.369819.1.4	6	3	0.37
INSIG1	201626_at	Hs.520819.1.23	2	9	0.37
SLC46A3	214719_at	Hs.117167.1.1	9	2	0.36
ANK3	209442_x_at	Hs.499725.1.10	9	2	0.35
NDUFB3	203371_s_at	Hs.109760.1.2	8	3	0.34
COPG	217749_at	Hs.518250.1.47	8	3	0.32
MAP3K7IP2	212184_s_at	Hs.269775.1.27	4	7	0.32
MAP3K7IP2	212184_s_at	Hs.269775.1.28	4	7	0.32
ERG	211626_x_at	Hs.473819.1.5	3	8	0.31
ERG	211626_x_at	Hs.473819.1.6	2	9	0.31
FBXL4	209943_at	Hs.536850.1.13	2	8	0.31
QRICH1	209174_s_at	Hs.297389.1.9	3	7	0.3
FST	204948_s_at	Hs.9914.1.10	6	5	0.3
OGT	207564_x_at	Hs.405410.1.62	4	6	0.3
PLEKHG3	212823_s_at	Hs.509637.1.31	9	2	0.29
ATP8A1	213106_at	Hs.435052.1.2	9	2	0.29
SPARC	200665_s_at	Hs.111779.1.7	6	5	0.28
RRP12	216913_s_at	Hs.434251.1.14	4	7	0.27

Table B.2. 3'-UTR lengthening genes in TNBC continued

Gene Name	Affymetrix Probeset ID	Poly(A) Site ID	Proximal Probes#	Distal Probes #	SLR (TNBC compared to controls)
NDN	209550_at	Hs.50130.1.3	6	5	0.27
RP6-213H19.1	218499_at	Hs.444247.1.29	7	4	0.27
RCN1	201063_at	Hs.97887.1.13	2	9	0.26
FUCA1	202838_at	Hs.370858.1.4	3	8	0.25
UBE2D2	201345_s_at	Hs.108332.1.7	5	6	0.23
CACYBP	211761_s_at	Hs.508524.1.11	2	9	0.21
ENTPD4	204076_at	Hs.444389.1.5	7	2	0.21
C1orf115	218546_at	Hs.519839.1.9	8	2	0.18
LRRC49	219338_s_at	Hs.12692.1.34	3	8	0.18
NDN	209550_at	Hs.50130.1.4	4	7	0.15
GPRC5B	203632_s_at	Hs.148685.1.3	2	8	0.11
RCN2	201486_at	Hs.79088.1.17	8	3	0.1
C16orf61	218447_at	Hs.388255.1.7	2	9	0.07

