AN INVESTIGATION OF POLY(A) SITE SELECTION LOCI IN ESTROGEN TREATED BREAST CANCER CELLS

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

BY

DAMLA BEKAR

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

OCTOBER 2019

Approval of the thesis:

AN INVESTIGATION OF POLY(A) SITE SELECTION LOCI IN ESTROGEN TREATED BREAST CANCER CELLS

submitted by **DAMLA BEKAR** in partial fulfillment of the requirements for the degree of **Master of Science in Biology Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Ayşe Gül Gözen
Head of Department, **Biology** _____

Prof. Dr. A. Elif Erson Bensan
Supervisor, **Molecular Biology and Genetics, METU** _____

Examining Committee Members:

Prof. Dr. Tolga Can
Computer Engineering, METU _____

Prof. Dr. A. Elif Erson Bensan
Molecular Biology and Genetics, METU _____

Assist. Prof. Dr. Bahar Değirmenci
Molecular Biology and Genetics, Bilkent University _____

Date: 01.10.2019

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, Surname: Damla Bekar

Signature :

ABSTRACT

AN INVESTIGATION OF POLY(A) SITE SELECTION LOCI IN ESTROGEN TREATED BREAST CANCER CELLS

Bekar, Damla
Master of Science, Biology
Supervisor: Prof. Dr. A. Elif Erson Bensan

October 2019, 80 pages

Alternative polyadenylation (APA) is an mRNA processing step implicated in 3'UTR (Untranslated Region) isoform diversity, which may have significant impacts on protein levels. Nearly 70% of known human genes harbor multiple polyA sites. Proliferative signals, developmental cues and tissue specificity can induce alternative selection of polyA sites, producing transcripts with different 3'UTR lengths.

Given that APA generates a vast isoform diversity, there are possible mechanistic explanations emerging on how APA might be regulated. To better understand the APA mechanism in the presence of proliferative signals, we chose a model system where we know APA is induced by Estradiol (E2) in Estrogen positive (ER+) breast cancer cell line. We analyzed existing Chromatin Immunoprecipitation- Sequencing (ChIP-Seq) data for certain histone marks that may overlap with polyA sites in E2 treated cells and performed experiments to confirm usage of these specific polyA sites.

Despite the low number of cases we could analyze in this study, our results may suggest a possible link between E2 regulated transcription and APA. Future high throughput experiments will be important to test how widespread these correlations are and what the underlying mechanisms are.

Keywords: APA, Transcription, E2, Poly(A) site selection, Breast Cancer

ÖZ

E2 MUAMELESİ YAPILAN MEME KANSERİ HÜCRELERİNDE POLİ (A) BÖLGESİ SEÇİLİMİNİN ARAŞTIRILMASI

Bekar, Damla
Yüksek Lisans, Biyoloji
Tez Danışmanı: Prof. Dr. A. Elif Erson Bensan

Ekim 2019, 80 sayfa

Alternatif poliadenilasyon (APA) 3'UTR (Untranslated Region) transkript çeşitliliğine sebep olduğu bilinen bir mRNA işlenmesi basamağıdır ve bu mekanizmanın protein düzeyine önemli etkileri olabilir. Bilinen insan genlerinin yaklaşık %70'i birden fazla poli A bölgesi bulundurur. Proliferatif sinyaller, gelişim işaretleri ve doku spesifitesi Poli A bölgelerinin alternatif seçilimine neden olabilir ve bu seçilim farklı 3'UTR uzunluğuna sahip transkriptler üretilmesini sağlar.

APA mekanizmasının izoform çeşitliliğine sebep olduğu bilinmektedir ve yapılan çalışmalarla APA'nın nasıl düzenlendiği ve işlediği ile ilgili muhtemel mekanistik açıklamalar ortaya çıkmaya başlamıştır. Çalışmamızda, proliferatif sinyallerin varlığında APA mekanizmasının işleyişini daha iyi anlayabilmek adına, östrojen pozitif (ER+) meme kanseri hücre hattında Estradiol (E2) tarafından indüklenen APA mekanizmasının incelendiği bir model sistemi seçilmiştir. Önceden varolan Kromatin İmmunopresipitasyon-Sekanslama (ChIP-Seq) dataları belirli histon işaretlerine bakılarak analiz edilmiş, E2 muamelesi yapılan hücrelerde histon işaretleri ve Poli A bölgeleri örtüşmelerine bakılmıştır. Bahsi geçen spesifik poli A bölgelerinin seçilimini doğrulamak için deneyler yapılmıştır.

Yaptığımız analizlerin az sayıda olmasına rağmen sonuçlarımız, E2 ile regüle edilen transkripsiyon ve APA arasında muhtemel bir bağlantıyı gösteriyor olabilir. Gelecekte yapılacak olan yüksek verimli deneyler bu korelasyonların ne kadar yaygın olduğunu test etmek ve bu bağlantıların altında yatan mekanizmaları anlamak açısından önemli olacaktır.

Anahtar Kelimeler: APA, Transkripsiyon, E2, Poly(A) bölgesi seçilimi, Meme Kanseri

To my family,

ACKNOWLEDGMENTS

First, I would like to thank and express my endless gratitude to my supervisor Prof. Dr. A. Elif Erson-Bensan for her peerless teaching, motivation and support throughout my study. I could not have presented this thesis successfully without her inspiring guidance and encouragement.

I would like to thank all my thesis committee members, Assist. Prof. Dr. Bahar Değirmenci and Prof. Dr. Tolga Can.

I would like to express my sincere gratitude to Prof. Dr. Tolga Can for his endless support in bioinformatic analyses. And, I am grateful to Prof. Dr. Mesut Muyan for his invaluable suggestions and advices.

I thank all my lab mates; Merve Öyken, Ayça Çırçır Hatıl, Gözde Köksal, Esra Çiçek, Mustafa Çiçek, Murat Erdem, İbrahim Özgül and Irmak Gürcüoğlu for answering all my questions in patience, for all the fun we had and for being my dearest friends.

I would like to thank specially to my Gözde for her precious friendship and support, she lightened my burden during my study and she was always there for me.

I am deeply grateful to Mert for his patience, support and encouragement throughout this hard process and for being beside me during all the happy and tough moments.

Lastly, I want to express my warmest gratitude to my family, especially to my mother and father whose support, love and guidance were with me all the time.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

1.1. Alternative Polyadenylation

1.1.1. 3’ End Processing of mRNA

Regulation of mRNA processing has long been known to have important roles in gene expression. Mature eukaryotic messenger RNAs (mRNAs) are generated from precursor mRNAs (pre-mRNAs) by a series of processing steps involving splicing, capping, editing, and polyadenylation which are occurring in nucleus. 3’-end processing is one of these essential processes for gene regulation. With the exception of the replication-dependent histone mRNAs [1], all eukaryotic mRNA 3’ UTRs undergo endonucleolytic cleavage of the primary transcript and subsequent poly (A) tail addition processes [2]. First, pre-mRNAs are cleaved endonucleolytically at a specific site and after that, adenosine residues are added to 3’end. These residues are termed as poly(A) tail and this process occurs by poly(A) polymerases (PAPs) in template-independent manner [3]. The activity of PAP is stimulated by Poly(A) binding protein (PABP) for catalyzing the adenosine tail addition [4].

1.1.2. Polyadenylation Mechanism

Poly(A) tail addition occurs almost synchronously with transcription. Therefore, mRNA goes through a very rapid and dynamic maturation process. Studies showed

that cleavage and poly(A) addition are closely coupled processes performed by a large and complex multi-subunit proteins [5].

These proteins involving the subunits of Cleavage and Polyadenylation Stimulatory Factor (CPSF), and Cleavage Factor Im and IIm (CFIm, CFIIm) and Cleavage Stimulatory Factor (CSTF) complexes with the addition of other factors, form the protein machinery which has an important role in recognition of poly(A) signal (PAS), binding to the PAS and pre-mRNA cleavage at a specific poly (A) site. The core protein complexes are CPSF, CSTF and CFIm which recruit other important proteins involving the scaffolding protein Symplekin, CFIIm, and poly(A) polymerase (PAP) (Figure 1.1).

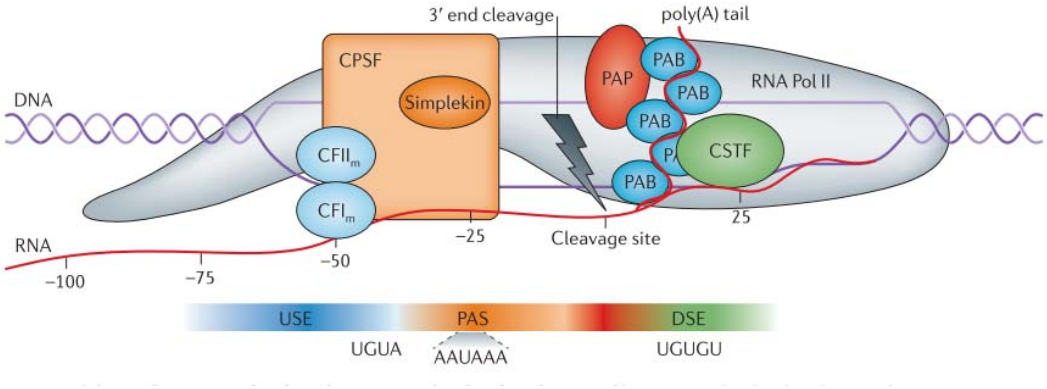


Figure 1.1. Polyadenylation Machinery

Cleavage and polyadenylation process involve several protein complexes. CPSF is important for PAS recognition. CSTF is responsible for PAS selection. CFIm which binds to U-rich/ UGUA USEs, has two polypeptides and two subunits. PAB is poly(A)-binding protein crucial for polyadenylation process. Figure is taken from [5].

Polyadenylation process is monitored by cis-acting elements that located upstream and downstream of the polyadenylation site. Cis elements are composed of hexamers

A[A/U]UAAA which are canonical poly (A) signals that highly conserved, non-canonical poly(A) signal variants, UGUA elements and U-rich elements.

The key cis-element that controls polyadenylation is 6 nucleotide long motif, termed as PAS, located ~15-30 nucleotides upstream of the poly(A) site which is recognized by CPSF and used as the cleavage site [6,7]. Additionally, there are 2 important cis-element located in the vicinity of the PAS and contribute to increase cleavage efficiency: U-/GU-rich downstream sequence elements (DSEs) and U-rich/UGUA upstream sequence elements (USEs). These sequence elements define the strength of a PAS [8]. After the recognition of PAS, CSTF complex subunit which has endonuclease activity binds to DSE and CFIm complex binds to USE to recruit other factors and mediate the cleavage process. At the end, CFIIm complex that recruited by CFIm helps the termination of transcription mediated by RNA Polymerase II and PAPs mediates untemplated adenosines addition [9].

1.1.3. Types of Polyadenylation

Latest studies done with high throughput methods have revealed that nearly 70% of known human genes' 3' UTRs contain poly(A) sites [8,10]. These alternative polyadenylation sites are used as different cleavage sites resulting in the generation of different 3'UTR transcripts with diverse lengths.

These alternative polyadenylation events can be categorized in three different groups:

1) Just one polyadenylation signal in 3'UTR resulting in an mRNA isoform and a protein (Figure 1.2 A).

2) More than one polyadenylation signal in 3'UTR resulting in different mRNA isoforms with different 3'UTRs but the same protein (Figure 1.2 B).

3) APA signals located in exons (Figure 1.2 C) or introns (Figure 1.2 D) where alternative polyadenylation occurs along with alternative splicing. Different protein products can be generated subjected to the location of the stop codon [11].

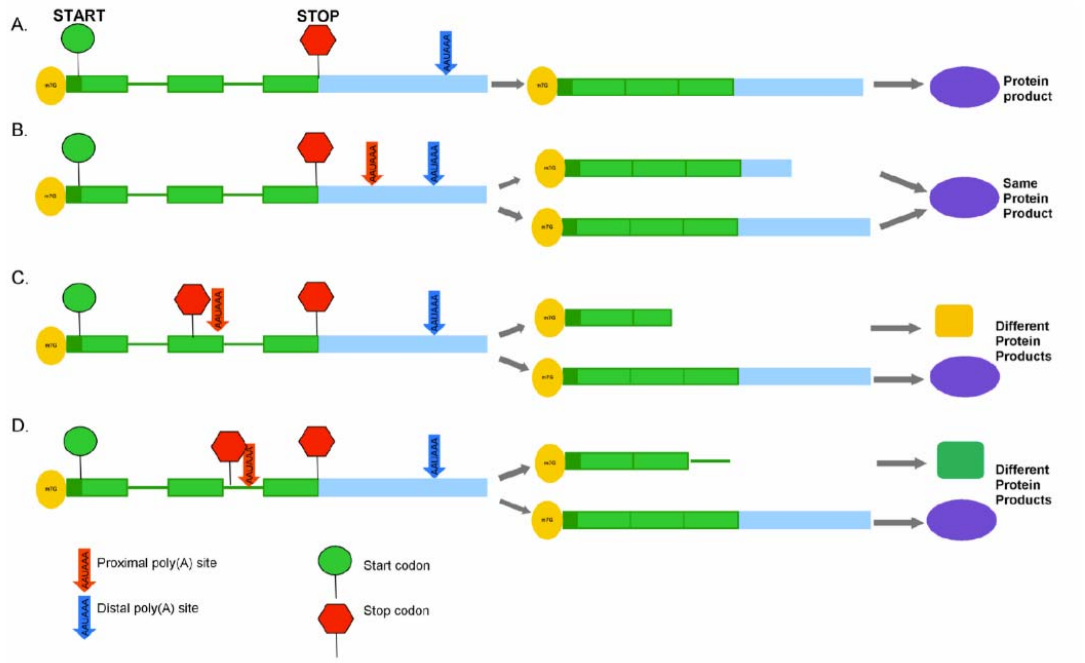


Figure 1.2. Alternative Polyadenylation Events

A) Only one polyadenylation signal in 3'UTR of mRNA. B) More than one polyadenylation signal in 3'UTR of mRNA. C) APA signals located in the exons of mRNA D) APA signals located in the introns of mRNA. Figure is taken from [11].

1.1.4. Consequences of APA

It is known that APA is a process that generates different isoforms. However, the functional importance of these isoforms and its relationship with differential usage of poly (A) sites is this unknown. For a better understanding of the effects of APA, the

alterations resulting from different poly (A) site usage can be examined in four categories: 1) coding sequence; 2) mRNA localization; 3) protein level; and 4) protein localization (Figure 1.3).

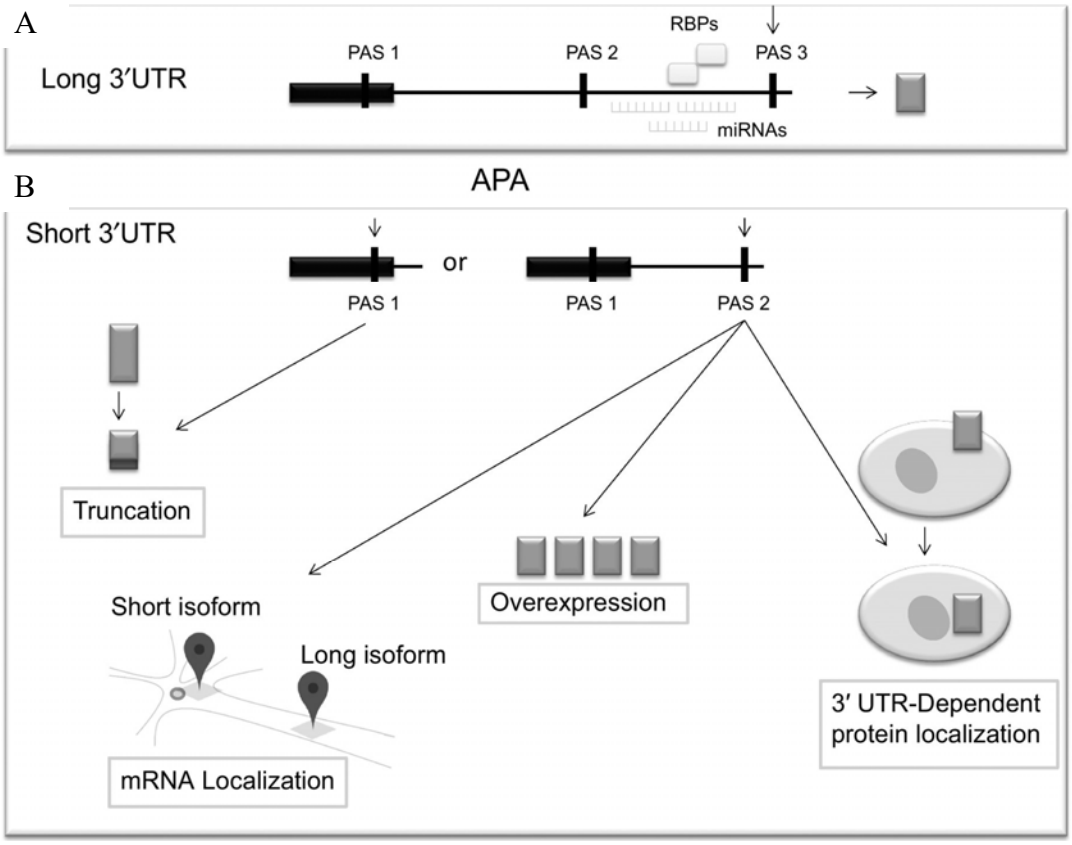


Figure 1.3. Consequences of APA

A. Transcript shows 3 PAS located differently on a hypothetical mRNA. B. Possible mRNA isoforms when PAS 1 or PAS 2 is selected. PAS 1 is located in the coding region and PAS 2 is located in 3'UTR like PAS 3. The selection of different ploy (A) sites may generate different consequences. Figure is taken from [12].

The position of PAS is the main reason why different mRNA isoforms are produced. Hence, it is important to take the position of PAS into consideration while talking about the consequences of APA. Let's think about a hypothetical mRNA having three

poly (A) sites (Figure 1.3.A): PAS 1 is located in coding sequence, PAS 2 is proximal and PAS 3 is distal and they are both located in untranslated region. According to PAS selection, APA will generate isoforms with different lengths. When PAS 1 is selected, coding region is altered and truncated mRNA is produced. As a result, a truncated protein is translated which may function different from the full-length protein. When PAS 2 or PAS 3 is selected, 3'UTR length of the mRNA changes. Coding region remains the same, but the resulting isoform may have different sub-cellular localization. Its translation efficiency may differ and this can cause different protein localization or overexpression.

With the selection of PAS 2, proximal poly (A) site, 3'UTR is shortened and the target sites for RNA binding proteins (RBPs), Adenylate/uridylate-rich elements binding proteins (AREBPs), long noncoding RNAs (lncRNA) and microRNAs (miRNAs) are not present in the transcript which affects the mRNA stability, translation, localization and exportation [13] (Figure 1.4) . When PAS 3 is selected, distal poly (A) site, longer 3'UTR isoform is generated and binding sites for factors like miRNAs or RBPs are included in the transcript.

miRNAs are non-coding RNAs which are negative regulators that inhibit translation by binding to 3'UTRs and destabilizing mRNA [14] .It is proposed that short 3'UTR isoforms escape from miRNA-mediated negative regulation [15], and this can explain the increase in the stability of short 3' UTR isoforms [16]. So, without transcriptional up-regulation, the absence of negative regulators may result in increased translation. However, miRNA-mRNA interactions are not fully understood, and the generality of shorter isoforms and stability relation is under debate. But it is known that APA generated isoforms have potentially different stabilities which affect protein levels.

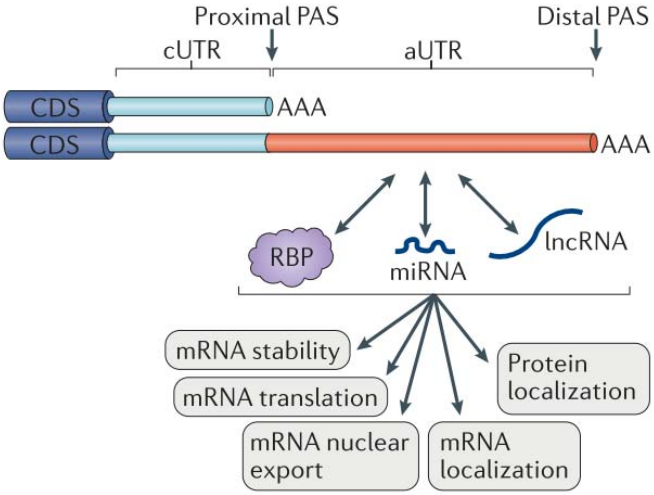


Figure 1.4. Possible Outcomes of miRNA, RBP and lncRNA Binding

Interactions of long non-coding RNAs (lncRNAs), RNA-binding proteins (RBPs) and microRNAs (mi RNAs) with UTR resulted in different functional outcomes. Adenosine residues (AAA) represent poly(A) tail. Figure is taken from [17].