Table B.3. Constant SLR genes in TNBC

Gene Name	Affymetrix Probeset ID	Poly(A) Site ID	Proximal Probes#	Distal Probes#	SLR (TNBC compared to controls)
CRYZ	202950_at	Hs.83114.1.8	2	9	1
SORT1	212797_at	Hs.485195.1.5	3	8	1
SLC16A1	202236_s_at	Hs.75231.1.5	4	6	1
EGLN1	221497_x_at	Hs.444450.1.11	4	7	1
OAT	201599_at	Hs.523332.1.4	4	7	1
TMEM9B	218065_s_at	Hs.501853.1.3	6	4	1
PSMA1	201676_x_at	Hs.102798.1.3	8	3	1
CAT	211922_s_at	Hs.502302.1.22	8	3	1
CCND1	208711_s_at	Hs.523852.1.12	3	8	1
ETS1	214447_at	Hs.369438.1.14	6	5	1
DYRK4	212954_at	Hs.439530.1.13	3	8	1
TWF1	201745_at	Hs.189075.1.2	7	3	1
GPR182	214506_at	Hs.483909.1.6	5	6	1
PIP4K2C	218942_at	Hs.144502.1.22	7	3	1
RB1	211540_s_at	Hs.408528.1.2	2	9	1
ATP5S	213995_at	Hs.438489.1.9	7	3	1
C14orf162	220887_at	Hs.458319.1.3	5	6	1
SLC30A4	207362_at	Hs.162989.1.4	5	5	1
FUS	215744_at	Hs.513522.1.31	7	4	1
POLR2C	216282_x_at	Hs.79402.1.16	2	8	1
CHST5	64900_at	Hs.156784.1.12	11	4	1
GAS7	210872_x_at	Hs.462214.1.7	8	3	1
GAS7	210872_x_at	Hs.462214.1.8	8	3	1
GGNBP2	218079_s_at	Hs.514116.1.33	2	8	1
CDC42EP4	218063_s_at	Hs.3903.1.8	6	4	1
GRB2	215075_s_at	Hs.444356.1.6	8	2	1
RAB40B	204547_at	Hs.484068.1.2	8	3	1
PTBP1	211270_x_at	Hs.172550.1.33	8	3	1
EBI3	219424_at	Hs.501452.1.12	8	3	1
GCDH	208369_s_at	Hs.532699.1.33	6	4	1
C19orf53	217926_at	Hs.231616.1.1	5	5	1
TRIB2	202479_s_at	Hs.467751.1.23	9	2	1
TEX261	212084_at	Hs.516087.1.4	9	2	1
CYTIP	209606_at	Hs.270.1.3	7	4	1
HNRNPA3	211932_at	Hs.516539.1.16	5	5	1
SLC5A1	206628_at	Hs.1964.1.12	7	4	1
SLC25A17	214210_at	Hs.474938.1.2	3	8	1
EP300	202221_s_at	Hs.517517.1.52	5	6	1

Table B.3. Constant SLR genes in TNBC continued

Gene Name	Affymetrix Probeset ID	Poly(A) Site ID	Proximal Probes#	Distal Probes#	SLR (TNBC compared to controls)
MOBP	207659_s_at	Hs.121333.1.11	6	5	1
NKTR	202379_s_at	Hs.529509.1.24	4	6	1
UBXN7	212840_at	Hs.518524.1.3	4	6	1
WFS1	202908_at	Hs.518602.1.21	3	7	1
CXCL5	214974_x_at	Hs.89714.1.4	6	5	1
CDH6	205533_s_at	Hs.171054.1.31	9	2	1
FBXO38	221257_x_at	Hs.483772.1.50	5	5	1
TRMT11	218877_s_at	Hs.404186.1.23	5	6	1
THBS2	203083_at	Hs.371147.1.4	5	6	1
PPIA	211765_x_at	Hs.356331.1.7	9	2	1
SEMA3C	203788_s_at	Hs.269109.1.7	2	9	1
UNKNOWN	203629_s_at	Hs.239631.1.2	7	3	1
CALD1	212077_at	Hs.490203.1.47	6	5	1
MGAM	206522_at	Hs.122785.1.16	6	5	1
SH2D4A	219749_at	Hs.303208.1.25	5	4	1
NRG1	208230_s_at	Hs.453951.1.21	7	4	1
RAB14	200928_s_at	Hs.371563.1.2	7	2	1
GPR143	206696_at	Hs.74124.1.7	2	7	1
PCTK1	208824_x_at	Hs.496068.1.31	5	5	1
WAS	38964_r_at	Hs.2157.1.19	13	3	1
SMC1A	201589_at	Hs.211602.1.5	3	7	1

APPENDIX C

SAM Results

Table C.1. SAM Results. Significantly altered 3'-UTR genes

Expected score: the random fluctuation when there is no difference between the two samples

Observed score: the relative difference of a protein between the control and TNBC samples.

Numerator: the difference between the means of the two groups. A larger absolute value of the numerator says that difference between the means of the two groups is greater

Denominator : the denominator of the T-statistic (approximately the standard deviation of the expression levels for each protein across all samples).