Localization of mRNA is also known to be an important regulator for protein levels and their functions. mRNA transcripts have different exportation efficiencies depending on their 3'UTR lengths. Longer 3'UTR isoforms show tendency to be located in nucleus more than shorter isoforms [18] and movement of the mRNA in the cytoplasm is reported to be affected by APA . Brain-derived neurotrophic factor (BDNF) generates different 3' UTR isoforms by APA. The short 3' UTR is kept in the neuron's soma however, the long isoform is located in dendrites and because of the their localizations they have different translation rates and functions [19]. Moreover, studies showed that it is not only the transcript localization that is affected by APA; the localization of the encoded protein can also be affected by the 3' UTR of the mRNA showing that differential selection of poly (A) sites can lead to different 3' UTR dependent protein localization [20].

Even though we have limited knowledge about the consequences of APA, novel evidence suggests that APA generated isoforms have a functional role in mRNA life cycle and gene expression because of their potentially different stabilities, sub-cellular localizations, translation efficiencies and functions. Moreover, APA has been implicated in many diseases and studies show that APA will emerge as an important mechanism in many diseases including cancer for its contribution to deregulation of gene expression.

1.1.5. Regulators of Poly A Site Usage

Although the increasing interest in APA has led to discoveries that unravel the complexity of the process, the exact mechanisms of APA are not fully understood, however; there are possible mechanistic explanations emerging about how APA is regulated including differentially expressed proteins important for polyadenylation, promoter sequence, transcription rate, splicing patterns, poly(A) signal integrity and local chromatin structure. In addition, proliferative signals, hormones and differentiation factors are also possible regulators of APA (Figure 1.5).

One explanation comprises the core processing machinery proteins that affect poly (A) site selection and deregulation of these proteins leads to global changes in 3'UTRs of mRNAs. Proteins involved in polyadenylation like CSTF2, CFIm complex are known to affect polyadenylation process since their depletion or overexpression leads to differential site selection [21].

Another explanation is that polyadenylation occurs co-transcriptionally which means that initiation of polyadenylation, RNA polymerase rate, and splicing are controlled by transcriptional machinery which in turn may affect polyadenylation efficiency and specificity [22]. Several studies have revealed that independently discovered mRNA processing reactions closely influence one another. In the light of this information,

recent findings showed a linkage between APA, transcriptional initiation and RBPs. ELAV is an RNA-binding protein fostering a very long 3'UTR formation by inhibiting RNA processing at proximal polyadenylation sites [23]. In a recent study, it has been shown that native promoters with a GAGA element show tendency to pause RNA Pol II and recruit ELAV for a longer 3' UTR isoform production [24]. Hence, it is important to investigate the relationship between poly(A) site selection and promoter to a better understanding of possible inducers of APA.

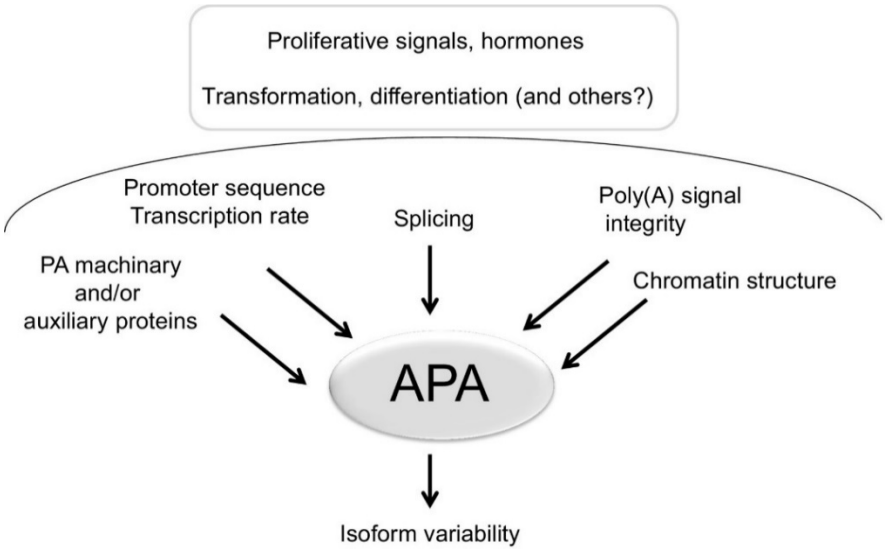


Figure 1.5. Regulators of APA.

Possible mechanistic explanations emerging about how APA is regulated including differentially expressed proteins important for polyadenylation, promoter sequence, transcription rate, splicing patterns, poly(A) signal integrity and local chromatin structure. In addition, proliferative signals, hormones and differentiation factors are also possible regulators of APA. Figure is taken from [12].

APA is also regulated by epigenetic and chromatin modifications like local chromatin structure, histone modifications, DNA methylation and positioning of nucleosomes [25]. For example, a study has shown that around PAS, there is a nucleosome depletion; on the contrary, at the downstream of PAS there is a nucleosome

enrichment. Depending on different genomic sequences of PAS, binding affinity of nucleosomes changes so, we can say that there is a link between nucleosome density and the strength of PAS. Therefore, for a better understanding of the mechanisms of APA and its relationship with cancer, the epigenetic marks in cancer cells should also be considered.

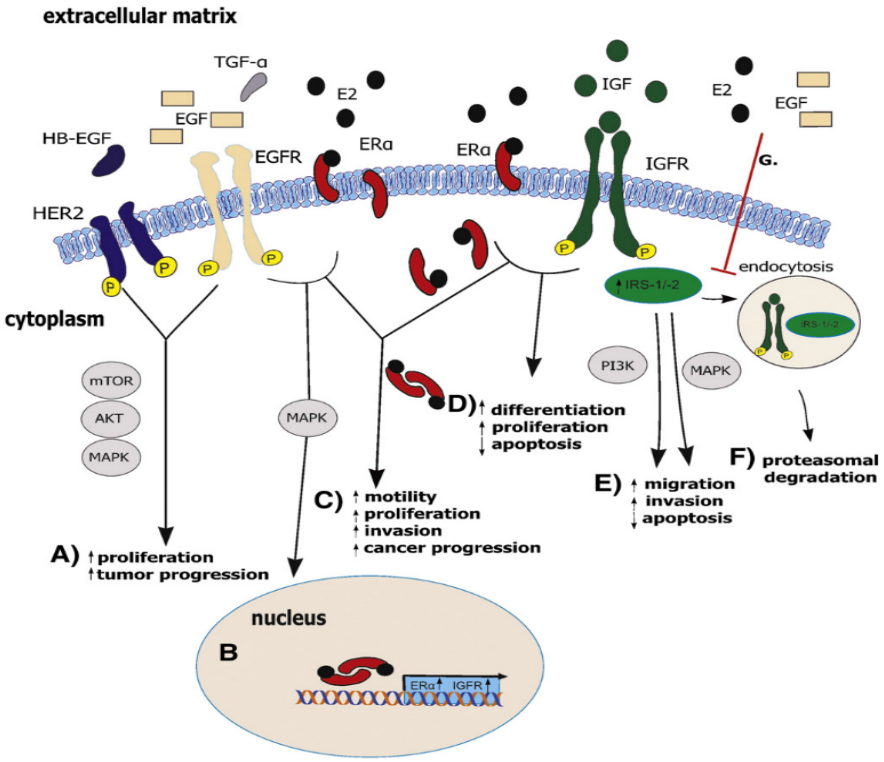


Figure 1.6. Outcomes of Proliferative Signals

Illustration of signaling pathways containing proliferative signals ERs–EGFR–IGFR crosstalk and their possible consequences. Figure is taken from [26].

Another possible mechanistic explanation about APA regulation can be the effect of proliferative signals and transcriptional response by cell-signaling pathways which have also different consequences for cell fate (Figure 1.6). In a recent study focusing on breast cancers, a link between APA and signaling pathways has been reported

where estrogen (E2) receptor positive (ER+) and Epidermal Growth Factor (EGF) receptor positive (EGFR+) breast cancer cells show an increase in proximal poly (A) site usage [25,26]. There is also a study which shows a link between APA and mTOR (mammalian target of rapamycin) pathway in addition to E2 and EGF. They observed an increase in 3' UTR lengthening when they knockdown mTOR. These findings suggest that mTOR may have a functional importance in the regulation of APA but the exact molecular mechanism of 3' UTR processing in relation with mTOR activity is still remaining elusive [30]. Although proliferative signals and hormones can alter 3'UTR length of some genes, further studies are needed to better understand the mechanisms underlying APA and 3'UTR shortening considering cancer-associated signaling pathways.

1.2. Transcription Coupled DNA Breaks

Transcription process is the first critical stage of gene expression that produces primary RNA transcript from a particular DNA segment. Transcription is known to be associated with DNA supercoiling and DNA breaks to open up the chromatin structure and increase DNA accessibility [31]. When a DNA break occurs, DNA damage response (DDR) is activated. This response involves recruitment of repair proteins to the sites of DNA damage and checkpoint response activation. Until the end of repair process, DDR arrests or slows down the cell-cycle progression, (Figure 1.7).

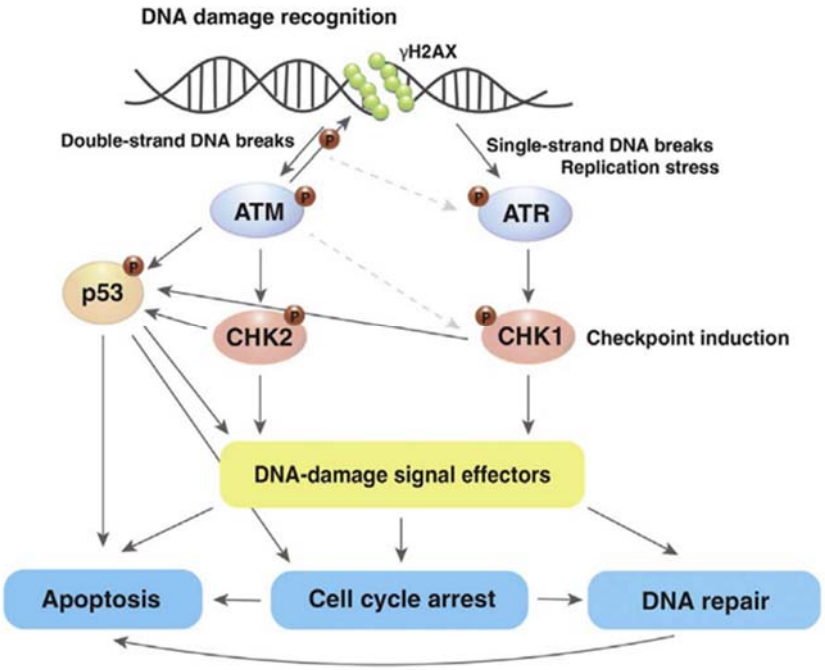


Figure 1.7. DNA Damage Response Pathway.

DNA damage and replication stress induce single and double strand breaks. This activates a response pathway. Figure is taken from [32]

PARP1 and MRE11- RAD50-NBS1 (MRN) complex recognizes the resulting DNA double-strand breaks (DSBs) and recruits the ataxia telangiectasia mutated (ATM) protein kinase which phosphorylates the H2AX histone variant and forms gamma H2AX (γH2AX). Formation of γH2AX recruits additional DDR factors [33] . γH2AX is considered as a sensitive biomarker for DSBs since it firstly appears when there is a DNA damage and (Figure 1.8).

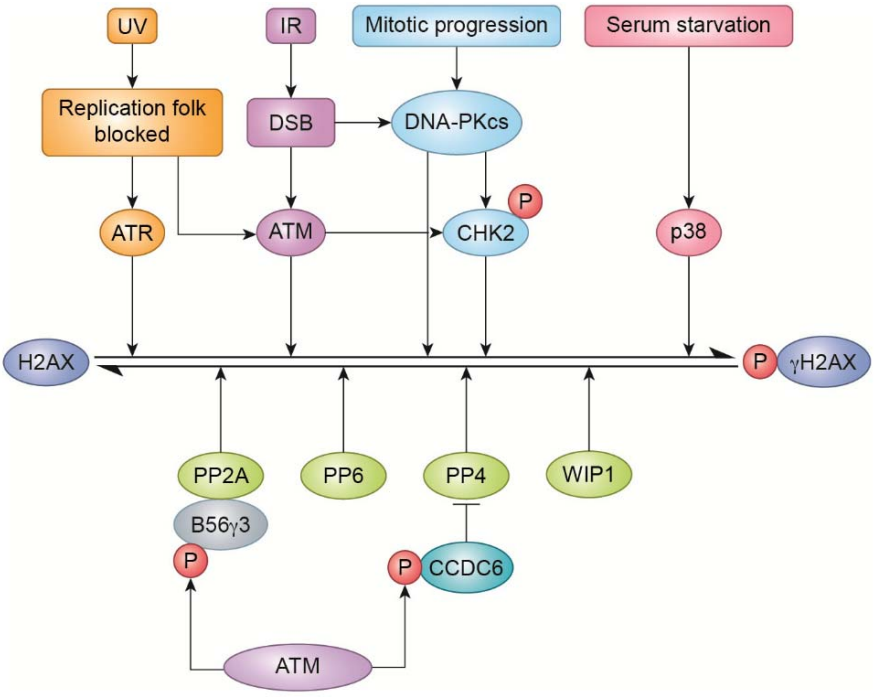


Figure 1.8. Regulation of H2AX Phosphorylation

In response to UV, ionizing radiation, mitotic progression and starvation γH2AX is formed. Figure is taken from [34].

The repair of DSBs involve non- homologous end joining (NHEJ) and homologous recombination (HR) pathways. In HR, an intact chromatid serves as a template for repair and it forms 3' single stranded DNA (ssDNA) overhangs. In NHEJ, the broken DNA ends are brought back together throughout whole cell cycle phases with the recruitment of 53BP1 mediator protein and Ku70-Ku80 heterodimer to DSB ends. This inhibits end resection and promote the joining of the broken ends [35] (Figure 1.9).

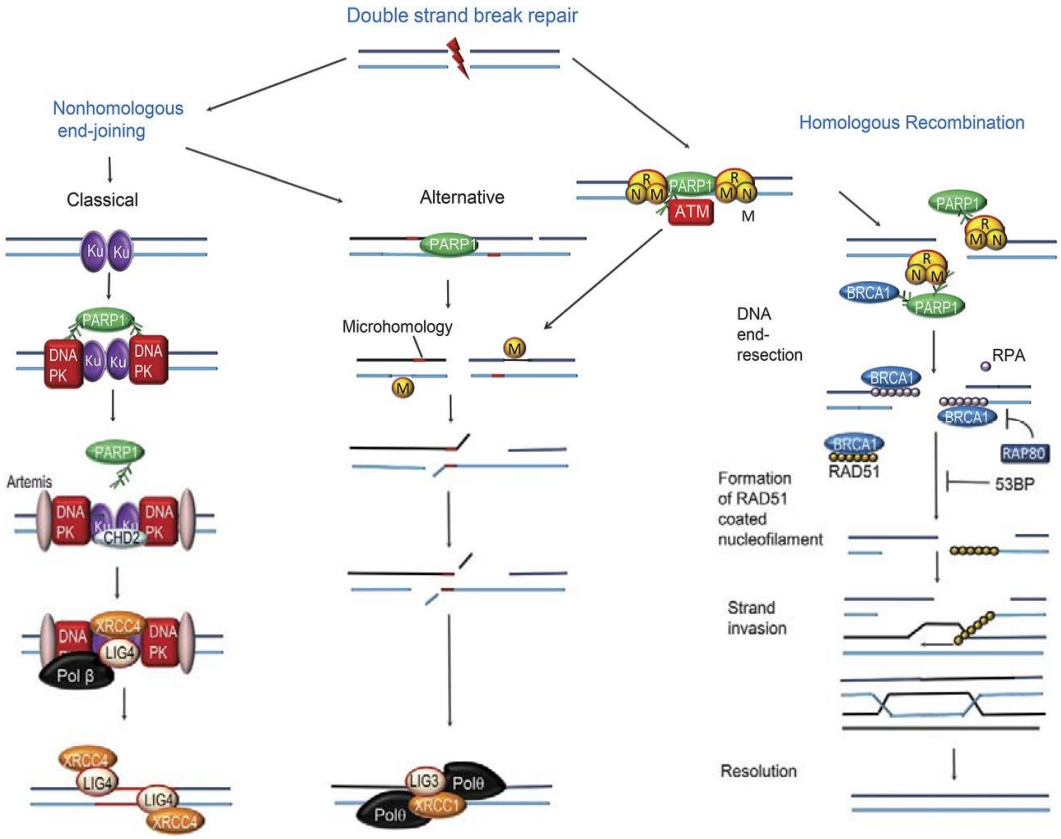


Figure 1.9. Double Strand Break Repair Mechanism.

(PARP)1 is an initial sensor of double strand DNA breaks. There are 2 major pathways for double strand DNA breaks: nonhomologous end joining and homologous recombination. Figure is taken from [36].

It has been discovered that gene transcription may create occasional DNA damage as a byproduct. Moreover, recent findings have suggested that DNA damage can also be generated by cells in a "scheduled" manner during transcription [37]. In a study focusing on estrogen-induced transcription associated with topoisomerase IIb, in not-induced state, the estrogen-dependent pS2 gene promoter contains topoisomerase IIb in a repressor complex associated with proteins such as nucleolin, PARP1 (poly (ADP ribose) polymerase 1), Hsp70, HDAC3 and N-Cor gene repressor [38]. When ER+ MCF7 breast cancer cells are treated with estrogen 17-estradiol (E2), estrogen receptor

(ER) binds to the promoter, most of the proteins in repressor complex are released and nucleosomes are deprived from gene promoter. They demonstrated estrogen-induced recruitment of proteins such as topoisomerase IIb, PARP1, DNA-PK, Ku70 to the estrogen responsive pS2 gene meaning that the E2 treatment is inducing dsDNA break in MCF7 cells.

1.3. Aim of the Study

APA has been implicated in 3'UTR isoform diversity which may have significant impact on protein levels. Despite increasing evidence of APA, it is not clear how proliferative signals such as estrogen might induce APA. Here we took an in-silico approach to study the effects of E2 treatment in ER+ breast cancer cells to identify whether E2 induced transcription patterns correlate with APA.

CHAPTER 2

MATERIALS AND METHODS

2.1. ChIP-Seq Datasets

Gene Expression Omnibus (GEO) is a public data repository that contains next-generation sequencing, microarray and other high-throughput functional genomics data [39]. GEO dataset GSE57426 [40] contains three datasets GSM1382415 gammaH2AX_ChIPSeq Vehicle treated, GSM1382416 gammaH2AX_ChIPSeq Estrogen treated, GSM1382417 gammaH2AX_ChIPSeq H2O2 treated.

GSE57426 contains experiments performed with MCF-7 breast cancer cell line. After starving cells for 72 hours in 10% dextran-coated charcoal-stripped FCS -DMEM without phenol-red, hormone depleted cells were treated with 100 nM estrogen, H2O2 or vehicle for 45 minutes. Illumina methodology was used to perform γH2AX ChIP-seq. Genome binding/occupancy profiling experiments were performed by high throughput sequencing.

2.2. Analyses for IGV Visualisation

Analyses were performed on The Cancer Genomics Cloud (CGC), in association with Seven Bridges. After creating an account in CGC, Apps were added from Public Apps Section to the created Project according to the intended file extension (Figure 2.1).