Fold change: the ratio of the averages of the two groups

Gene Name	Expected Score (dExp)	Observed Score(d)	Numerator (r)	Denominator (s+s0)	Fold Change
IQCK	-0.254341	-16,111,7	-12,019,764	0.07460259	0.5368711
USP9X	13,262,457	-14,197,8	-0.61748123	0.0434912	0.6617657
RPL13	-0.2057283	-1,277,725	-0.7393458	0.057864234	0.6026497
SNX3	0.60942376	-11,825,29	-0.63826764	0.053974785	0.6829249
FAM3C	0.7829962	-10,865,78	-0.26181254	0.024095137	0.8368929
SNX3	0.6084021	-10,666,62	-0.5349355	0.050150383	0.7134191
TRAK2	0.0868179	-10,445,30	-0.2505425	0.023986144	0.8426916
TOP2A	-0.1750641	-10,039,98	-0.49413595	0.049216837	0.6942739
PAOX	-0.7282895	-9,941,904	-0.4348116	0.043735243	0.7363496
MED14	13,176,409	-9,782,698	-0.5702762	0.05829437	0.6720389

Table C.1. SAM Results continued

Gene Name	Expected Score (dExp)	Observed Score(d)	Numerator (r)	Denominator (s+s0)	Fold Change
WDTC1	-16,168,592	-9,268,944	-0.4928903	0.053176533	0.7264393
YME1L1	-0.8426271	-9,013,944	-0.8436321	0.0935919	0.5417494
STX6	-10,037,969	-8,809,629	-0.38182044	0.04334126	0.7641353
STX6	-10,013,411	-8,809,629	-0.38182044	0.04334126	0.7641353
STX6	-0.9989751	-8,809,629	-0.38182044	0.04334126	0.7641353
STX6	-0.9925757	-8,809,629	-0.38182044	0.04334126	0.7641353
STX6	-10,082,314	-8,809,629	-0.38182044	0.04334126	0.7641353
STX6	-10,057,963	-8,809,629	-0.38182044	0.04334126	0.7641353
STX6	-0.9965391	-8,809,629	-0.38182044	0.04334126	0.7641353
STX6	-0.9944790	-8,809,629	-0.38182044	0.04334126	0.7641353
MGAT2	-0.4104367	-8,802,423	-0.45177406	0.05132383	0.7450818
ADSL	0.20973574	-8,699,552	-0.4924091	0.056601666	0.7177865
SMARCA4	-0.0632866	-8,563,689	-0.3259505	0.038061924	0.8042794
WDR77	-11,474,493	-850,525	-0.19266793	0.022652823	0.8757393
SAG	0.11245041	-8,468,288	-0.31884277	0.037651382	0.8016577
VASP	-0.0352199	-8,363,911	-0.510409	0.06102516	0.6872744
SLC16A3	-0.1251066	-8,320,704	-0.6676481	0.08023937	0.5983475
CNBP	0.26808056	-8,297,722	-0.54125124	0.065228894	0.6820235
COIL	-0.1489165	-8,182,343	-0.3182803	0.038898434	0.8048259
CNBP	0.26704356	-81,679,12	-0.45978683	0.05629184	0.7341786
LDB1	-0.773548	-81,636	-0.41307402	0.050599493	0.7692259
TIMM9	-0.3952051	-8,097,305	-0.3758144	0.04641228	0.7788818
WARS	-0.3591779	-8,060,402	-0.49357802	0.061234914	0.6883553
HNRNPA	-0.5392660	-7,971,873	-0.5462264	0.068519205	0.6713845
HNRNPA	-0.5379276	-7,971,873	-0.5462264	0.068519205	0.6713845
HNRNPA	-0.540326	-7,971,873	-0.5462264	0.068519205	0.6713845
CLU	0.8588983	-7,778,625	-0.32378927	0.041625515	0.8010767
DNPEP	0.10524803	-7,776,565	-0.33454585	0.043019745	0.8132095
RPL19	-0.1775339	-7,704,579	-0.4492266	0.058306444	0.7104866