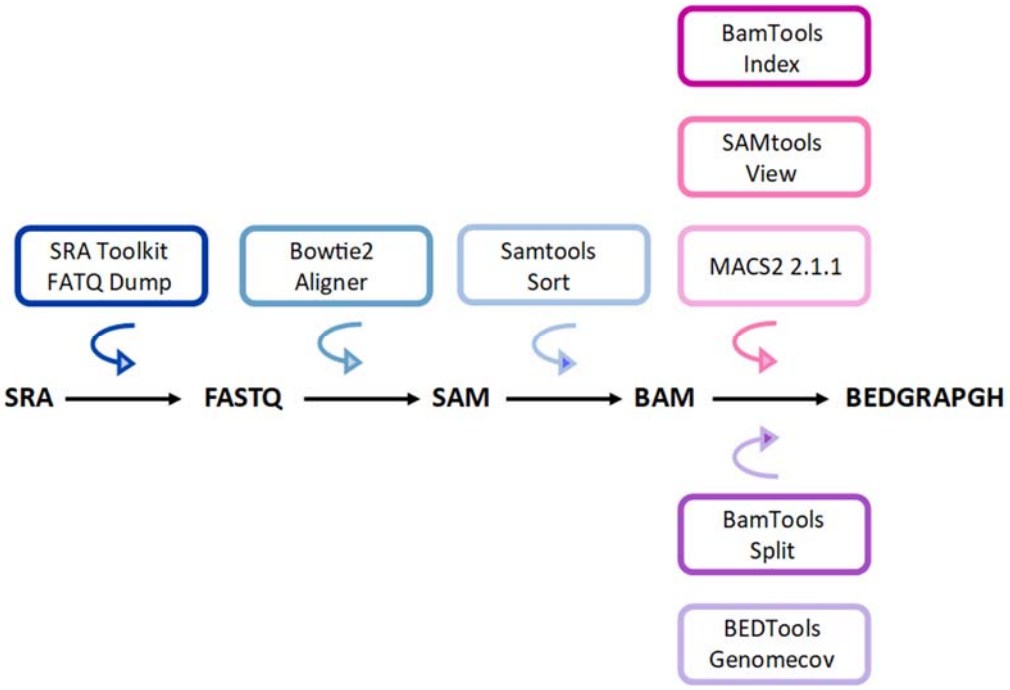


Figure 2.1. ChIP-Seq Analysis Pipeline.

Step by step SRA to BEDGRAPH conversion is illustrated. All analyses were performed on the CGC platform.

First step, after addition of SRA Toolkit fastq-dump app, was uploading sra file of data set taken from GEO by creating a ftp link (https://ftp.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR number/…) to “Add File” section and import under FTP/ http in CGC or write SRA accession number as SRA# to SRA Toolkit fastq-dump application. After importing input file, we defined “App Settings” by selecting ‘no_paired_end’ option for ‘Read ids’ and ‘Split reads’. SRA Toolkit is a collection of tools and libraries which uses data in the International Nucleotide Sequence Database Collaboration (INSDC) Sequence Read Archives format (SRA). Tool fastq-dump converts SRA data into FASTQ format. Fastq-dump tool accepts file

as input, however it can also accept accession number of the SRA. As an option, output can be selected as FASTA or FASTQ/FASTA.

Second step, after addition of “Bowtie2 Aligner App”, was giving fastq file as read sequence input and ‘human_g1k_v37_decoy.phiX174_bowtie2-2.2.6.tar’ to Bowtie index archive as reference genome input. Bowtie2 Aligner is a tool that aligns sequencing reads by taking long reference sequences as base. Output alignments are in SAM format.

Third step, after adding Samtools sort App, was giving “fastq.sam” file to “Samtools Sort” as input file. The app sorts alignments according to leftmost coordinates, or when -n is used it sorts by read name. Output is in BAM format. Now, the sra file that is uploaded at the beginning has fastq.sorted.bam format.

Fourth step, after adding BamTools Index, was giving fastq.sorted.bam as input. It creates BAM file as output and BAI file as generated index file.

Fifth step, after adding SAMtools View App, was giving “fastq.sorted.bam” as input file and “fastq.sorted.bam.bai” as index file. The tool filters SAM or BAM formatted data, uses options and arguments to understand what data to select and passes only that data through. It can be used just to look at the raw file contents, to convert a subset of data into a new file and to convert between BAM and SAM formats. From app settings, Output BAM file is marked as TRUE. The output file will be in filtered BAM format as fastq.sorted.filtered.bam.

Sixth step, after adding MACS2 2.1.1 App, was preparing data before giving it to MACS2 app. In Files section, click … sign (more actions) at the right top, select ‘export metadata manifest’ and download. Open downloaded manifest.csv file with wordpad. Write ,chip-seq at the end of the first line of the document and write ,sample at the end of the remaining lines which contain , (comma) at the end. Then, save the changed csv file and upload this file to CGC by clicking … (more actions) under Files

section. Select ‘import metadata manifest’ and import the changed csv file. Then, select all the files and click Edit. Delete ‘sample’ from control file. Do not change ‘Treatment chip-seq : sample’. Now, the csv file that contains metadata fields will be applied to the other files.

The last step was giving sorted and filtered bam files to MACS2 2.1.1 App. First, choose treatment and control file as Chip-seq treatment. Then, define app settings by selecting Analysis Type as Broad peaks and Format of Input as BAM. Select TRUE option in ‘If True, MACS will save signal per million’ section. Adjust ‘Name of organism to calculate effective genome size’ as 'hs' for human (2.7e9). Select TRUE option in ‘Save extended fragment pileup, and local lambda tracks into a bedgraph file.’ Section. Model-based Analysis of ChIP-Seq (MACS) is an algorithm and it identifies peaks by using ChIP-seq data. MACS2 evaluates the significance of enriched ChIP regions and enhances the binding sites’ resolution by using the information that contains the combination of orientation and sequencing tag position. The analysis result consists of several different file formats like narrowpeaks, broadpeaks, R and BDG format. BDG (bedgrapgh) format is used for IGV visualization. BDG file can be uploaded to IGV and peaks can be examined.

In addition to these steps, BEDGRAPH files can also be created without giving BAM files to MACS2 but in this way peak calling will not be applied to ChIP-seq data. In this case, there is no need to run SAMtools View App and BamTools Index App since they are preparing files to be imported as input to MACS. For the convertion of BAM files to BEDGRAPH files without MACS2 App, first, BamTools Split App should be added to the Project. After running Samtools sort App (Third step written above), import fastq.sorted.bam files to BamTools Split App as input. From App Settings, define ‘Reference prefix’ as ‘hg19_chr’ and change ‘Split Options’ by choosing ‘reference’ option. BamTools Split App splits a BAM file depending on the properties specified by user and creates new BAM output based on split options. Output files

will be in BAM format and named as fastq.sorted.splitted.hg19_chrX.bam according to the used reference (chrX changes according to the expected chromosome number).

Next step was to add BEDTools Genomecov App. Then, add 'hg19.chrom.sizes' file as Genome file and 'fastq.sorted.splitted.hg19_chrX.bam' as Input file which must be grouped by chromosome. We Chose TRUE option for 'Depth bedgraphformat' from App Settings. In a genome, BEDTools Genomecov App calculates the coverage of a feature file. It generates BED or BEDGRAPH file as output.

After converting file extensions of ChIP-seq data, Integrative Genome Viewer IGV 2.3 was used to generate ChIP-seq data snapshots by adjusting track height and data range accordingly.

2.3. Probe Secreening from Affymetrix

NetAffx Query was used to search probe sets under NetAffx™ Analysis Center.
Gene name was written to 'Search on probe sets for Affymetrix Gene Chip Catalog Arrays' section and Human Genome U133 Plus 2.0 Array (HGU133Plus2, GEO Accession number: GPL570) was selected as Gene Chip Array (Figure 2.2). From Results, after clicking on one of the Probe Set IDs and then 'View on UCSC Browser', probe sets based on poly (A) site locations were obtained.

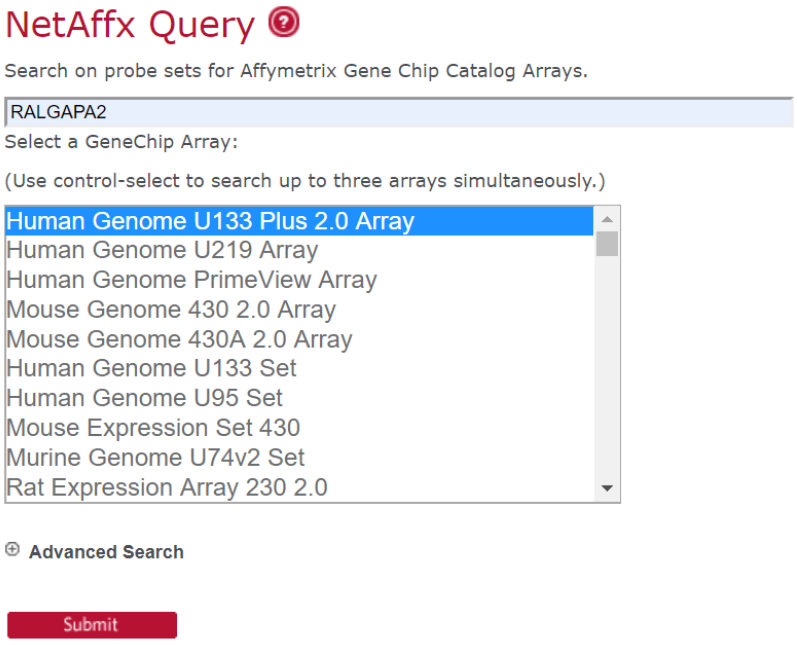


Figure 2.2. Probe Secreening from Affymetrix

2.4. Microarray Datasets Comparison in GEO2R

GEO2R is a tool that compares two or more groups of samples with GEO Series accession. It is used to identify differential expressions of genes in different experiments. Microarray datasets GSE11324 [41] and GSE8597 [42] were used in GEO2R comparison. Experiment datasets are summarized in Table A.1 and Table A.2 in Appendix A.

GEO accessions are written to GE02R web tool and all results are saved. Then, experiments are selected and grouped under ‘define groups’ part. After entering the corresponding probe ID in Profile Graph section, the gene expression profile is generated.

2.5. Statistical Analyses

Graphpad Prism was used for graphic plotting. The graph type was column bar graph plotted by taking mean with SD. Unpaired t-test was used for the comparison of treatment and control samples' mean values.

2.6. Cell Lines and Cell Culture

MCF7 cell line was cultured in Dulbecco's Modified Eagle Medium (DMEM) high glucose with glutamine (REF 01-052-1A, BI) containing 10% FBS (cat # S1810-500, biowest), 1% Penicillin/Streptomycin Solution (P/S) (REF 03-031-1B, BI), 1% Sodium Pyruvate Solution (REF 03-042-1B, BI). Incubation conditions were 37 ºC temperature with 5% CO_2 and 95% humidified air. Cell were grown as monolayers. Cryopreservation process is performed with 5% DMSO (dimethylsulfoxide) (cat # 154938, Sigma) in liquid nitrogen when cells reach 70-80% confluency at T75 flasks.

2.7. E2 Treatment

MCF7 cells were grown until confluency reaches to 40-60% in T75 flasks. Then, growth medium was cleaned and cells were cultured in starvation medium (phenol red-free medium containing 10% dextran-coated-charcoal stripped FBS) for 72 hours. After deprivation of stimulants, MCF7 cells were treated with 100 nM 17 β-Estradiol (E2) (Sigma-Aldrich) or ethanol (vehicle control) for 12 hours. After the treatment, cells are collected as E2 treated and control samples.

2.8. RNA Isolation

High Pure RNA Isolation Kit (REF 11828665001) was used to perform total RNA isolation by considering the instructions of manufacturers. Eluted RNA was incubated overnight with DNase I Recombinant Enzyme (cat # 04716728001, Roche) for DNase treatment (Reaction conditions are summarized in Table 2.1). 3M Sodium Acetate (NaAc), Phenol/Chloroform/Isoamyl Alcohol (25:24:1) and ethanol were used after incubation process. Elimination of DNA contamination was confirmed by PCR performed with GAPDH (Glyceraldehyde 3-phosphate dehydrogenase primers. PCR reaction was performed under following conditions: denaturation at 94°C- 10 minutes (min) incubation, 40 cycles at 94°C- 30 seconds (sec), 56°C (annealing temperature)- 30 sec, 72°C- 30 sec, and final extension at 72°C- 5 min. MCF7 cDNA was used as positive control and no template control was used to check contamination (Figure C.1. in Appendix C).

Table 2.1. DNase Treatment Reaction Conditions

Components	Amount
Total RNA	10 µL
10X Incubation Buffer (400 mM Tris-HCl, 100 mM NaCl, 60 mM MgCl2, 10 mM CaCl2, pH 7.9)	2 µL
DNase I recombinant, RNase-free (10 units/µl)	10 µL
RNase-free Water	up to 100 µl

After the confirmation of the absence of DNA contamination, MaestroNano Spectrophotometer (cat # MN-913, Maestrogen) was used to measure the

concentration and purity of RNA samples. A260/A280 and A260/A230 ratios of all RNA samples were fitting to purity criteria (Figure C.2. in Appendix C).

2.9. cDNA Synthesis and PCR for Confirmation

cDNAs were synthesized by RevertAid First Strand cDNA Synthesis Kit (cat # K1622, Thermo Scientific) by using 1 µg cleaned RNA (Reaction conditions are given in Table 2.2).

Table 2.2. cDNA Synthesis Conditions

Components	Amount
Total RNA	1 µg
Oligo(dT)$_{18}$ primer, 100 µM	1 µL
RNase-free Water	up to 12 µl
Incubation at 70 °C for 5 minutes	
5X Reaction Buffer (250 mM Tris-HCl (pH 8.3), 250 mM KCl, 20 mM MgCl2, 50 mM DTT)	4 µl
10 mM dNTP Mix	2 µl
RiboLock RNase Inhibitor (20 U/µL)	1 µl
RevertAid RT (200 U/µL)	1 µl
Incubation for 1 hour at 42°C and for 5 minutes at 70°C	

Prepared cDNAs were used to confirm the success of E2 treatment with a known E2 responsive gene *TFF1* (trefoil factor 1, pS2) primers [43]. PCR reaction was performed under following conditions: denaturation at 94˚C- 5 min incubation, 18 cycles at 94˚C- 30 sec, 61˚C (annealing temperature)- 30 sec, 72˚C- 30 sec, and final extension at 72˚C- 10 min. E2 treated MCF7 cDNA was used as positive control and no template control was used to check contamination (Figure C.1. in Appendix C).

2.10. RT-qPCR Analysis

RT-qPCR was performed with CFX Connect Real-Time PCR Detection System (Bio-Rad). iTaq™ Universal SYBR® Green Supermix (cat # 1725121, Bio-Rad) was used in reactions. 10 µL reactions were prepared with 500 nM of *RALGAPA2* short primers and 400 nM of *RALGAPA2* long primers, 300 nM of *TMEM 164* short primers and 400 nM of *TMEM 164* long primers. All RT-qPCR reactions were performed in following conditions but variable annealing temperatures: incubation at 94˚C-10minutes, 30 cycles of 94˚C- 30 sec, X˚C(annealing temperature)- 30 sec and 72˚C-30 sec, and final extension at 72˚C- 5 min. Primers are given in Table B.1. in Appendix B.

RALGAPA2 transcripts were amplified with *RALGAPA2*_Short F (RACE F2): 5'-GACCTGCCTCTGCTGTCATT-3', R: 5'- GATGAGGTGAGTGTGGGTGG-3' (product size: 120 bp, annealing temperature: 61˚) and *RALGAPA2*_Long F: 5'-GCCAGACTCACTCTTGGGAC-3', R: 5'-TTTGGGGCACCCTCATTCTC-3' (product size: 179 bp, annealing temperature: 56˚).

TMEM164 transcripts were amplified with *TMEM164*_Short F: 5'-TGGTAAACACTCGGCTGCTC-3', R: 5'- CTGAGGGGCTCTGGAGTGTA -3' (product size: 121 bp, annealing temperature: 62˚) and *TMEM164*_Long F: 5'-

TCTTTGAAGGCAGGGCCAAA-3', R: 5'-TGTAGCAGTTTGACGGTGGG-3' (product size: 112 bp, annealing temperature: 62˚).

The fold changes of the transcripts were normalized with the fold changes of reference gene *RPLP0*. *RPLP0* transcript was amplified with RPLP0_F: 5'-GGAGAAACTGCTGCCTCATA-3', RPLP0_R: 5'- GGAAAAAGGAGGTCTTC TCG-3' (annealing temperature: 60˚).

TFF1 expression was used as a E2 treatment positive control. *TFF1* was amplified with *TFF1*_F: 5'- TTGTGGTTTTCCTGGTGTCA -3' and *TFF1*_R: 5'-CCGAGCTCTGGGACTAATCA -3' (annealing temperature: 56˚).

The reaction efficiency combined ΔΔCq formula was used for the calculations of relative quantification.

2.11. Rapid Amplification of cDNA Ends

By using the oligo dT-anchor primer, RACE ready cDNA synthesis was performed with 5µg total RNA (DNase treated) taken from E2 and EtOH treated MCF7 cells. After cDNA synthesis, 3'RACE PCR was performed with gene specific forward primer and anchor sequence specific reverse primer. Gene Specific Primers are *TMEM164*_3'RACE_Forward 1: 5'- TACACTCCAGAGCCCCTCAG -3' (product size: 437 bp), *TMEM164*_3'RACE_Forward 2: 5'- GGGAAGCTGGTCATCCTGTT -3' (product size: 227 bp), *RALGAPA2* 3'RACE Forward 2: 5'-GACCTGCCTCTGCTGTCATT -3' (product size: 393 bp). All of them were designed and used as forward primers since the anchor reverse primer is the same for all the 3'RACE reactions.

For *TMEM164*, first round 3'-RACE PCR was performed using the *TMEM164*_3'RACE_Forward 1 and Anchor-R primers with the following PCR conditions; 95˚C- 3 min,35 cycles of 95˚C- 30 sec, 68˚C- 30 sec, 72˚C -30 sec and

final extension at 72˚C- 5 min. After running the RACE product on 1.5% agarose gel and extracting the correct sized band with Nucleospin® Gel and PCR Clean-up kit (REF 740609.50, MN), nested 3'RACE PCR was performed using the *TMEM164*_3'RACE_Forward 2 and Anchor-R primers with 2 µl extracted RACE product as template. Same reaction conditions were used. The expected sized band having 227 bp length was observed after running the product on 1.5% agarose gel.

For *RALGAPA2*, first round 3'-RACE PCR was performed using the *RALGAPA2* 3'RACE Forward 2 and Anchor-R primers with the following PCR conditions; 95˚C - 3 min,35 cycles of 95˚C- 30 sec, 68˚C- 30 sec, 72˚C- 30 sec and final extension at 72˚C- 5 min. After running the RACE product on 1.5% agarose gel and extracting the correct sized band with Nucleospin® Gel and PCR Clean-up kit (REF 740609.50, MN), second round 3'RACE was performed by using extracted RACE products as template with the same reaction conditions. The expected sized band having 393 bp length was observed after running the product on 1.5% agarose gel.

2.12. Cloning of 3'UTR Isoforms

3'RACE product of TMEM164 was extracted from 1.5% agarose gel after running. The purified product was quantified by MaestroNano Spectrophotometer. In order to ligate the RACE product into empty vector, the amount and volume of insert is calculated with the equations below:

$$\text{Amount of insert (ng)} : = \frac{\text{Amount of vector (ng) x Molecular Weight of insert (kb)}}{\text{Molecular Weight of vector (kb)}} x \frac{3}{1}$$

$$\text{Volume of insert (µl)} = \frac{\text{Amount of insert (ng)}}{\text{Concentration of extracted product } \left(\frac{\text{ng}}{\text{µl}}\right)}$$

*’3/1’ is the ratio between amount of insert and vector.

50 ng pGEM®-T Easy Vector (cat # A1360, Promega) and RACE product were ligated by using T4 DNA ligase (cat # 04716728001, Roche) at 4˚C for 16h. After ligation, transformation of the ligation product (10 µl) was performed with competent TOP10 *E.coli* cells. After the incubation of bacterial transformation plates, the colonies were observed and colony PCR was performed to confirm the existence of the insert with *TMEM164*_3'RACE_Forward2: 5’- GGGAAGCTGGTCATCCTGTT-3’ and Anchor_R:5’-GACCACGCGTATCGATGTCGAC-3’ primers. Conditions: incubation at 94˚C- 10 min, 35 cycles of 94˚C- 30 sec, 58˚C- 30 sec and 72˚C- 30 sec, and final extension at 72˚C- 10 min. According to colony PCR results, a colony having *TMEM164* 3’RACE product was taken and plasmid isolation was performed. Purified DNA was sent to sequencing.

CHAPTER 3

RESULTS AND DISCUSSION

3.1. IGV Snapshots

E2 is a proliferative signal, inducing an E2 responsive transcriptional pattern. E2 effects on an APA regulated *CDC6* was reported in our lab [28]. Interestingly, transcription induced DNA breaks are known to open up chromatin structure [31]. We hypothesized these breaks to aid APA. In light of these, we wanted to further investigate this potential link between E2 induced genes, APA and transcription coupled DNA breaks.