Table C.1. SAM Results continued

Gene Name	Expected Score (dExp)	Observed Score(d)	Numerator (r)	Denominator (s+s0)	Fold Change
METTL9	-0.2517458	-76,991,39	-0.6317172	0.08205036	0.6444195
PVR	-0.0386672	-76,506,15	-0.33535433	0.04383364	0.8002608
DCK	0.35735914	-76,339,35	-0.40556866	0.053127076	0.7500256
WIZ	-0.0568416	-75,367,77	-0.32215977	0.042745028	0.8058656
C1orf50	-1,411,343	-74,597,98	-0.23879981	0.032011565	0.8517495
TIMM9	-0.3961015	-73,965,37	-0.35316646	0.04774754	0.7881455
TBX2	-0.1405598	-73,470,07	-0.43836612	0.059665944	0.7333537
MYCL1	-14,438,697	-71,806,66	-0.26157352	0.03642747	0.8332668
CHAF1A	-0.0712966	-69,167,64	-0.26091325	0.037721865	0.8360262
NRG1	0.87017876	-6,904,762	-0.31199974	0.04518617	0.795895
MPZL2	-0.5912790	-68,929,04	-0.24499717	0.035543386	0.8448772
GLRX3	-0.7343216	-68,447,61	-0.51038164	0.07456529	0.6870627
GLRX3	-0.7373921	-68,447,61	-0.51038164	0.07456529	0.6870627
GLRX3	-0.7359134	-68,447,61	-0.51038164	0.07456529	0.6870627
PCMT1	0.64650166	-6,746,752	-0.26386	0.03910919	0.8455017
SLN	-0.6038076	-67,091,62	-0.251678	0.037512578	0.8405620
DDX28	-0.2241590	-6,679,067	-0.21907867	0.03280079	0.8659199
POFUT1	0.13437016	-6,674,212	-0.22419333	0.033590984	0.8620382
MYST2	-0.1575581	-66,528	-0.3723063	0.055962343	0.7890135
STAM	-0.852548	-6,645,086	-0.28471488	0.04284593	0.8249137
BGN	19,204,733	-6,594,874	-0.5634702	0.08544063	0.6135187
CNBP	0.26905373	-6,582,866	-0.29481313	0.04478492	0.8063773
C1QB	-17,467,387	-6,561,255	-0.3837508	0.058487408	0.7653926
LXN	0.29292747	-6,525,892	-0.2449694	0.037538074	0.8557231
YKT6	0.70146495	-6,523,273	-0.16953544	0.025989322	0.8933415
MAPK13	0.562155	-6,348,319	-0.27442306	0.043227673	0.82785
AURKA	0.15314405	-6,311,019	-0.32688197	0.051795434	0.7917885
PNRC1	0.5992627	-62,932,30	-0.74635315	0.11859619	0.6322606
G3BP2	0.36444932	-62,787,69	-0.2651658	0.04223213	0.8335069

Table C.1. SAM Results continued

Gene Name	Expected Score (dExp)	Observed Score(d)	Numerator (r)	Denominator (s+s0)	Fold Change
PRSS21	-0.2685292	-6,226,554	-0.28096974	0.045124438	0.8410842
LRCH3	0.313558	-62,254,18	-0.29426083	0.04726764	0.8057463
LMAN2	0.51339805	-61,995,58	-0.33772698	0.054475974	0.7748762
YARS2	-0.5552688	-6,184,235	-0.42176354	0.06819979	0.7114825
CUGBP2	-0.8716821	-61,332,83	-0.3197884	0.052139834	0.8185179
WBP5	14,881,585	-6,118,063	-0.35941458	0.058746465	0.7607961
YARS2	-0.5561949	-61,150,79	-0.4282841	0.07003737	0.7280733
APP	0.16287881	-60,980,16	-0.31520772	0.051690206	0.8116812
FMR1	1,840,941	-60,867,90	-0.23166332	0.03806001	0.8504520
ADH5	0.3812159	-60,116,88	-0.46114326	0.07670777	0.7430675
NNT	0.43693483	-600,528	-0.34096974	0.056778327	0.7913724
ACTL6A	0.30171627	-59,800,23	-0.26471192	0.044266034	0.8550475
CEP76	-0.1147291	-59,735,46	-0.31634387	0.052957464	0.7917626
C14orf143	-0.3680390	-5,955,388	-0.2620963	0.044009943	0.8301186
CARTPT	0.446916	-5,877,257	-0.22913474	0.03898668	0.8566038
GPR183	-0.4432443	-58,498,27	-0.26824215	0.045854714	0.8355819
TACC1	0.8721644	-58,243,62	-0.35644305	0.06119864	0.7812182
NPAT	-0.6023504	-5,804,364	-0.35885653	0.061825294	0.7706594
GRB2	-0.1323400	-5,798,092	-0.11226365	0.019362172	0.9252777
DDX27	0.14899814	-5,768,587	-0.20983383	0.036375254	0.8704795
CBX7	0.20737271	-5,764,834	-0.32557055	0.05647527	0.792248
MTMR2	-0.6126951	-57,634,06	-0.25415388	0.044097856	0.8313457
NIPBL	0.43181333	-5,760,743	-0.24960965	0.043329418	0.8491897
PDPK1	-0.2694239	-5,722,204	-0.15112032	0.02640946	0.9083882
FAM82B	0.9144773	-56,936,52	-0.25112227	0.044105653	0.8398594
C19orf42	-0.0561780	-56,240,00	-0.23169819	0.04119811	0.8434832
FDPS	-10,687,627	-55,334,22	-0.2514432	0.04544081	0.8498482
POT1	0.7846358	-54,899,67	-0.25032094	0.04559607	0.8391146
UBAC1	12,273,506	-54,452,97	-0.11108930	0.02040096	0.9265168

Table C.1. SAM Results continued

Gene Name	Expected Score (dExp)	Observed Score(d)	Numerator (r)	Denominator (s+s0)	Fold Change
DDOST	-17,634,492	-54,425,30	-0.32427585	0.059581816	0.7998743
HMOX2	-0.2645481	-539,314	-0.3397755	0.063001424	0.774485
NR0B1	1,290,503	-5,357,871	-0.15431595	0.028801728	0.9013405
RPL31	0.034591738	-5,326,513	-0.21977735	0.041261017	0.8606526
METTL9	-0.2525664	-53,188,75	-0.5164222	0.09709237	0.6679594
CETN1	-0.1203739	-5,273,159	-0.22035006	0.041787107	0.8591937
KRT12	-0.1735704	-52,731	-0.1765013	0.03347202	0.8868810
MRP63	-0.4756131	-5,255,841	-0.1890404	0.035967678	0.8907715
NEIL3	0.4151995	-5,249,213	-0.23341033	0.044465773	0.8502091
TIMM8A	14,740,407	-52,458,98	-0.22659364	0.043194443	0.8511301
IVL	-10,969,262	-52,361,00	-0.2481777	0.047397427	0.8273365
PGLYRP4	-1,093,854	-5,234,573	-0.22923401	0.043792304	0.8536569
NR3C1	0.501726	-52,174,55	-0.28241932	0.054129705	0.8070416
TBCC	0.56988657	-51,928,69	-0.28490084	0.054863855	0.82078
TCF3	-0.0830295	-51,902,03	-0.40772966	0.07855754	0.7256905
HNRNPA	-0.541407	-51,819,62	-0.20700794	0.03994779	0.87044805

APPENDIX D

GOTERM IDs and p Values

GOTERM ID	Biological Process	Log (p value)
GO:0018193	peptidyl-amino acid modification	2.66
GO:0019538	protein metabolic process	2.01
GO:0006996	organelle organization	1.88
GO:0044267	cellular protein metabolic process	1.73
GO:0051276	chromosome organization	1.7
GO:0036211	protein modification process	1.51
GO:0034645	cellular macromolecule biosynthetic process	1.51
GO:0009059	macromolecule biosynthetic process	1.5
GO:0043284	biopolymer modification	1.34
GO:0016043	cellular component organization	1.33
GO:0045039	protein import into mitochondrial inner membrane	1.33
GOTERM ID	**Molecular Function**	**Log (p value)**
GO:0003723	RNA binding	2.27
GO:0016887	ATPase activity	2.16
GO:0004002	ATPase activity	1.69
GO:0003727	single-stranded RNA binding	1.68
GO:0004386	helicase activity	1.53
GO:0008022	protein C-terminus binding	1.52
GO:0003682	chromatin binding	1.49
GO:0009055	electron carrier activity	1.49
GO:0042802	identical protein binding	1.47
GO:0000166	nucleotide binding	1.47
GO:0070035	purine NTP-dependent helicase activity	1.33
GO:0008026	ATP-dependent helicase activity	1.33
GO:0042393	histone binding	1.39