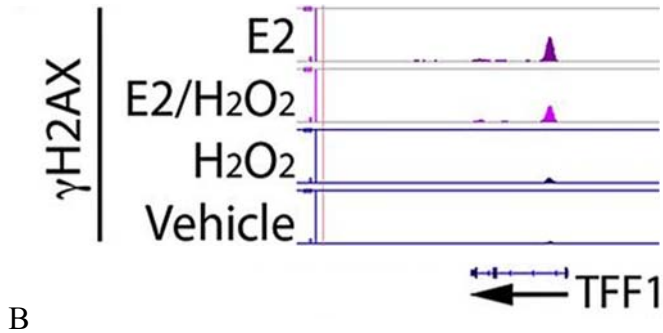
We first analyzed GSE57426 [40] dataset which contains γH2AX ChIP-Seq data in MCF7 cell line treated with E2, H_2O_2 and Ethanol (vehicle). γH2AX is a marker for double strand DNA breaks and its accumulation is a prerequisite for the oncoming DNA repair machinery [34].

We selected candidate genes according to their E2/ H_2O_2 peak ratios to study the effect of E2 induced DNA break patterns at around potential polyA sites. Two genes caught our attention in this data set; *TMEM164* and *RALGAPA2*, we observed >7.7 fold (for *TMEM164)* and >8.7 fold (for *RALGAPA2)* higher peaks of γH2AX in E2 treated samples around poly (A) sites in compared to H_2O_2 treated and EtOH (vehicle) samples, which indicates the presence of DNA strand breaks specified by γH2AX marker in that specific region (Figure 3.2, 3.3). EtOH (vehicle) samples were used as control groups and H_2O_2 treatment was used for a better comparison of γH2AX levels since H_2O_2 causes random dsDNA breaks [40].

To first confirm the functionality of E2 activity, we turned into the *TFF1* locus known to harbor E2 induced dsDNA breaks at its promoter region [41]. *TFF1* gene which was used as a positive control, had accumulation of γH2AX at its promoter region in E2 treated samples compared to H_2O_2 treated cells (Figıre 3.1. B). γH2AX peak ratio of E2/ H_2O_2 was 279/55 for *TFF1* gene meaning that γH2AX accumulation was 5 folds higher in E2 sample compared to H_2O_2.

Of note, we did not observe γH2AX peaks at around poly (A) site of *TFF1* gene which is predicted to have one polyA site (Hs.162807.1.1) according to polyA_DB [48].

A



B

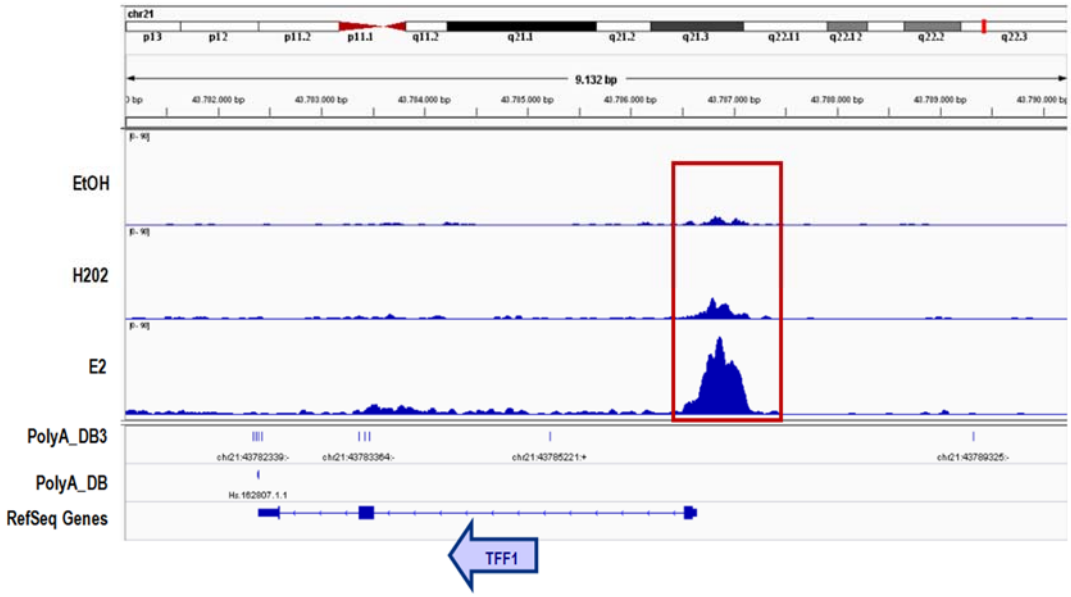
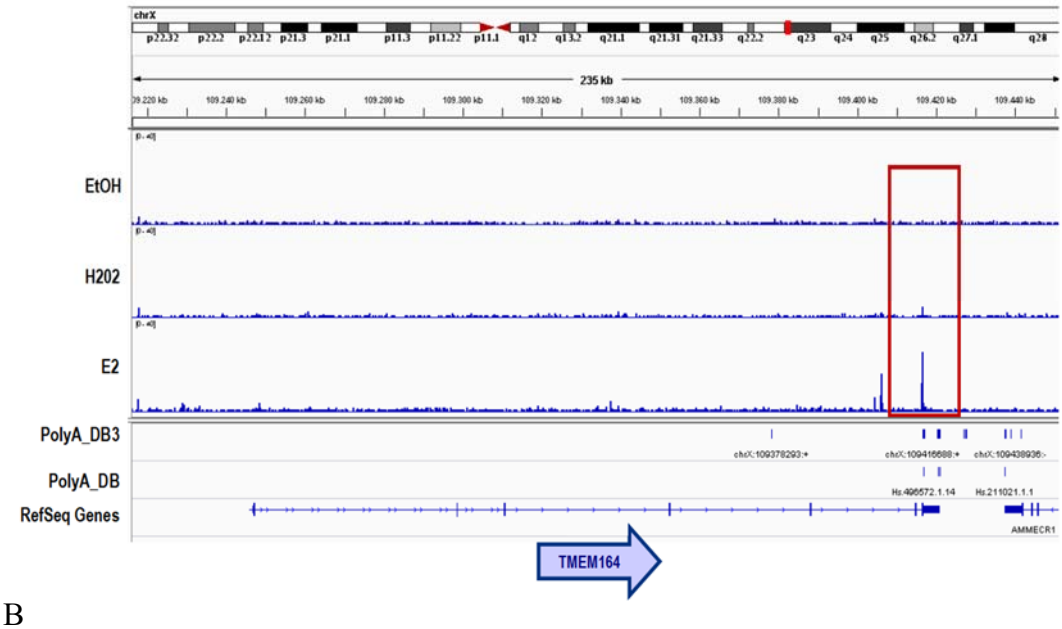


Figure 3.1. *TFF1* IGV Snapshot

A.IGV snapshot taken from the original study [40] showing γH2AX peaks of *TFF1* when cells were treated with 100nM E2, H_2O_2 and EtOH for 45 min. Data range of IGV:1-800. B. IGV snapshot of the ChIP-Seq data showing γH2AX peaks at *TFF1* promoter when cells were treated with 100 nM E2, H_2O_2 or vehicle for 45 minutes. Peak ratios (E2/ H_2O_2= 279/55) of the samples at the promoter region indicated in red box. Data range of IGV is adjusted to 0-300.

Our first candidate, *TMEM164* has 4 poly A sites according to polyA_DB [48]. Interestingly, we did not observe any accumulation of γH2AX at the promoter of *TMEM164*. Intriguingly, the peaks were mapping at around the 3'UTR. γH2AX peaks were co-localized to polyA site Hs.496572.1.14, in E2 treated sample compared to H_2O_2 or EtOH samples (Figure 3.2).

A



B

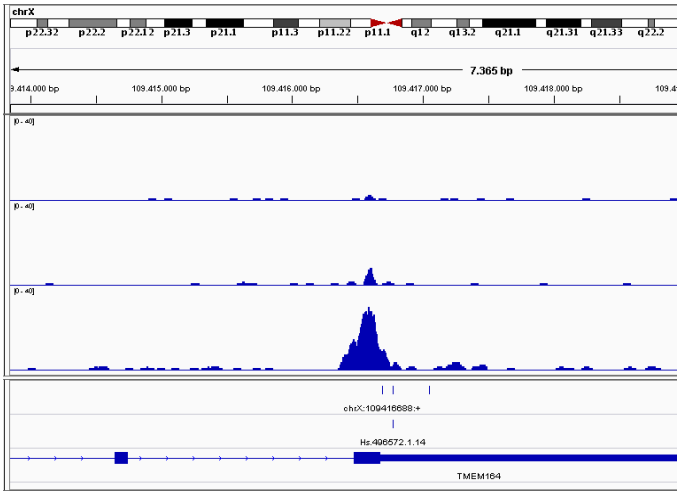


Figure 3.2. *TMEM164* IGV Snapshot

A.IGV snapshot showing γH2AX peaks of *TMEM164* in cells treated with 100nM E2, H_2O_2 and EtOH for 45 min. Peak ratios (E2/ H_2O_2= 100/13) of the samples at the 3'UTR region around Hs.496572.1.14 poly A site is indicated in the red box. Data range of IGV is adjusted to 0-110. B. IGV snapshot zoomed to the 3'UTR region and poly A sites of *TMEM164*.

Although E2 responsiveness of *TMEM164* has been shown in previous studies [45], we did not observe considerable peaks at the promoter region of *TMEM164* in contrast to *TFF1* gene. However, the peak ratio of E2/H_2O_2 was 100/13 showing that with E2 treatment, the availability of γH2AX increases at the proximal polyA site (Hs.496572.1.14). Interestingly, since γH2AX is the marker of DNA break, the result suggested the presence of a potential dsDNA break at the poly A site.

To test whether overall transcription of these candidates were even responsive to E2, we used the GSE62789 dataset [47] of RNA-Seq data performed for 5, 10, 20, 40, 80, 160, 320, 640, 1280 minutes of E2 (10nM) treatment in MCF7 samples.

For RNA-Seq analysis, we first converted FASTQ files to BEDGRAPH files, using CGC tools (http://www.cancergenomicscloud.org/). According to IGV snapshots, the expression of *TMEM164* shows a gradual increase with increasing E2 treatment time points (Figure 3.3). There was 5 folds of increase in 1280 min E2 treated sample compared to untreated sample (1280min E2/untreated= 291/ 58).

The studies showing E2 responsiveness and the RNA-Seq result of *TMEM164* support the idea that we can study E2 induced transcription with *TMEM164* since it responds to E2 with upregulated expression (Figure 3.3).

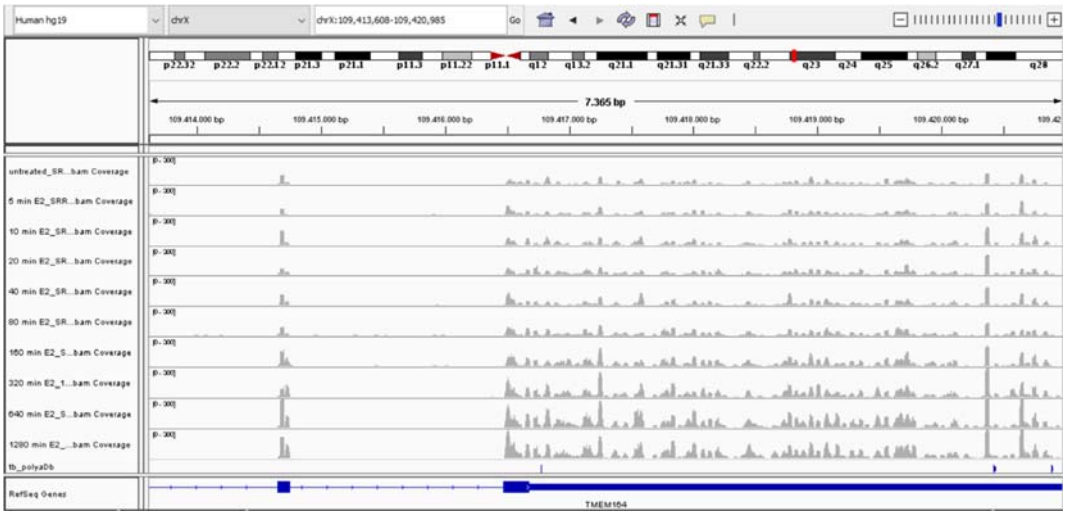


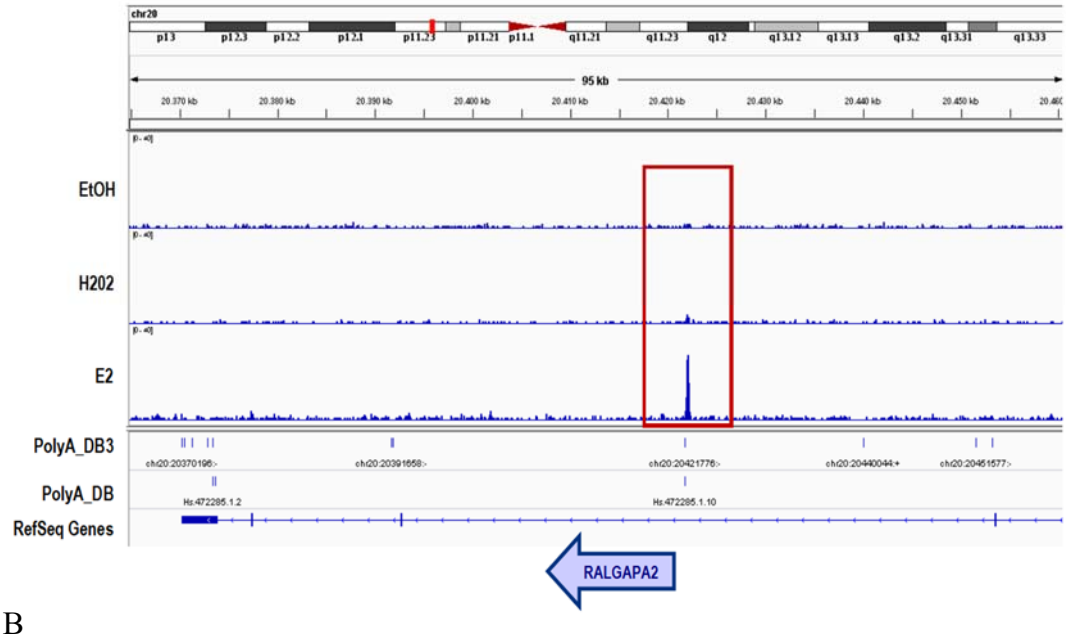
Figure 3.3. *TMEM164* RNA-Seq IGV Snapshot

IGV snapshot showing *TMEM164* expression in for 5, 10, 20, 40, 80, 160, 320, 640, 1280 min E2 treated MCF7 samples. Peak ratios of 1280min E2/untreated is 291/ 58. Data range of IGV is adjusted to 0-300.

For the *RALGAPA2* gene, there are 5 reported poly A sites according to polyA_DB [48]. We observed higher peaks around poly A site Hs.472285.1.10, for E2 sample compared to H_2O_2 and EtOH samples (Figure 3.4). E2/H_2O_2 ratio of γH2AX peaks was 114/13. In this case, poly A site having the highest peak was located in intron. Similar to *TMEM164*, we did not observe γH2AX peaks at the promoter region of *RALGAPA2* despite its E2 responsiveness which was shown in a study that demonstrated *RALGAPA2* expression to increase more than 3 folds after 40 minutes of E2 treatment [46].

To this end, we observed two candidate genes that had γH2AX accumulation at around poly A sites following E2 treatment.

A



B

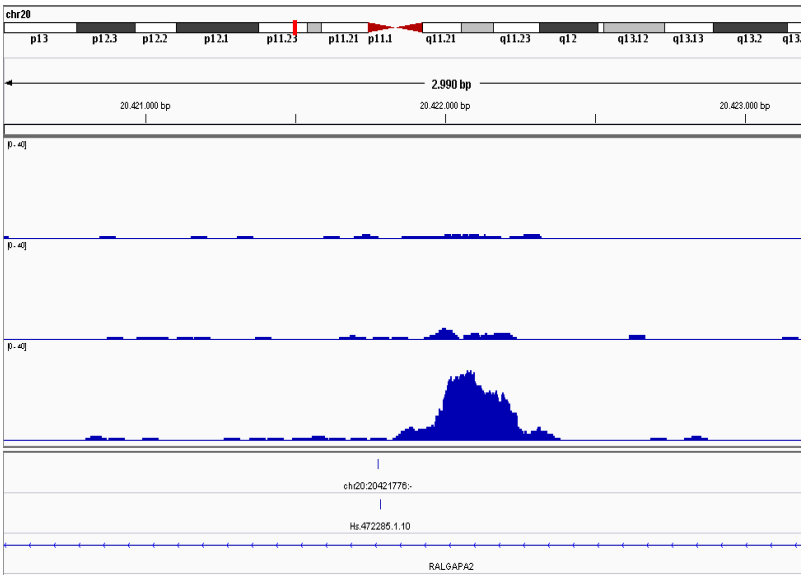
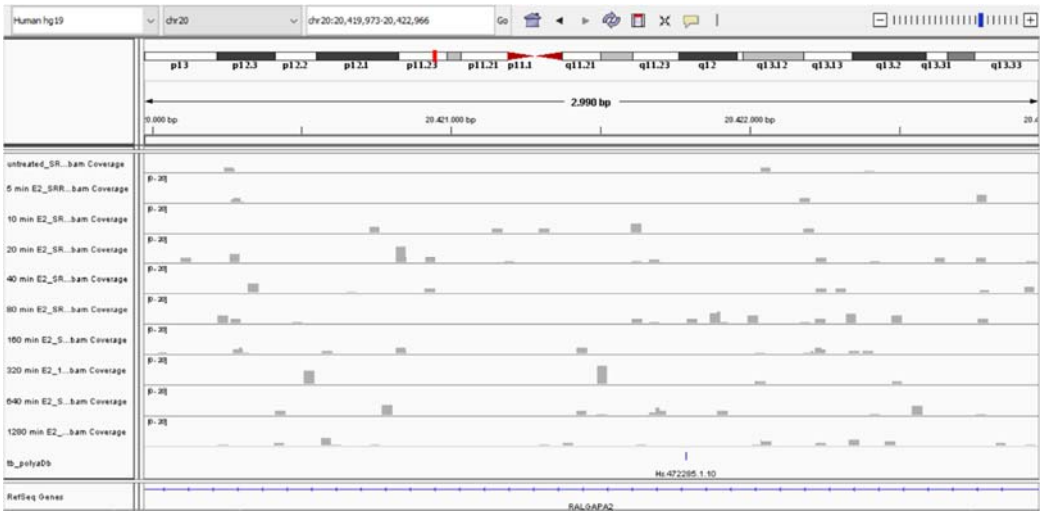


Figure 3.4. *RALGAPA2* IGV Snapshot

A. IGV snapshot showing γH2AX peaks of *RALGAPA2* when cells are treated with 100nM E2, H_2O_2 and EtOH for 45 min. Peak ratios (E2/ H_2O_2= 114/13) of the samples at around Hs.472285.1.10 poly A site is indicated in red box. Data range of IGV is adjusted to 0-120. B. IGV snapshot zoomed to the poly A site located in intron of *RALGAPA2*.

However, for *RALGAPA2*, while overall mRNA levels were increasing (although not significantly) in response to E2 treatment, we did not observe reads from the intronic region that would have been expressed if Hs.472285.1.10 was chosen. (Figure 3.5.A). However, we observed reads from 38/40 exon located after intronic Hs.472285.1.10. The expressions of mRNA levels were increasing until 40 min E2 treatment, then the expressions started to decrease with increasing E2 treatment time points.