APPENDIX E

DNA contamination, and Quantification Assessment

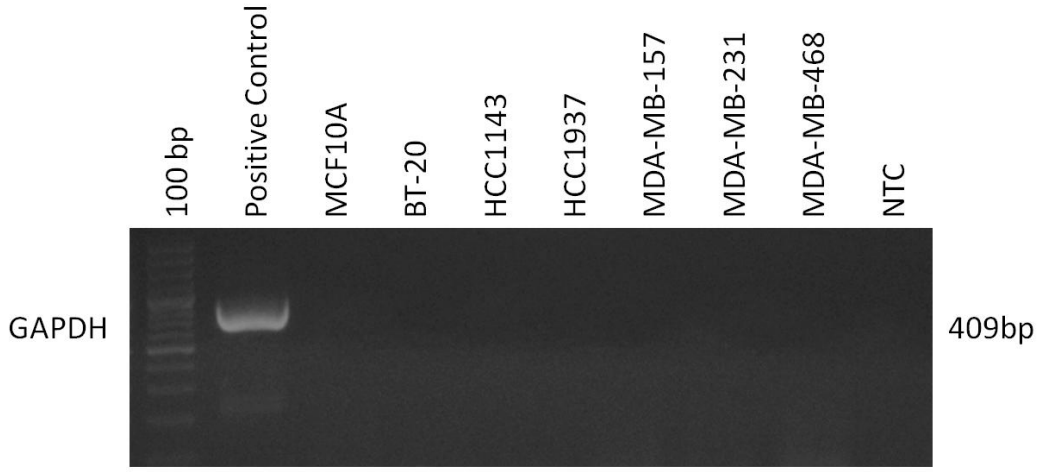

Figure E.1. Lack of DNA contamination in RNA samples was assessed. PCR was performed using *GAPDH* specific primers. GAPDH_F: 5'-GGGAGCCAAAAGGGTCATCA-3' and GAPDH_R: 5'-TTTCTAGACGGCAGGTCA GGT-3' (product size: 409 bp). Following conditions were used for the PCR reactions: incubation at 94^0C for 10 minutes, 40 cycles of 94^0C for 30 seconds, 56^0C for 30 seconds, and 72^0C for 30 seconds, and final extension at 72^0C for 5 minutes. MDA-MB-231 cDNA was used as a positive control.

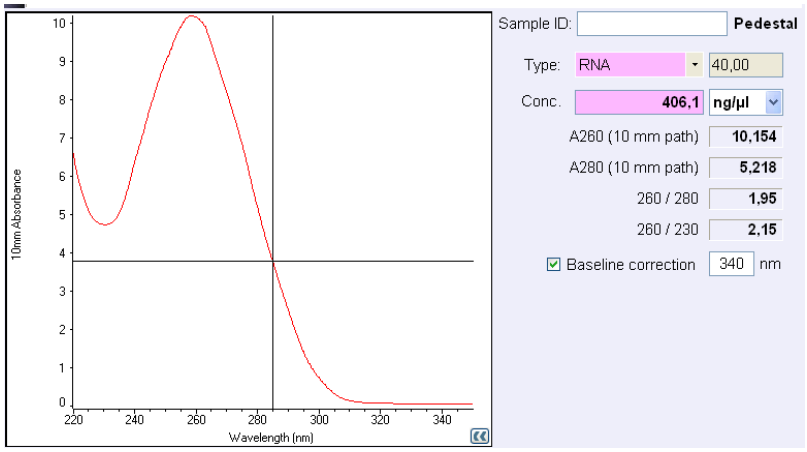

Figure E.2. RNA concentrations were determined using NanoDrop ND1000 (Thermo Scientific). Purity was determined by A260/A280 and A260/A230 ratios.

APPENDIX F

PRIMERS

Table F.1. Primers used in the study.