A



B

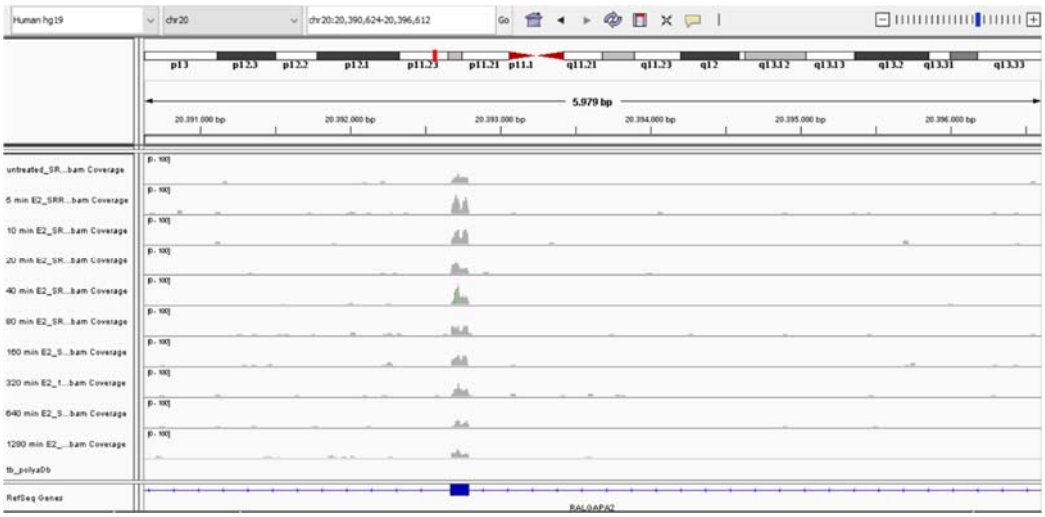


Figure 3.5. RNA-Seq IGV Snapshot

A.IGV snapshot showing *RALGAPA2* expression at around poly A site Hs.472285.1.10 for 5, 10, 20, 40, 80, 160, 320, 640, 1280 min E2 treated MCF7 samples. Data range of IGV is adjusted to 0-20. B.IGV snapshot showing *RALGAPA2* expression at exon (38/40) for 5, 10, 20, 40, 80, 160, 320, 640, 1280 min E2 treated MCF7 samples. Data range of IGV is adjusted to 0-100

To further investigate the expressions of candidate genes in the presence of E2, we also turned to existing microarray datasets.

3.2. Probe Distributions and Microarray Datasets Comparison in GEO2R

To test whether the mentioned poly A sites are indeed differentially used in response to E2, we sought to detect 3'UTR isoforms that may form in response to APA. To call for such isoforms, we turned to existing microarray data. Therefore, first, to find evidence about the expression of isoforms generated by the poly A sites, NetAffx Query was used to search probe sets based on poly (A) site locations according to Human Genome U133 Plus 2.0 Array (HGU133Plus2, GEO Accession number: GPL570). Probes obtained from NetAffx were viewed on UCSC Browser (Figure 3.6, Figure 3.7).

For *TMEM164*, 3 probe sets were determined: 223201_s_at, 223202_s_at, 220486_x_at.

Figure 3.6. Probe Distributions of *TMEM164*

Snapshot taken from UCSC Browser. 4 poly A sites are shown. The poly A site (Hs.496572.1.14) having the highest peak is circled in red. Red line indicates the separation of probe sets based on the location of circled poly A site. Probe sets based on Human Genome U133 Plus 2.0 Array are 223201_s_at, 223202_s_at, 220486_x_at.

However, all the probes were located after the poly A site Hs.496572.1.14. Therefore, these probe sets were not appropriate to distinguish 3'UTR isoforms.

For *RALGAPA2*, 6 probe sets were determined; 225499_at, 231826_at, 1559699_at, 232500_at, 234934_at, 239660_at. Three essential probe sets for comparison were chosen as 1559699_at, 225499_at and 232500_at which mapped to proximal and distal regions of the polyA site Hs.472285.1.10.

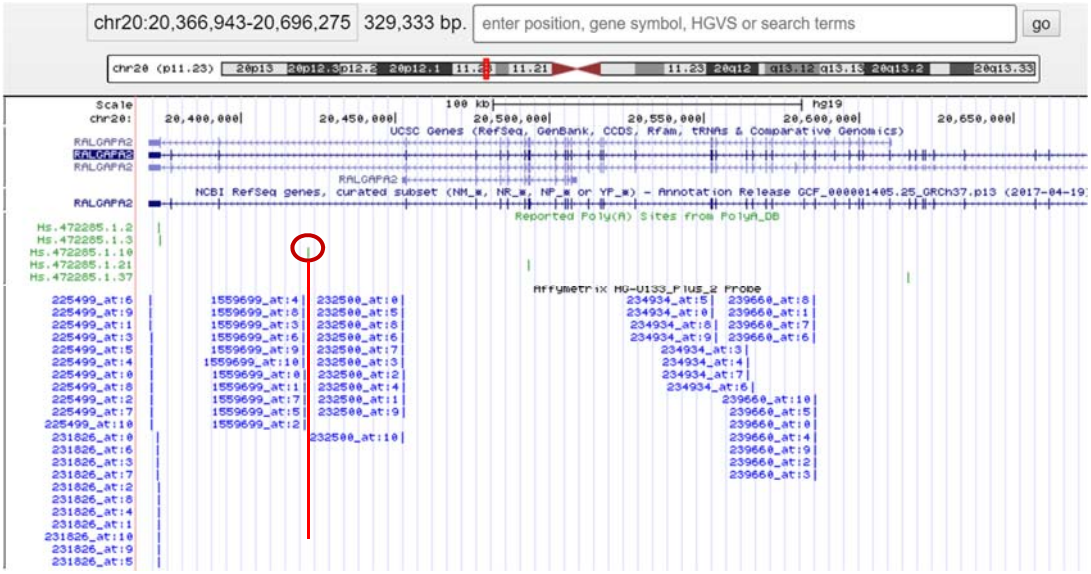


Figure 3.7. Probe Distributions of *RALGAPA2*

Snapshot taken from UCSC Browser. 5 poly A sites are shown. The poly A site Hs.472285.1.10 is circled in red. Red line indicates the separation of probe sets based on the location of circled poly A site. Essential probe sets for comparison based on Human Genome U133 Plus 2.0 Array are 1559699_at, 225499_at and 232500_at.

Then, to investigate and identify differential expressions of our candidate genes in different experiments, the obtained probe sets were used for GEO2R comparison with microarray datasets GSE11324 [41] and GSE8597 [42] which have E2 treated and control samples of MCF7 (Table A.1,Table A.2 in Appendix A).

3.2.1. *TMEM164* Expression in GSE11324

In GSE11324 microarray data set 0h, 3h, 6h,12h E2 (100nM) treated MCF7 samples were used. GEO2R comparison was performed with 223201_s_at, 223202_s_at and 220486_x_at probe sets by using GSE11324 data set (Figure 3.8). Since all the probe sets (Figure 3.8.A-C) were located after the poly A site Hs.496572.1.14, we could not distinguish between potential isoforms that may use these 3 distal poly A sites.

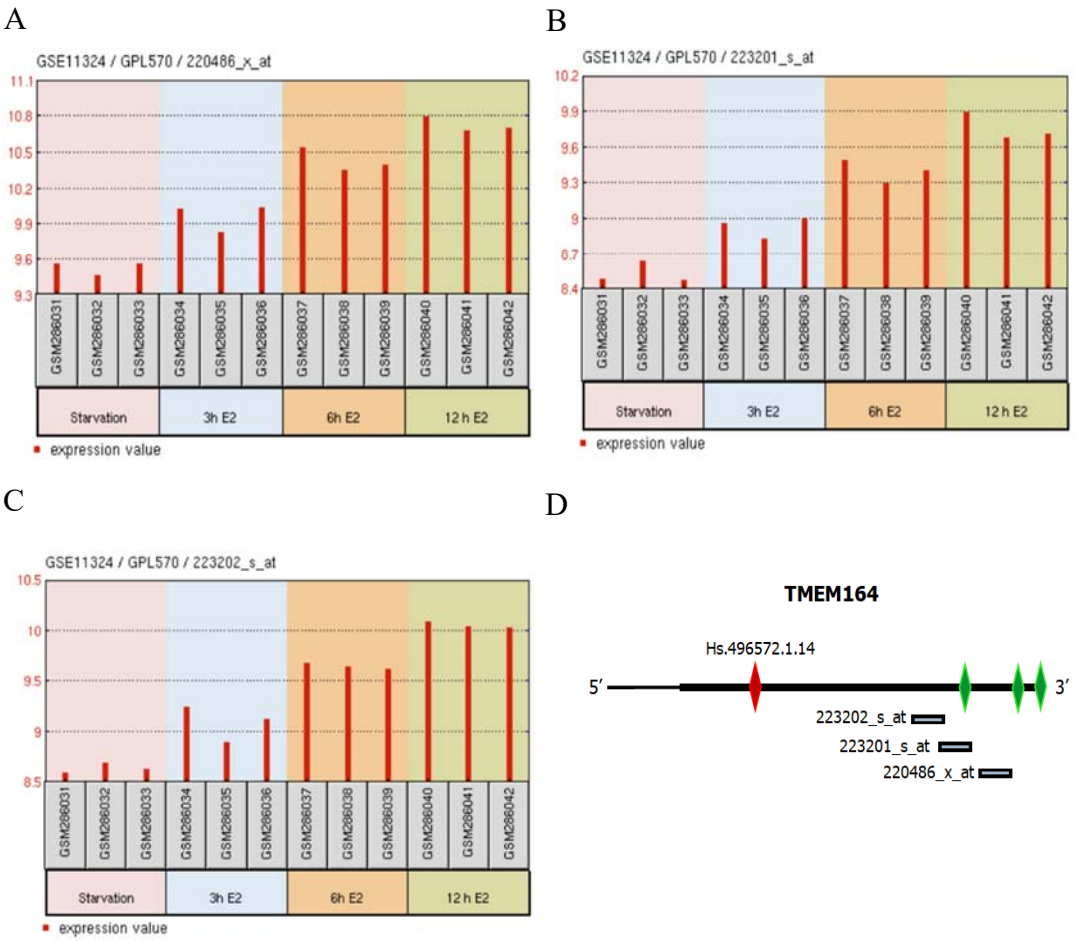


Figure 3.8. GEO2R Comparison of GSE11324 for *TMEM164*

GSE11324 data set containing 0h, 3h, 6h,12h E2 (100 nM) treated MCF7 samples. A. Data set compared according to 220486_x_at probe sets. B. Data set compared according to 223201_s_at probe sets. C. Data set compared according to 223202_s_at probe sets. D. *TMEM164* poly A sites and probe locations. Poly A site Hs.496572.1.14 is indicated in red, others are indicated in green.

Nevertheless, it was clear that the expression of *TMEM164* increased 1.15 fold in response to E2 treatment compared to control samples.

3.2.2. *TMEM164* Expression in GSE8597

In GSE8597 data set 24h E2 (25 nM) and EtOH treated MCF7 samples were used. GEO2R comparison was performed with 223201_s_at, 223202_s_at, 220486_x_at probe sets by using GSE8597 data set (Figure 3.9).

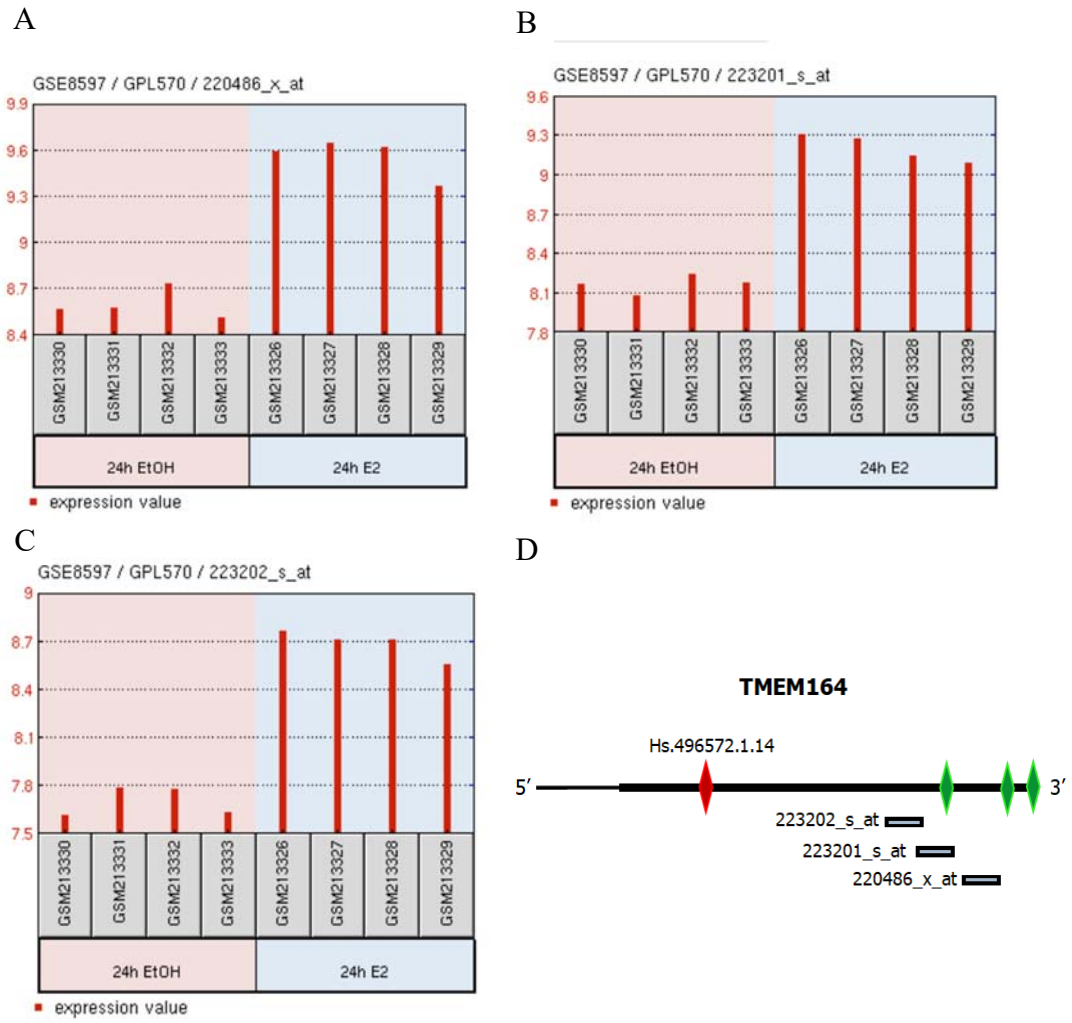


Figure 3.9. GEO2R Comparison of GSE8597 for *TMEM164*

GSE8597 data set containing 24h E2 (25 nM) and EtOH treated MCF7 samples. A. Data set compared according to 220486_x_at probe sets. B. Data set compared according to 223201_s_at probe sets. C. Data set compared according to 223202_s_at probe sets. D. *TMEM164* poly A sites and probe locations. Poly A site Hs.496572.1.14 is indicated in red, others are indicated in green.

The expression of *TMEM164* was 1.2 fold higher in E2 treated samples for all the probes when compared to EtOH treated ones according to probes (Figure 3.8.A-C) located after the poly A site Hs.496572.1.14. These results suggests that the expression of *TMEM164* increased with E2 treatment.

3.2.3. *RALGAPA2* Expression in GSE11324

In GSE11324 data set 0h, 3h, 6h,12h E2 (100 nM) treated MCF7 samples were used. GEO2R comparison was performed with 1559699_at, 225499_at, 232500_at probe sets by using GSE11324 data set (Figure 3.10).

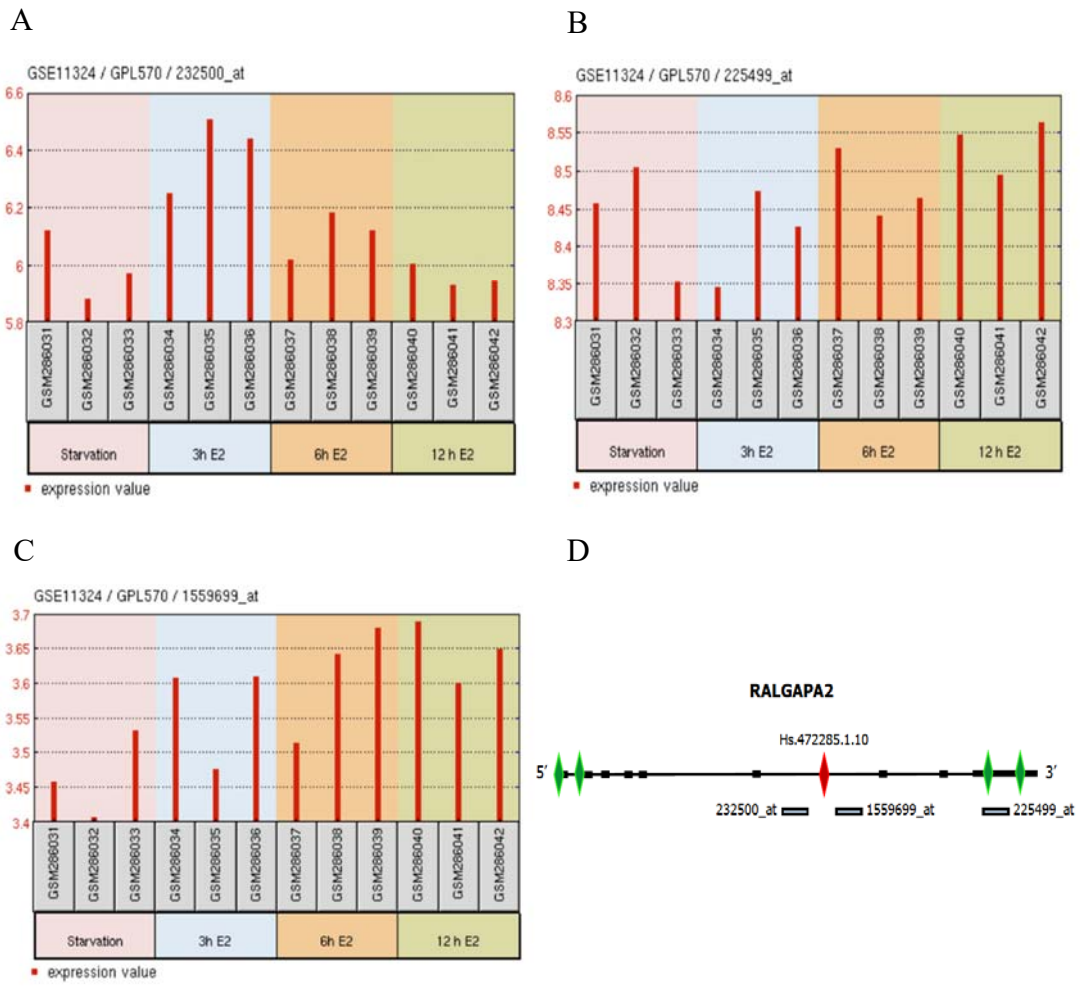


Figure 3.10. GEO2R Comparison of GSE11324 for *RALGAPA2*

GSE11324 data set containing 0h, 3h, 6h,12h E2 (100nM) treated MCF7 samples. A. Data set compared according to 232500_at probe sets. B. Data set compared according to 225499_at probe sets. C. Data set compared according to 1559699_at probe sets. D. *RALGAPA2* poly A sites and probe locations. Poly A site Hs.472285.1.10 is indicated in red, others are indicated in green.

According to probe set 232500_at (Figure 3.10.A) located before the poly A site Hs.472285.1.10, the expression of the isoform first increased with E2 treatment until 3h, then, decreased with increasing E2 treatment time points. On the other hand, the isoforms generated by the recognition of other poly A sites in 3'UTR (Hs.472285.1.2, Hs.472285.1.3) increased with increasing E2 treatment time points. These isoforms are recognized by the 225499_at probe set located in 3'UTR (Figure 3.9.C). However, the isoform generated by the recognition of intronic poly A site Hs.472285.1.10, can be recognized by only the 232500_at probe set and not by 225499_at or 1559699_at probe sets (Figure 3.10.B-C).

By looking at these results, we can say that more than 3 hours of E2 treatment may favor the recognition of poly A sites in 3'UTR (Hs.472285.1.2, Hs.472285.1.3) but not the intronic poly A site Hs.472285.1.10. The RNA-Seq result in Figure 3.5.A which we did not see the expression of isoform generated by the recognition of Hs.472285.1.10, also supports the evidence that the expression of this isoform was very low.

3.2.4. *RALGAPA2* Expression in GSE8597

GSE8597 data set has expression data from 24h E2 (25 nM) and EtOH treated MCF7 cells. GEO2R expression analysis was performed with with 1559699_at, 225499_at, 232500_at probe sets of the GSE8597 data set.

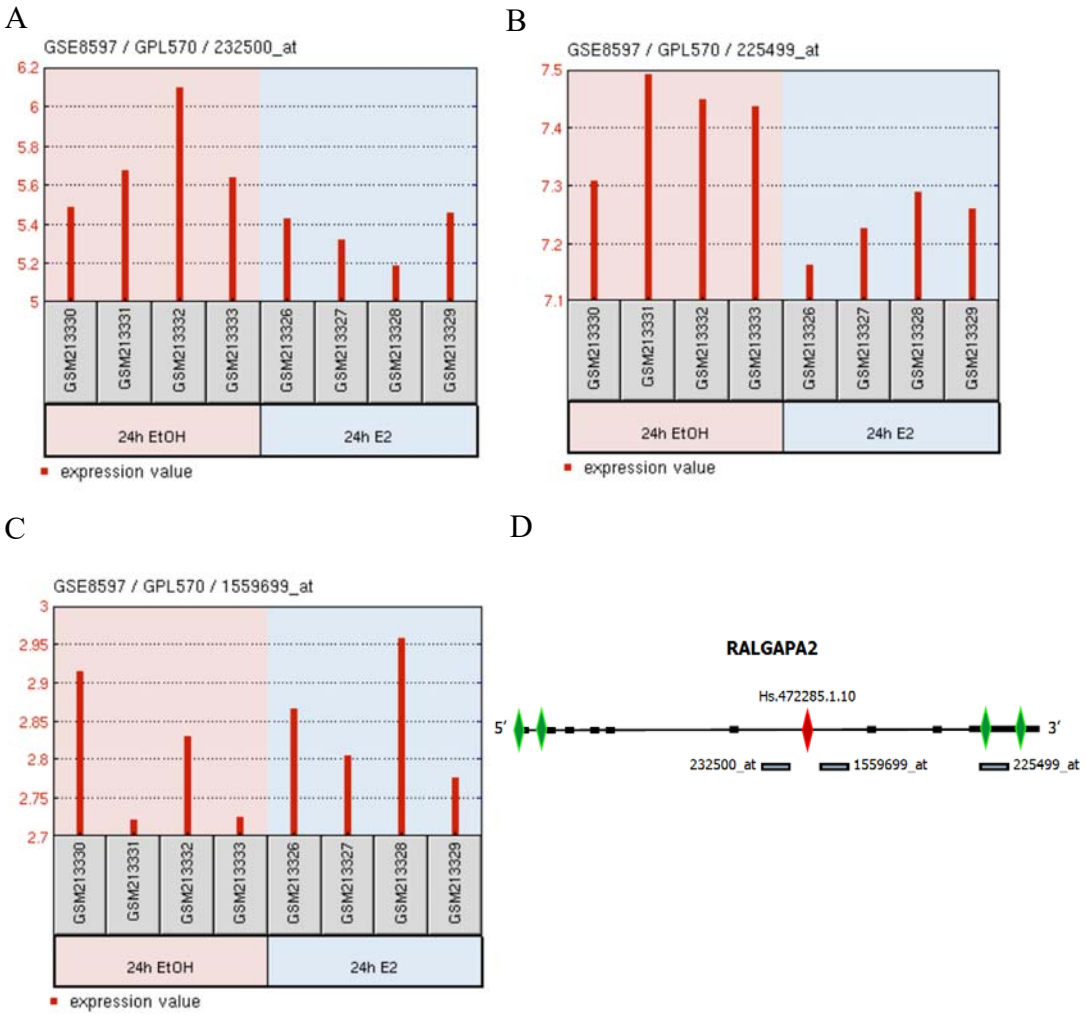


Figure 3.11. GEO2R Comparison of GSE8597 for *RALGAPA2*

GSE8597 data set containing 24h E2 (25 nM) and EtOH treated MCF7 samples. A. Data set compared according to 232500_at probe sets. B. Data set compared according to 225499_at probe sets. C. Data set compared according to 1559699_at probe sets. D. *RALGAPA2* poly A sites and probe locations. Poly A site Hs.472285.1.10 is indicated in red, others are indicated in green.

According to probe set 232500_at (Figure 3.11.A) mapping proximal to the intronic poly A site Hs.472285.1.10, the expression of the isoform generated by that poly A site was one fold lower in E2 treated samples compared to EtOH (vehicle) treated samples. On the other hand, the isoforms generated by the recognition of other poly A sites in 3'UTR (Hs.472285.1.2, Hs.472285.1.3) had approximately same expressions in E2 and EtOH treated samples detected by the 225499_at probe set (Figure 3.11.C).

These results may suggest that, the expression of *RALGAPA2* isoform generated by the recognition of intronic poly A site Hs.472285.1.10 was decreasing in the presence of E2. However, this decrease may not be significant enough to make a conclusion about the expressions of isoforms. Therefore, we needed to confirm these in silico results with experiments.

3.3. E2 Treatment

To analyze the expression levels and confirm APA events in the presence of E2 for the *TMEM164* and *RALGAPA2* transcripts, first, ER+ breast cancer cell line MCF7 was treated with 100 nM E2 after growing cells in serum-deprived medium for 72 hours. MCF7 cells were treated with E2 for 12 hours based on RNA-seq and/or microarray data presented in 3.2. After RNA isolation from E2 treated cells and cDNA synthesis, the treated cDNAs were used as template to check the success of E2 treatment. Since *TFF1* is a known E2 responsive gene [43], the effect of E2 treatment was confirmed in E2 treated ER+ cells with the observation of approximately 3 folds upregulation in *TFF1* expression. EtOH treated samples were expressing less *TFF1* than E2 treated samples as can be seen from the band intensities (Figure 3.12).

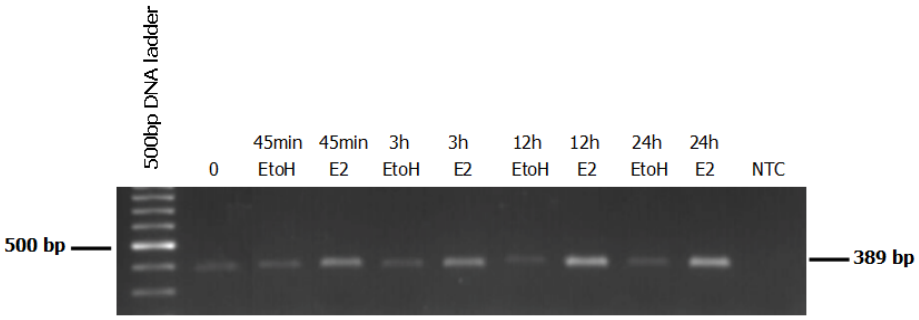


Figure 3.12. *TFF1* Upregulation with E2 Treatment

MCF7 cells were treated with vehicle (ethanol) or 100 nM E2 for 45min, 3h, 12h and 24h. Expected *TFF1* product size was 389 bp.

After the confirmation of E2 treatment in MCF7 cells, we proceeded with 3'RACE and then RT-qPCR quantification of *TMEM164* and *RALGAPA2* short and long isoforms by using the E2 treated cDNA samples as template.

3.4. 3'RACE

According to the results of ChIP-Seq and RNA-Seq data, 3'RACE forward primers are designed and 3'RACE was performed to confirm the existence of isoforms generated by the recognition of poly A site having the highest γH2AX peak among others (Figure 3.2, Figure 3.4). First, RACE ready cDNAs were synthesized with oligo dT-anchor primers. Then, anchor sequence and designed forward primers are used to amplify cDNA ends.

For *TMEM164*, first round 3'RACE was performed with *TMEM164*_3'RACE_Forward 1(F1) primer by using 12h E2 race ready cDNAs and 437 bp product was observed. After using the extracted 3'RACE F1 product as template, nested 3'RACE was performed with *TMEM164*_3'RACE_Forward 2 (F2) and the expected 227 bp sized band was observed (Figure 3.13). After nested 3'RACE

we could still observed the F1 product but the F2 bands were bright enough to be cut and extracted for cloning. With 3'RACE results of *TMEM164*, the isoforms generated by the recognition of poly A site Hs.496572.1.14 were detected.

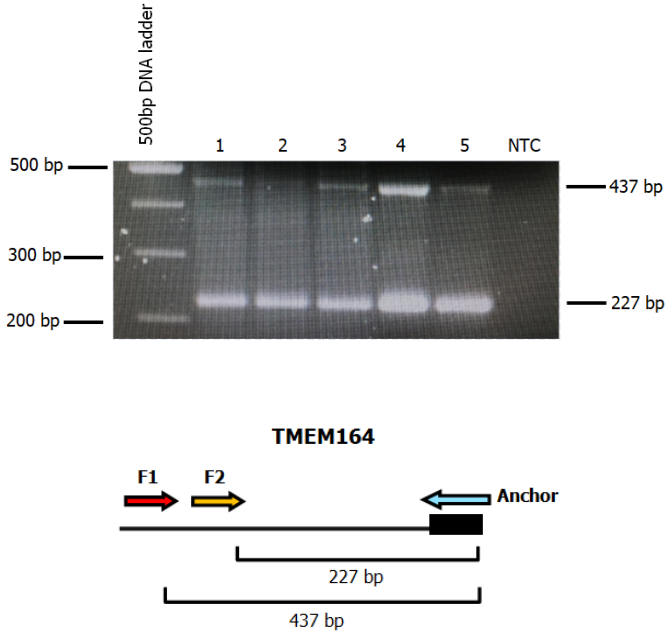


Figure 3.13. 3'RACE of *TMEM*164 F1-F2

*TMEM164*_3'RACE_Forward 1(F1) primer and Anchor-R primer are used. After 3'RACE with F1 primer, the products are used as template for F2 RACE. Lane 1 ,2 and 3 represents template volumes of F1 PCR products as 1 uL, 1.5 uL and 2 uL respectively. Lane 4 and 5 represents template volumes of extracted F1 products as 2 uL and 4uL respectively. F1 product size: 437 bp, F2 product size: 227 bp.

Sequencing of F2 product was needed to verify the presence of *TMEM164* isoform generated by the selected poly A site after E2 treatment. Therefore, we cloned the purified products into pGEM®-T Easy Vector with ligation and bacterial transformation. TOP10 E.coli colonies were obtained and colony PCR was performed with *TMEM164* F2 primer. The presence of *TMEM164* F2 product in

PGEMT was confirmed. After performing plasmid isolation from colony #3, the isolated DNA was sent to sequencing to make sure that DNA sequence of the isoform was correct (Appendix H).

Figure 3.14. Sequencing of *TMEM164* 3'RACE F2 Product

Used tmem164 F2 primer and RACE anchor reverse primer were highlighted in blue and yellow. Hs 496572.1.15 poly A site was indicated in red.

The sequencing result verified the existence of *TMEM164* F2 product generated by the recognition of poly A site Hs.496572.1.14 (Figure 3.14).

For *RALGAPA2*, first round 3'-RACE PCR was performed using the *RALGAPA2* 3'RACE_Forward_2 (F2) and Anchor-R primers by using 12h E2 race ready cDNAs and 393 bp product was observed (Figure 3.15). After extracting the 3'RACE band having 390 bp size, second round 3'RACE was performed by using extracted RACE product as template with the same reaction conditions since we could not observe a bright correct sized band and there were other nonspecific bands in the first 3'RACE. The expected sized band having 390 bp length was observed again after second round RACE. With 3'RACE results of *RALGAPA2*, the isoforms generated by the recognition of poly A site Hs.472285.1.10 were detected and they were ready for cloning.

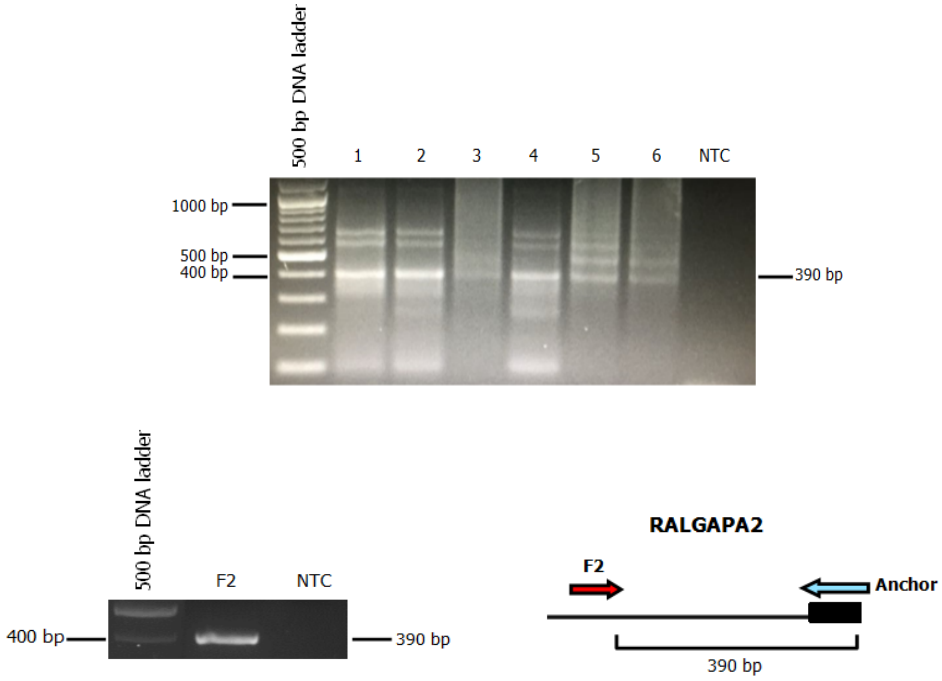


Figure 3.15. *RALGAPA2* F2

RALGAPA2 3'RACE Forward 2 and Anchor-R primers are used. Upper gel photo represents the first round 3' RACE. Lane 1 ,2, 3, 4, 5 and 6 represents template volumes of 12h E2 cDNA as 1 uL, 1.5 uL, 2 uL, 2.5 uL, 3 uL and 3.5 uL respectively. The gel photo below indicates the second round RACE result performed by using the first round RACE products (2 ul) as template. F2 product size: 390 bp.

Sequencing of F2 product was needed to verify the presence of *RALGAPA2* isoform generated by the selected poly A site after E2 treatment. We sent *RALGAPA2* F2 product to sequencing after cloning into pGEM®-T Easy Vector, but the sequencing results did not belong to *RALGAPA2* isoform. So, it means that we could not clone RALGAPA2 isoform which is observed in 3'RACE. We think this isoform is not very abundantly expressed.

3’RACE results of both *TMEM164* and *RALGAPA2* showed the existence of isoforms generated by the recognition of poly A site having the highest γH2AX peak (Figure 3.2, Figure 3.4).

As a result, we can say that *TMEM164* isoform generated by the recognition of poly A site Hs.496572.1.14 having the highest γH2AX peak in the presence of E2, exists and we verified its existence with sequencing result. But for *RALGAPA2,* even though we observed the correct sized band with 3’RACE, we cannot be sure about the presence of the isoform generated by the recognition of poly A site Hs.472285.1.10. We need to clone and sequence that isoform to ensure its existence.

The preliminary in-silico analyses and 3’RACE experiments may show a possible correlation between E2 induced transcription and poly A sites. However, we need to find more evidence to show that correlation.

Next, we proceeded with RT-qPCR to observe the expressions quantitatively

3.5. Expression Analysis

According to the information given in MIQE guidelines [44], optimization of RT-qPCR was performed. Reaction melt analyses were controlled accordingly to ensure that the reactions were lack of primer dimerization or other non-specific product formation. RT-qPCR assay results are shown in Appendix D.

For *TMEM164* results, there was >2.5 folds increase in short isoforms of E2 treated samples. Expression level of the long isoform of Hs.496572.1.20/21/22 also increased by 2 folds in E2 treated samples compared to EtOH treated cells (Figure 3.16).

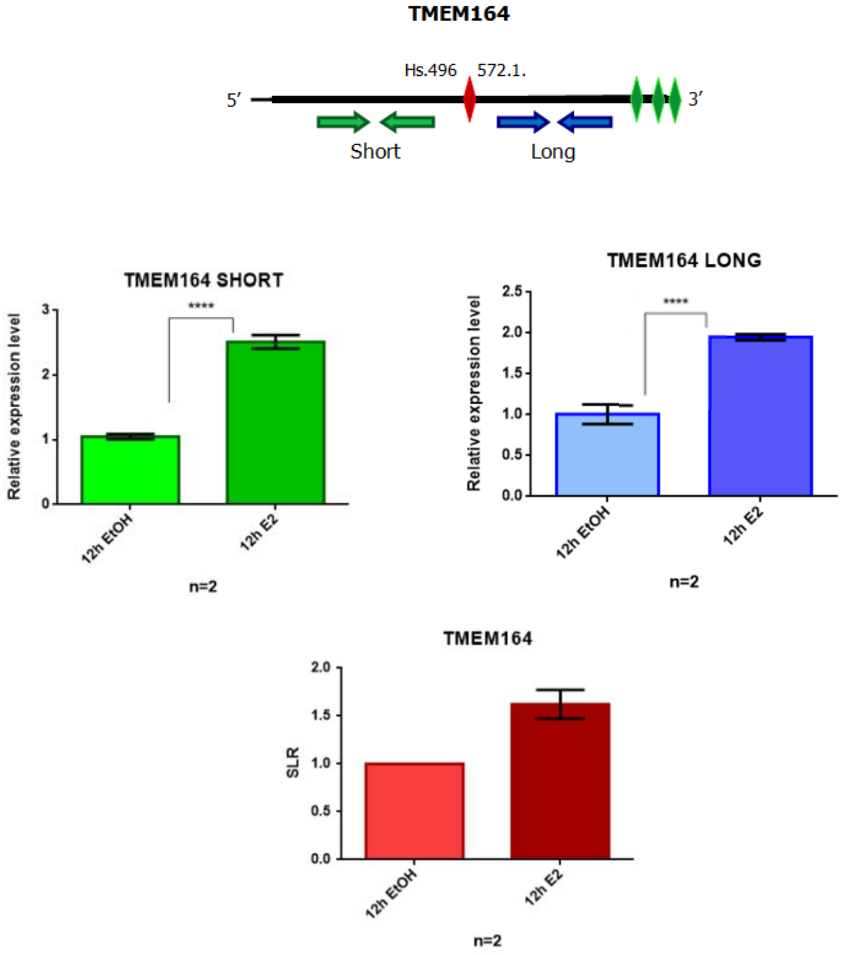


Figure 3.16. Relative Quantification of *TMEM164* Short and Long Isoforms

MCF7 cells were treated with 100 nM E2 and EtOH for 12 hours. The fold change for the isoforms was normalized against the reference gene; RPLP0. Quantification was done using the reaction efficiency correction and ΔΔCq method. **** indicates significant difference between E2 and EtOH isoforms' expression, $p<0.001$.

In agreement with in-silico results, the short 3'UTR isoform ending at the Hs.496572.1.14 poly site which is confirmed by 3'RACE (Figure 3.13) shows 2.5 folds increase of short isoform in E2 treated MCF7 cells compared to EtOH (vehicle) treated MCF7 cells. Short to long ratio (SLR) was 1.5 folds more in E2 treated cells, which may suggest a 3'UTR shortening event in *TMEM164*. These results may suggest that poly A site Hs.496572.1.14 with γH2AX accumulation in the presence of E2, may be favored with E2 treatment instead of other poly A sites Hs.496572.1.20/21/22 located on the pre-mrna.

For *RALGAPA2* results, there was a decrease in short isoform ending at intronic poly A site Hs.472285.1.10 and long isoform ending at poly A site Hs.472285.1.2 or Hs.472285.1.3 for E2 treated samples (Figure 3.17). The decrease in short isoform was one fold higher than the decrease in long isoform.

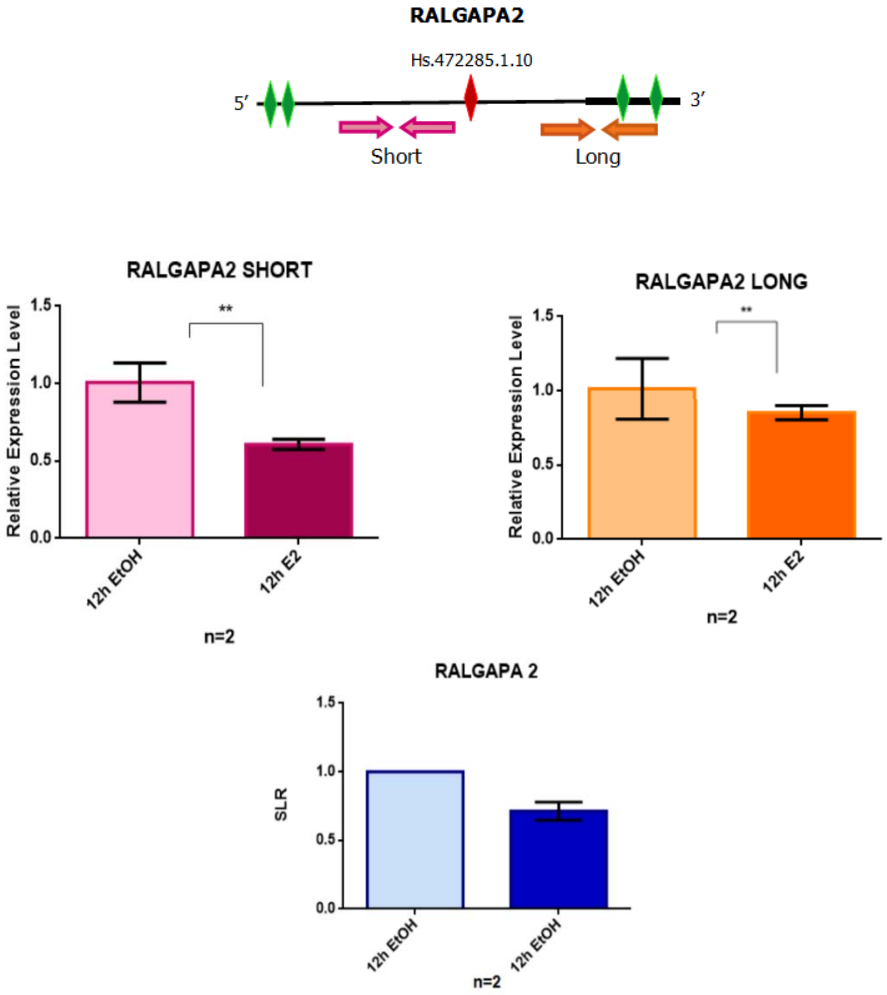


Figure 3.17. Relative Quantification of *RALGAPA2* Short and Long Isoform

MCF7 cells were treated with 100 nM E2 and EtOH for 12 hours. The fold change for the isoforms was normalized against the reference gene; RPLP0. Quantification was done using the reaction efficiency correction and ΔΔCq method. ***indicates significant difference between E2 and EtOH isoforms' expression, $p<0.001$.