Primer Name	Primer Sequence (5'→3')	Product size
GAPDH_F	GGGAGCCAAAAGGGTCATCA	409bp
GAPDH_R	TTTCTAGACGGCAGGTCAGGT	409bp
SDHA_F	TGGGAACAAGAGGGCATCTG	86bp
SDHA_R	CCACCACTGCATCAATTCATG	86bp
USP9X_S_F	GAAAGAAACAGCCCCCAGAA	190bp
USP9X _S_R	TCAGTTGGTGGAGGGCTTAG	190bp
USP9X _L_F	TCCAGGTCCTCCAATCAAATAC	135bp
USP9X _L_R	GGATTGGGAAGAGGAGGATTAC	135bp
YME1L1_S_F	TGATATGGATGCTTGCTGGT	187bp
YME1L1_S_R	GGGATGCTAATTTGCAATAGG	187bp
YME1L1_L_F	AGCACTTTTCAGAGGCCAAG	177bp
YME1L1_L_R	CAGGCTCAAGCAATTCTCCT	177bp

CURRICULUM VITAE

Taner Tuncer

Middle East Technical University,
Department of Biological Sciences,
06800 Ankara/TURKEY
Tel: 0090 3122106480
Phone: 0090 5333990918
Fax: 0090 3122107976
tanertuncer1@gmail.com

Education:

PhD, **2006-2014**

Department of Biotechnology,

Middle East Technical University, Ankara, Turkey

Thesis title: "Alternative polyadenylation dependent 3'-UTR changes in Triple

Negative Breast Cancers"

MSc, **2003-2006**

Department of Biotechnology,

Middle East Technical University, Ankara, Turkey

Thesis title: "Transformation of tobacco (*Nicotiana tabaccum*) with antimicrobial *pflp* gene and analysis of transgenic plants"

BS with honors, **1999-2003**

Department of Molecular Biology and Genetics,

Middle East Technical University, Ankara, Turkey

Research Experience and Employment Record:

Teaching Assistant **2003-present**

Department of Biological Sciences,

Middle East Technical University, Ankara, Turkey

Erasmus PhD Exchange **9/2008-3/2009**

iNANO Centre for Nanotechnology,

Aaurhus University, Aarhus, Denmark

Special Project Research (Prof Dr. Hüseyin Avni Öktem's Lab) **2002-2003**

Department of Biological Sciences,

Middle East Technical University, Ankara, Turkey

Internship **6/2002-9/2002**

Selçuk University Faculty of Medicine, Konya, Turkey

Books:

1. **Taner Tuncer**, Begüm Akman. Code of Life: Human Genome Project. METU Publishing, 2007, ISBN : 9789944344296

Presentations/Conference Papers:

1. Akman BH, **Tuncer T**, Öyken M, Can T, Erson-Bensan AE. Alternative polyadenylation dependent 3'-UTR changes in triple negative breast, 3th International Congress of the Molecular Biology Association of Turkey, İzmir, Turkey, September 10-12, 2014
2. Akman BH, **Tuncer T**, Can T, Erson-Bensan AE. Alternative Polyadenylation Dependent 3' UTR Length Changes in Triple Negative Breast Cancers, EMBO Cellular Signaling and Cancer Therapy, Cavtat, Croatia, May 23-27, 2014
3. **Tuncer T,** Öktem HA, Nanoparçacık Tabanlı Genetiği Değiştirilmiş Organizma (GDO) Tanı Sistemleri **(oral presentation)**, October 29-31, 2007

4. **Tuncer T**, Öktem HA, Antimikrobiyal *Pflp* Geninin Tütüne Aktarılması ve Transgenik Bitkilerin Geliştirilmesi, XIV. National Biotechnology Congress, Eskişehir, Turkey, August 31- September 2, 2005

Awards and Scholarships:

1. **Erasmus Scholarship** **9/2008-3/2009**
   European Union Student Exchange Programme

2. **PhD Scholarship** **2007-2012**
   The Scientific and Technological Research Council of Turkey

3. **BSc Scholarship** **1999-2003**
   The Scientific and Technological Research Council of Turkey