In agreement with in-silico results shown in Figure 3.10 and 3.11, we can say that the expression of *RALGAPA2* short isoform was not increasing with E2 treatment. According to the GEO2R comparison results, the expression of probe sets located after the intronic poly A site Hs.472285.1.10 were increasing while the expression of probe sets located before the intronic poly A site Hs.472285.1.10 were decreasing with E2 treatment. These results may suggest that poly A site Hs.472285.1.10 which contains high levels of γH2AX accumulation in the presence of E2, may not be favored with E2 treatment.

According to these two different RT-qPCR results of two different genes, γH2AX accumulation with E2 treatment may or may not have a functional role in APA event.

For *TMEM164*, we might say that there is a correlation between dsDNA breaks and the site selection of the poly A site having γH2AX accumulation by looking at in-silico and experimental results. But for *RALGAPA2*, the poly A site having γH2AX accumulation was not selected over the other poly A sites which do not have γH2AX accumulation. There can be a case specific/gene specific effect or consequence of E2 induced γH2AX accumulation. It is also possible that RALGAPA was not a good candidate to test the hypothesis given its low expression.

To find more evidence about the mechanism underlying E2 induced APA and dsDNA breaks, we wanted to look at the survival plots of the genes to see whether their functional importance in survival of breast cancer may give us a clue about their functions and/or different expression levels.

3.6. Survival Plots

The survival plots were obtained from Kaplan-Meier (KM) Plotter which assesses the effect of genes on survival in different cancer types [50]. We wanted to use KM Plotter to assess correlation between our candidate genes' expression and survival of breast cancer.

We tested 2061 ER+ breast cancer patients for the effect of low and high expression of *TMEM164* isoform (short isoform) ending at poly A site Hs.496572.1.14 on relapse free survival (RFS). However, there was not any probe sets located before the poly A site Hs.496572.1.14 detecting short isoform. Therefore, we could only tested ER+ breast cancer patients for the effect of low and high expression of *TMEM164* long isoform ending at poly A site Hs.496572.1.20/21/22.

The RFS probability of patients was higher when *TMEM164* long isoform was highly expressed in compared to the low expression of *TMEM164* long isoform (Figure 3.18) with p value:0.048<0.05.

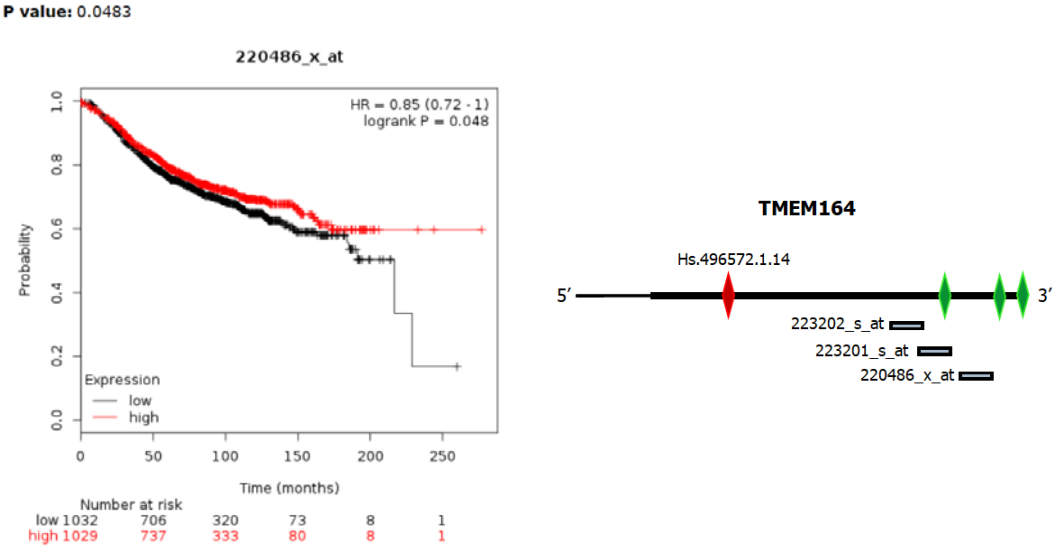


Figure 3.18. KM Plot of *TMEM164*

220486_x_at probe set expression was plotted in 2061 ER+ breast cancer patients. p value:0.048<0.05

This result suggested that the high expression of long isoform of *TMEM164* was useful for the patient survival so, the long isoform will probably not be favored by the cancer cells since it acts as a negative regulator for cancer / does not act as a positive regulator for cancer. The short isoform may be favored by cancer cells in agreement with the in-silico and experimental results suggested.

We tested 762 ER+ breast cancer patients for the effect of low and high expression of *RALGAPA2* short isoform ending at poly A site Hs.472285.1.10 on relapse free survival (RFS). We used probe set (232500_at) located in intron, before the poly A site Hs.472285.1.10 detecting short isoform (Figure 3.19.A) and probe set (225499_at) located in 3'UTR, after the poly A site Hs.472285.1.10 detecting long isoform (Figure 3.19.B).

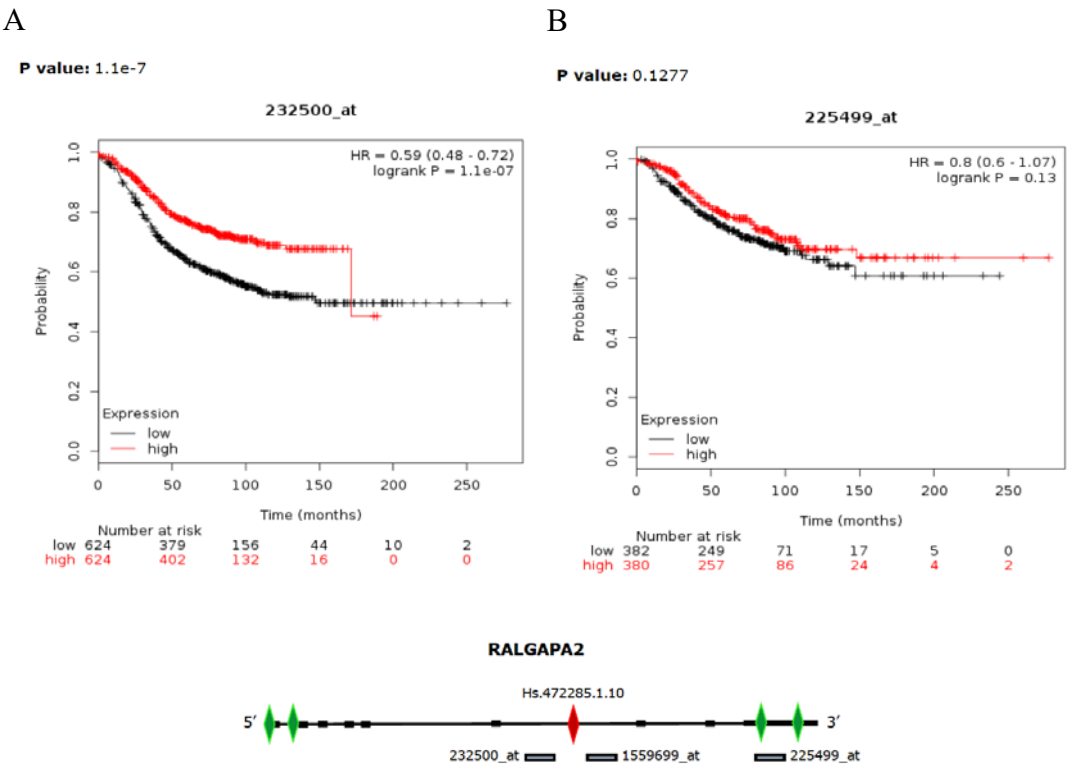


Figure 3.19. KM Plot of *RALGAPA2*

A.232500_at probe set expression was plotted in 762 ER+ breast cancer patients. p value:1.1e-7<0.05.
B. 225499_at probe set expression was plotted in 762 ER+ breast cancer patients. p value:0.13>0.05.

The RFS probability of patients was higher until 150 months when *RALGAPA2* short isoform was highly expressed in compared to the low expression of *RALGAPA2* short isoform p=1.1e-7 (Figure 3.19.A).

The RFS probability result of patients with *RALGAPA2* long isoform expression was not significant with p: 0.13>0.05 (Figure 3.19.B). The long isoform was detected by 225499_at probe set however, that probe set does not recognize just one specific long isoform generated by the recognition of Hs.472285.1.2 or Hs.472285.1.3. The probe set partially recognizes the long isoform.

Given the low expression levels we detected in RALGAPA2, it is hard to conclude whether the intronic isoform exists or is functional.

CHAPTER 4

CONCLUSION

APA is a mechanism that generates 3'UTR isoform diversity which may have significant effects on protein levels and functions. APA process occurs almost synchronously with transcription and studies have revealed that proliferative signals such as estrogen (E2) might induce APA. There are possible mechanistic explanations emerging about how APA is regulated but it is not clear how these mechanisms are regulated especially in the presence of E2.

To better understand the APA mechanism in the presence of E2, we took an in-silico approach to study the effects of E2 treatment in ER+ breast cancer cells to identify whether E2 induced transcription events correlate with APA. Of great interest transcription coupled DNA breaks in response to E2 [40] was an interesting observation that we wanted to test whether it has a role in APA. For this purpose, ChIP-Seq data for γH2AX were analyzed, which is performed with ER+ breast cancer cell line MCF7, after E2 and H_2O_2 treatment. The resulting ChIP-Seq data were visualized on Integrative Genome Viewer (IGV) and different γH2AX peak ratios were observed depending on E2 treatment. Candidate genes were selected according to their γH2AX peak ratios of E2/H_2O_2 samples to study the effect of E2 induced transcription patterns on polyA site selection. Interestingly, in our candidate genes, *TMEM164* and *RALGAPA2*, we observed higher peaks of γH2AX in E2 treated samples at around poly (A) sites in compared to H_2O_2 treated and EtOH (vehicle) samples, which indicates the presence of DNA strand breaks specified by γH2AX marker in that specific region. Isoforms generated by the recognition of poly (A) sites having the highest γH2AX accumulation were confirmed with 3'RACE and RT-qPCR.

This was a very preliminary study trying to associate E2 induced transcription and dsDNA breaks with poly A site selection. Ideally, ChIP for DNA break proteins, E2 treatment, 3'RACE and RT-qPCR experiments should be done synchronously to test our hypothesis.

.

REFERENCES

[1] P. A. Krieg and D. A. Melton, “Formation of the 3’ end of histone mRNA,” *Gene*, vol. 308, no. 5955, pp. 203–206, 1999.

[2] N. Proudfoot, “Poly (A) Signals Minireview,” *Cell Press*, vol. 64, pp. 671–674, 1991.

[3] W. C. Merrick, “Mechanism and Regulation of,” *Synthesis (Stuttg).*, no. June, pp. 291–315, 1992.

[4] U. Kühn, M. Gündel, A. Knoth, Y. Kerwitz, S. Rüdel, and E. Wahle, “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,” *J. Biol. Chem.*, vol. 284, no. 34, pp. 22803–22814, 2009.

[5] R. Elkon, A. P. Ugalde, and R. Agami, “Alternative cleavage and polyadenylation: extent, regulation and function,” *Nat. Rev. Genet.*, vol. 14, no. 7, pp. 496–506, Jun. 2013.

[6] Y. Shi *et al.*, “Molecular architecture of the human pre-mRNA 3’ processing complex,” *Mol. Cell*, vol. 33, no. 3, pp. 365–376, 2009.

[7] H. H. Ağuş and A. E. Erson Bensan, “Mechanisms of mRNA polyadenylation,” *Turkish J. Biol.*, vol. 40, no. 3, pp. 529–538, 2016.

[8] A. Derti *et al.*, “A quantitative atlas of polyadenylation in five mammals.,” *Genome Res.*, vol. 22, no. 6, pp. 1173–83, Jun. 2012.

[9] N. J. Proudfoot, “Ending the message: poly (A) signals then and now,” *Genes Dev*, vol. 25, pp. 1770–1782, 2011.

[10] B. Tian, J. Hu, H. Zhang, and C. S. Lutz, “A large-scale analysis of mRNA polyadenylation of human and mouse genes,” *Nucleic Acids Res.*, vol. 33, no.

1, pp. 201–212, 2005.

[11] H. B. Akman and A. E. Erson-Bensan, “Alternative Polyadenylation and Its Impact on Cellular Processes,” *MicroRNA*, vol. 3, no. 1, pp. 2–9, 2014.

[12] A. E. Erson-Bensan and T. Can, “Alternative Polyadenylation: Another Foe in Cancer,” *Mol. Cancer Res.*, vol. 14, no. 6, pp. 507–517, 2016.

[13] N. Spies, C. B. Burge, and D. P. Bartel, “3′ UTR-Isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts,” *Genome Res.*, vol. 23, no. 12, pp. 2078–2090, 2013.

[14] H. Guo, N. T. Ingolia, J. S. Weissman, and D. P. Bartel, “Mammalian microRNAs predominantly act to decrease target mRNA levels,” *Nature*, vol. 466, no. 7308, pp. 835–840, 2010.

[15] J. Nam *et al.*, “Global analyses of the effect of different cellular contexts on microRNA targeting,” vol. 53, no. 6, pp. 1031–1043, 2015.

[16] C. Mayr and D. P. Bartel, “Widespread shortening of 3’UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.,” *Cell*, vol. 138, no. 4, pp. 673–84, Aug. 2009.

[17] B. Tian and J. L. Manley, “Alternative polyadenylation of mRNA precursors,” *Nat. Rev. Mol. Cell Biol.*, vol. 18, no. 1, pp. 18–30, 2016.

[18] Y. Lubelsky and I. Ulitsky, “Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells,” *Nature*, vol. 555, no. 7694, pp. 107–111, 2018.

[19] N. Dimitrova *et al.*, “Distinct Role of Long 3′UTR BDNF mRNA in Spine Morphology and Synaptic Plasticity in Hippocampal Neurons,” *PLoS One*, vol. 32, no. 7, pp. 736–740, 2017.

[20] B. D. Berkovits and C. Mayr, “Alternative 3’ UTRs act as scaffolds to regulate membrane protein localization.,” *Nature*, vol. 522, no. 7556, pp. 363–7, Jun.

2015.

[21] A. R. Gruber, G. Martin, W. Keller, and M. Zavolan, “Cleavage factor I m is a key regulator of 3’ UTR length,” *RNA Biol.*, vol. 9, no. 12, pp. 1405–1412, 2012.

[22] N. J. Proudfoot, A. Furger, M. J. Dye Sir, and W. Dunn, “Review Integrating mRNA Processing with Transcription,” *Cell*, vol. 108, pp. 501–512, 2002.

[23] S. B. Lemke and M. Levine, “ELAV mediates 39 UTR extension in the Drosophila nervous system,” pp. 2259–2264, 2012.

[24] N. Dimitrova *et al.*, “ELAV links paused Pol II to alternative polyadenylation in the Drosophila nervous system,” *Mol. Cell*, vol. 57, pp. 341–348, 2015.

[25] H. HUANG, H. LIU, and X. SUN, “Nucleosome Distribution near the 3′ Ends of Genes in the Human Genome,” *Biosci. Biotechnol. Biochem.*, vol. 77, no. 10, pp. 2051–2055, 2013.

[26] S. S. Skandalis *et al.*, “Cross-talk between estradiol receptor and EGFR/IGF-IR signaling pathways in estrogen-responsive breast cancers: Focus on the role and impact of proteoglycans,” *Matrix Biol.*, vol. 35, pp. 182–193, 2014.

[27] B. Nuvoli and R. Galati, “Cyclooxygenase-2, Epidermal Growth Factor Receptor, and Aromatase Signaling in Inflammation and Mesothelioma,” *Mol. Cancer Ther.*, vol. 12, no. 6, pp. 844–852, 2013.

[28] B. H. Akman, T. Can, and A. Elif Erson-Bensan, “Estrogen-induced upregulation and 3’-UTR shortening of CDC6,” *Nucleic Acids Res.*, vol. 40, no. 21, pp. 10679–10688, 2012.

[29] H. B. Akman, M. Oyken, T. Tuncer, T. Can, and A. E. Erson-Bensan, “3’UTR shortening and EGF signaling: implications for breast cancer.,” *Hum. Mol. Genet.*, vol. 24, no. 24, pp. 6910–20, Dec. 2015.

[30] J. W. Chang *et al.*, “MRNA 3′-UTR shortening is a molecular signature of

mTORC1 activation," *Nat. Commun.*, vol. 6, pp. 1–9, 2015.

[31] M. Drolet, "Growth inhibition mediated by excess negative supercoiling: The interplay between transcription elongation, R-loop formation and DNA topology," *Mol. Microbiol.*, vol. 59, no. 3, pp. 723–730, 2006.

[32] O. Tšuiko *et al.*, "A speculative outlook on embryonic aneuploidy: Can molecular pathways be involved?," *Dev. Biol.*, vol. 447, no. 1, pp. 3–13, 2019.

[33] E. P. Rogakou, C. Boon, C. Redon, and W. M. Bonner, "Megabase chromatin domains involved in DNA double-strand breaks in vivo," *J. Cell Biol.*, vol. 146, no. 5, pp. 905–915, 1999.

[34] W. Z. Tu *et al.*, "γh2AX foci formation in the absence of DNA damage: Mitotic H2AX phosphorylation is mediated by the DNA-PKcs/CHK2 pathway," *FEBS Lett.*, vol. 587, no. 21, pp. 3437–3443, 2013.

[35] K. Sugars, "The mechanism of DSB repair by the NHEJ," *Mol. Microbiol.*, no. 3, pp. 181–211, 2011.

[36] E. A. Coon and E. E. Benarroch, "DNA damage response," *Neurology*, vol. 90, no. 8, pp. 367 LP – 376, Feb. 2018.

[37] S. K. Calderwood, "A critical role for topoisomerase IIb and DNA double strand breaks in transcription," *Transcription*, vol. 7, no. 3, pp. 75–83, 2016.

[38] B. G. Ju *et al.*, "A topoisomerase IIβ-mediated dsDNA break required for regulated transcription," *Science (80-.).*, vol. 312, no. 5781, pp. 1798–1802, 2006.

[39] T. Barrett *et al.*, "NCBI GEO: Archive for functional genomics data sets-10 years on," *Nucleic Acids Res.*, vol. 39, no. SUPPL. 1, pp. 1005–1010, 2011.

[40] M. Periyasamy *et al.*, "APOBEC3B-Mediated Cytidine Deamination Is Required for Estrogen Receptor Action in Breast Cancer," *Cell Rep.*, vol. 13, no. 1, pp. 108–121, 2015.

[41] J. S. Carroll *et al.*, "Genome-wide analysis of estrogen receptor binding sites," *Nat. Genet.*, vol. 38, no. 11, pp. 1289–1297, 2006.

[42] V. Bourdeau, J. Deschênes, D. Laperrière, M. Aid, J. H. White, and S. Mader, "Mechanisms of primary and secondary estrogen target gene regulation in breast cancer cells," *Nucleic Acids Res.*, vol. 36, no. 1, pp. 76–93, 2008.

[43] H. Alotaibi, E. Ç. Yaman, E. Demirpençe, and U. H. Tazebay, "Unliganded estrogen receptor-α activates transcription of the mammary gland Na+/I-symporter gene," *Biochem. Biophys. Res. Commun.*, vol. 345, no. 4, pp. 1487–1496, 2006.

[44] S. A. Bustin *et al.*, "The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments," *Clin. Chem.*, vol. 55, no. 4, pp. 611–622, 2009.

[45] M. Liu et al., "Transcriptional profiling of Chinese medicinal formula Si-Wu-Tang on breast cancer cells reveals phytoestrogenic activity," *BMC Complement. Altern. Med.*, vol. 13, 2013.

[46] C. M. Manville et al., "Genome-wide ChIP-seq analysis of human TOP2B occupancy in MCF7 breast cancer epithelial cells," Biol. Open, vol. 4, no. 11, pp. 1436–1447, 2015.

[47] A. Honkela et al., "Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays," Proc. Natl. Acad. Sci. U. S. A., vol. 112, no. 42, pp. 13115–13120, 2015.

[48] H. Zhang et al., "PolyA_DB: A database for mammalian mRNA polyadenylation," *Nucleic Acids Research*, vol. 33, pp. 116–120, 2005.

[49] Zhang, Yong et al. "Model-based analysis of ChIP-Seq (MACS)," *Genome biology*, vol. 9, no. 9, pp. 137-143, 2008.

[50] Goel, Manish Kumar et al. “Understanding survival analysis: Kaplan-Meier estimate.” *International journal of Ayurveda research*, vol. 1, no.4, pp. 274-278, 2010.

APPENDICES

A. DATASETS

Table A.1. Experiments of GSE11324 Dataset

GSM number	Sample Name
GSM286031	0hr 1
GSM286032	0hr 2
GSM286033	0hr 3
GSM286034	3hr 1
GSM286035	3hr 2
GSM286036	3hr 3
GSM286037	6hr 1
GSM286038	6hr 2
GSM286039	6hr 3
GSM286040	12hr 1_redo2
GSM286041	12h 2
GSM286042	12h 3

Table A.2. Experiments of GSE8597 Dataset

GSM number	Sample Name
GSM213326	MCF7_E2_24h_rep1
GSM213327	MCF7_E2_24h_rep2
GSM213328	MCF7_E2_24h_rep3
GSM213329	MCF7_E2_24h_rep4
GSM213330	MCF7_EtOH_24h_rep1
GSM213331	MCF7_EtOH_24h_rep2
GSM213332	MCF7_EtOH_24h_rep3
GSM213333	MCF7_EtOH_24h_rep4

B. PRIMERS

Table B.1. PCR Primers and RT-Qpcr Primers

Primer Name	Primer Sequence (5’ to 3’)	Experiment
GAPDH_F	GGGAGCCAAAAGGGTCATCA	PCR
GAPDH_R	TTTCTAGACGGCAGGTCAGGT	PCR
TFF1_F	CCATGGAGAACAAGGTGATCTGC	PCR
TFF1_R	GTCAATCTGTGTTGTGAGCCGAG	PCR
RPLP0_F	GGAGAAACTGCTGCCTCATA	RT-qPCR
RPLP0_R	GGAAAAAGGAGGTCTTCTCG	RT-qPCR
TFF1_F	TTGTGGTTTTCCTGGTGTCA	RT-qPCR
TFF1_R	CCGAGCTCTGGGACTAATCA	RT-qPCR
RALGAPA2_Short_F	GACCTGCCTCTGCTGTCATT	RT-qPCR
RALGAPA2_Short_R	GATGAGGTGAGTGTGGGTGG	RT-qPCR
RALGAPA2_Long_F	GCCAGACTCACTCTTGGGAC	RT-qPCR
RALGAPA2_Long_R	TTTGGGGCACCCTCATTCTC	RT-qPCR
TMEM164_Short_F	TGGTAAACACTCGGCTGCTC	RT-qPCR
TMEM164_Short_R	CTGAGGGGCTCTGGAGTGTA	RT-qPCR
TMEM164_Long_F	TCTTTGAAGGCAGGGCCAAA	RT-qPCR
TMEM164_Long_R	TGTAGCAGTTTGACGGTGGG	RT-qPCR

Table B.2. 3’RACE Primers

Primer Name	Primer Sequence (5’ to 3’)
RACE_OligodT	TGACCACGCGTATCGATGTCGACTTTTTTTT TTTTTTTV
Anchor_R	GACCACGCGTATCGATGTCGAC
TMEM164_3'RACE_Forward 1	TACACTCCAGAGCCCCTCAG
TMEM164_3'RACE_Forward 2	GGGAAGCTGGTCATCCTGTT
RALGAPA2 3'RACE Forward 2	GACCTGCCTCTGCTGTCATT

C. DNA CONTAMINATION AND RNA EXAMINATON

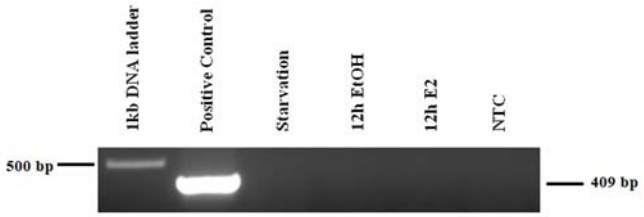


Figure C.1. Absence of DNA Contamination.

1kb DNA ladder is used as marker. MCF7 cDNA is used as positive control. NTC (no template control) lacking template. Cleaning of RNA samples are controlled by PCR performed with GAPDH specific primers. Expected size is 409 bp length.

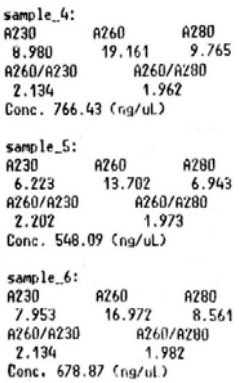
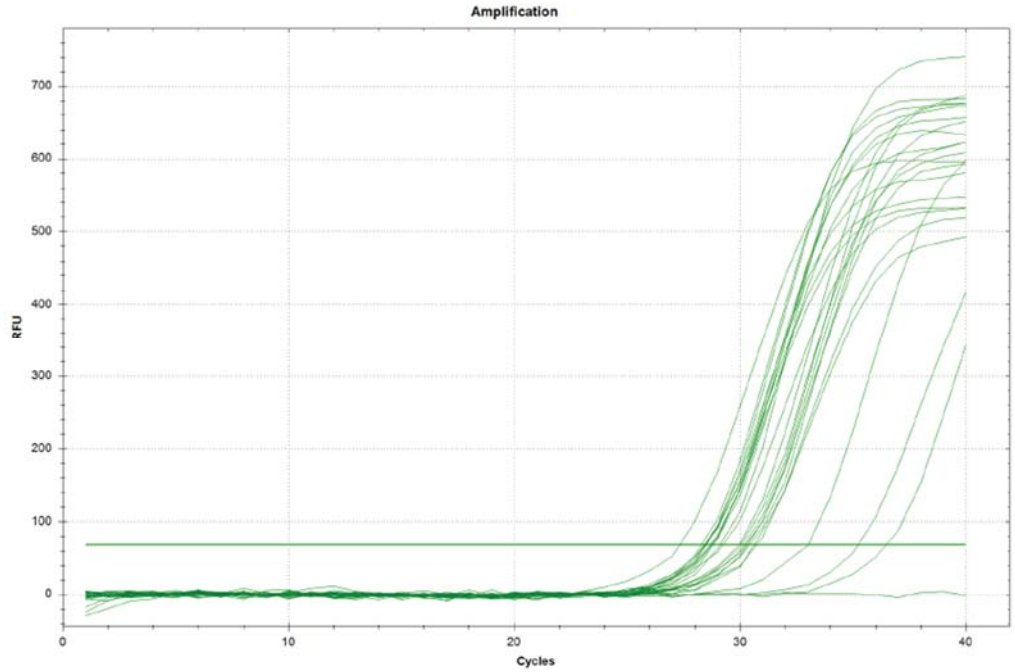
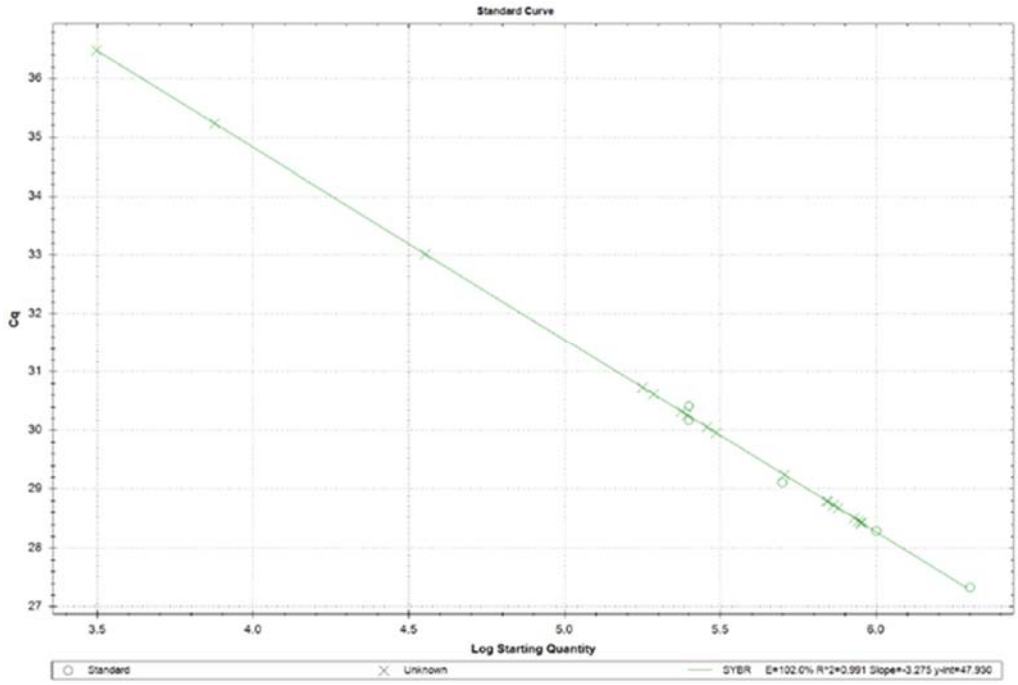


Figure C.2. Concentrations of RNA Samples After DNase treatment

MaestroNano Spectrophotometer was used to measure the concentration and purity of RNA samples. Sample_4 refers to Starvation, sample_5 refers to 12h EtOH, sample_6 refers to 12h E2 treatment RNA samples. A260/A280 and A260/A230 ratios were in between 1.8-2.2 range showing their purity.

D. QRT- PCR REPORTS





Quantification Data

Well	Fluor	Target	Content	Sample	Cq	Cq Mean	Cq Std. Dev	Starting Quantity (SQ)	Log Starting Quantity	SQ Mean	SQ Std. Dev
D10	SYBR		Unkn	045E2 RS	30.27	30.27	0.000	2.474E+05	5.393	2.47E+05	0.00E+00
E10	SYBR		Unkn	045E2 RS	29.96	29.96	0.000	3.061E+05	5.486	3.06E+05	0.00E+00
F10	SYBR		Unkn	045E2 RS	30.06	30.06	0.000	2.864E+05	5.457	2.86E+05	0.00E+00
A10	SYBR		Unkn	045ETOH RS	28.46	28.46	0.000	8.792E+05	5.944	8.79E+05	0.00E+00
B10	SYBR		Unkn	045ETOH RS	28.73	28.73	0.000	7.292E+05	5.863	7.29E+05	0.00E+00
C10	SYBR		Unkn	045ETOH RS	28.51	28.51	0.000	8.522E+05	5.931	8.52E+05	0.00E+00
B11	SYBR		Unkn	3E2 RS	30.33	30.33	0.000	2.370E+05	5.375	2.37E+05	0.00E+00
C11	SYBR		Unkn	3E2 RS	30.74	30.74	0.000	1.778E+05	5.250	1.78E+05	0.00E+00
D11	SYBR		Unkn	3E2 RS	30.62	30.62	0.000	1.927E+05	5.285	1.93E+05	0.00E+00
A11	SYBR		Unkn	3ETOH RS	33.02	33.02	0.000	3.568E+04	4.552	3.57E+04	0.00E+00
G10	SYBR		Unkn	3ETOH RS	36.48	36.48	0.000	3.139E+03	3.497	3.14E+03	0.00E+00
H10	SYBR		Unkn	3ETOH RS	35.23	35.23	0.000	7.521E+03	3.876	7.52E+03	0.00E+00
C12	SYBR		NTC		N/A	0.00	0.000	N/A	N/A	0.00E+00	0.00E+00
E09	SYBR		Std	1:10	29.11	29.11	0.000	5.000E+05	5.699	5.00E+05	0.00E+00
B09	SYBR		Std	1:2,5	27.33	27.33	0.000	2.000E+06	6.301	2.00E+06	0.00E+00
G09	SYBR		Std	1:20	30.18	30.18	0.000	2.500E+05	5.398	2.50E+05	0.00E+00
H09	SYBR		Std	1:20	30.42	30.42	0.000	2.500E+05	5.398	2.50E+05	0.00E+00
C09	SYBR		Std	1:5	28.30	28.30	0.000	1.000E+06	6.000	1.00E+06	0.00E+00
A12	SYBR		Unkn	12E2 RS	28.80	28.80	0.000	6.936E+05	5.841	6.94E+05	0.00E+00
B12	SYBR		Unkn	12E2 RS	28.79	28.79	0.000	6.997E+05	5.845	7.00E+05	0.00E+00
H11	SYBR		Unkn	12E2 RS	29.24	29.24	0.000	5.075E+05	5.705	5.07E+05	0.00E+00
E11	SYBR		Unkn	12ETOH RS	28.44	28.44	0.000	8.956E+05	5.952	8.96E+05	0.00E+00
F11	SYBR		Unkn	12ETOH RS	28.43	28.43	0.000	8.993E+05	5.954	8.99E+05	0.00E+00
G11	SYBR		Unkn	12ETOH RS	28.68	28.68	0.000	7.564E+05	5.879	7.56E+05	0.00E+00

Melt Curve

Step #: 7

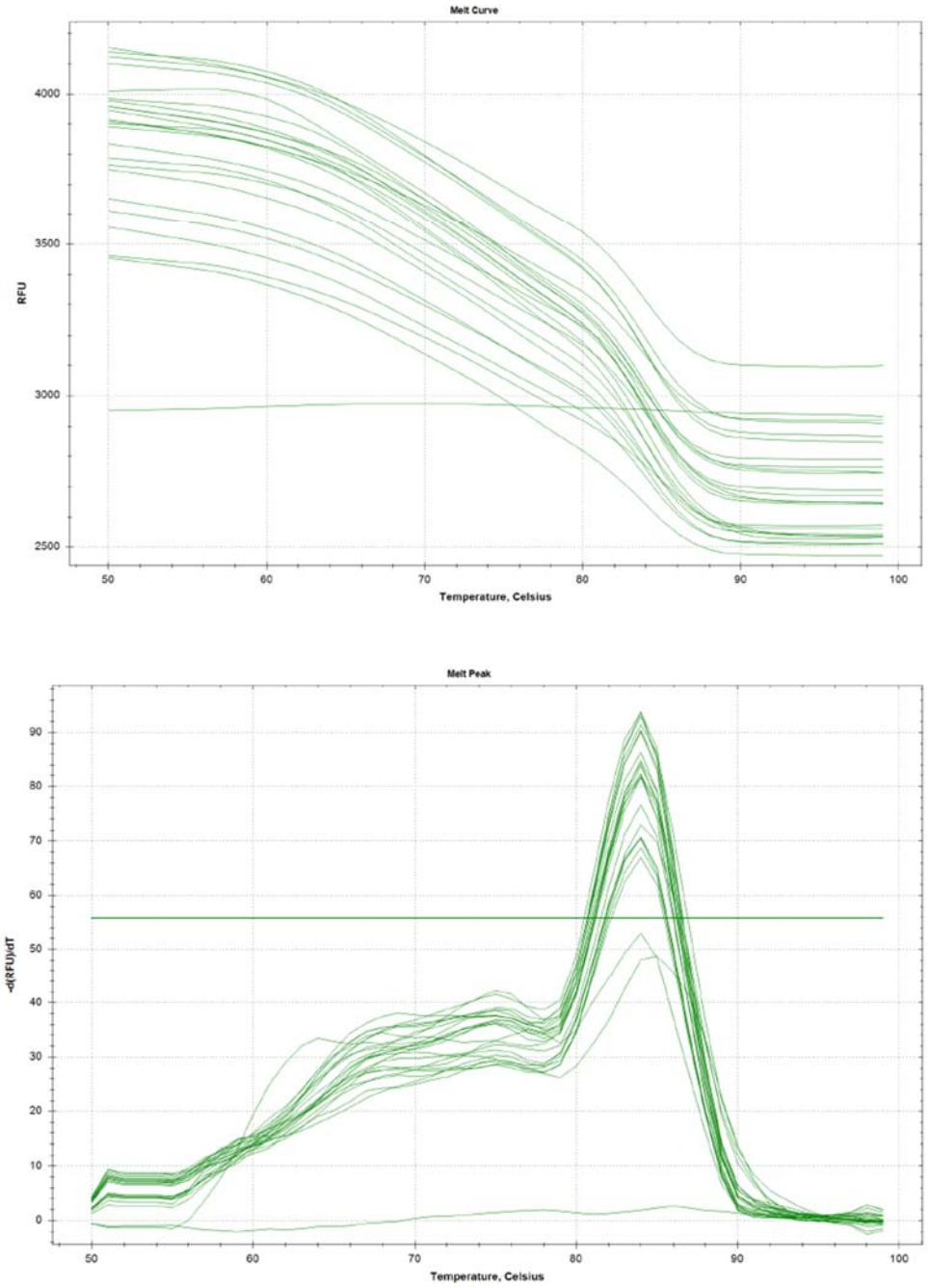


Figure D.1. *RALGAPA2* Short Primers RT-qPCR Assay Report containing Amplification, Standard Curve, Quantification Data, Melt Curve and Melt Peak

E. GENE DIAGRAMS

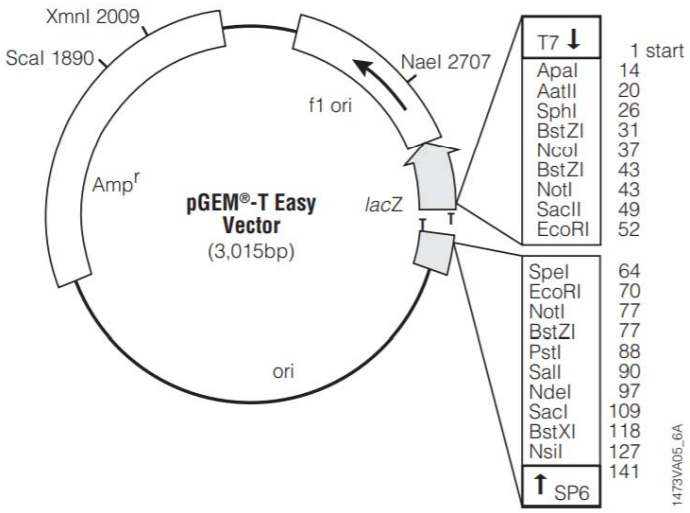


Figure E.1. PGEM-T Easy Vector

F. MARKERS

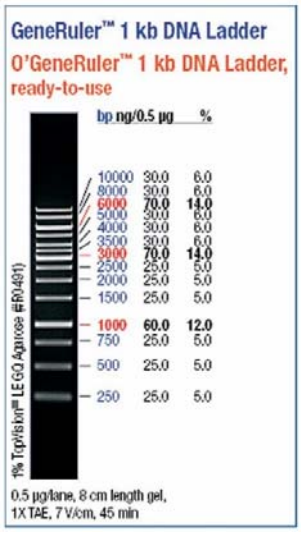


Figure F.1. 1kb DNA Ladder

Cat # FERSM1163, Thermo Scientific

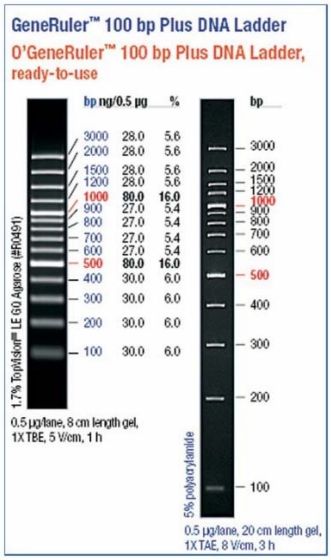


Figure F.2. 100 bp Plus DNA Ladder

Cat # 10364280, Thermo Scientific