UTILIZATION OF TRANSCRIPT VARIANTS OF *CXXC5* GENE TO PREDICT ITS PROMOTER

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

BY

KERİM YAVUZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
MOLECULER BIOLOGY AND GENETICS

AUGUST 2020

Approval of the thesis:

UTILIZATION OF TRANSCRIPT VARIANTS OF *CXXC5* GENE TO PREDICT ITS PROMOTER

submitted by **KERİM YAVUZ** in partial fulfillment of the requirements for the degree of **Master of Science** i**n Molecular Biology and Genetics, Middle East Technical University** by,

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

Prof. Dr. Ayşe Gül Gözen
Head of the Department, **Biological Sciences** _____

Prof. Dr. Mesut Muyan
Supervisor, **Biological Sciences, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Erkan Kiriş
Biological Sciences, METU _____

Prof. Dr. Mesut Muyan
Biological Sciences, METU _____

Asst. Prof. Dr. Murat Alper Cevher
Molecular Biology and Genetics, Bilkent University _____

Date: 14.08.2020

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

Name Last name : Kerim Yavuz

Signature :

ABSTRACT

UTILIZATION OF TANSCRIPT VARIANTS OF *CXXC5* GENE TO PREDICT ITS PROMOTER

Yavuz, Kerim
Master of Science, Molecular Biology and Genetics
Supervisor: Prof. Dr. Mesut Muyan

August 2020, 77 pages

CXXC5 is an estrogen responsive gene whose deregulations are implicated in a number of diseases including cancers. It is crucial to understand the mechanisms through which *CXXC5* expression is regulated for the characterization its role in physiologies and pathophysiologies. To gain an insight in the *CXXC5* gene expression, we wanted to identify its promoter. We hypothesized that the promoter of *CXXC5* could reside in close proximity to the transcription start site of its main transcript variant. In this study, we wanted to identify the main transcript(s) of *CXXC5* and its transcription start site(s). *CXXC5* has 14 transcript variants all of which encode the same protein. We found here that *CXXC5* Transcript Variant 2 is the variant that has the highest expression levels in both MCF7 and HL60 cells, thus, it is the main transcript variant. *CXXC5* Transcript Variant 2 has a broad transcription initiation region and by investigating the ranges of this region, we identified a novel transcript variant of *CXXC5* as well.

Keywords: *CXXC5,* Transcript Variant, Promoter, Gene Regulation

ÖZ

CXXC5 GENİNİN TRANSKRİPT VARYANTLARININ PROMOTÖR BÖLGESİNİN TAHMİNİ İÇİN KULLANILMASI

Yavuz, Kerim
Yüksek Lisans, Moleküler Biyoloji ve Genetik
Tez Yöneticisi: Prof. Dr. Mesut Muyan

Ağustos 2020, 77 sayfa

CXXC5, deregülasyonunun kanser de dahil olmak üzere farklı hastalıklarda etkisinin olduğunun gösterildiği östrojen yanıtlı bir gendir. *CXXC5*'in fizyolojik ve patofizyolojik durumlardaki rolünü anlamak için gen ifadesinin nasıl düzenlendiğini anlamak elzemdir. *CXXC5* gen ifadesine yönelik fikir edinmek için promotör bölgesini tanımlamak istedik. *CXXC5* promotörünün ana transkript varyant(lar)ının transkripsiyon başlangıç yer(ler)inin yakınında olabileceğini tahmin ettik. Bu çalışmada, *CXXC5* ana transkript varyantını ve bu varyantın transkripsiyon başlangıç yerini belirlemek istedik. *CXXC5*'in hepsi aynı proteini kodlayan 14 tane transkript varyantı vardır. *CXXC5* Transkript Varyant 2'nin MCF7 ve HL60 hücrelerinde en fazla ifade seviyesine sahip olduğunu, bu sebeple ana transkript varyantı olduğunu bulguladık. *CXXC5* Transkript Varyant 2'nin geniş bir transkripsiyon başlangıç bölgesine sahip olduğunu ve bu bölgeden daha önce tanımlanmamış bir *CXXC5* transkript varyantının sentezlendiğini gösterdik.

Anahtar Kelimeler: *CXXC5,* Transkript Varyantı, Promotör, Gen İfadesi

To all resisting souls

ACKNOWLEDGMENTS

I would like to express my sincerest gratitudes to my supervisor Prof. Dr. Mesut Muyan for his guidance and support in my initial steps into academia. I've learned priceless things and skills under his supervision and I am greatly thankful for his support throughout my Master's journey.

I would like to thank my thesis committee members Assoc. Prof. Dr. Erkan Kiriş and Asst. Prof. Dr. Murat Cevher for their kind interest and feedbacks.

I would like to thank my laboratory mates for their help, support and friendship. I am thankful for Çağla Ece Olgun, Gizem Kars, Gizem Turan and Öykü Deniz Demiralay. I would also like to thank my former laboratory mates, Gamze Ayaz, Negin Razizadeh and Burcu Karakaya. I cannot thank Pelin Yaşar enough for her help.

I would like to thank my family for being by my side at all times.

I would like to thank ODTU-BAP for supporting this study through YLT-108-2018-3754 and TUBİTAK for supporting this study through 117Z213 and 118Z957.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

1.1 Estrogen-Estrogen Receptor Signaling

As members of the steroid hormone family, estrogens play important roles in processes such as regulation of reproductive cycle and differentiation of tissues and organs [1]. 17β-estradiol (E2) is the main circulating estrogen hormone [2].

E2 exerts its cellular effects through estrogen receptors α and β (ERα and β) that function as transcription factors. ERα and β are members of the nuclear hormone receptor family and like the other nuclear hormone receptors; they are modular proteins having distinct functional regions. ERα and β comprise of six regions named as A/B, C, D and E/F. The A/B region is localized in the N-terminus of the proteins and exhibits a ligand-dependent transactivation function, also called activation function 1 (AF1). In addition, the A/B region is involved in the interaction of ERs with coregulatory proteins that are important in ER-mediated transcription. The C region is the DNA Binding Domain (DBD) and contains two zinc-binding motifs with two perpendicular α-helices important in the binding of estrogen receptors to DNA. The D region of ERα and β is the hinge region that acts as a linker between DBD and the carboxyl-terminus. D region not only provides flexibility to the protein but also contains the nuclear localization signal. E/F region is a multifunctional domain involved in ligand binding, receptor dimerization and ligand-dependent transactivation (activation function 2, AF2) (Figure 1.1) [3].

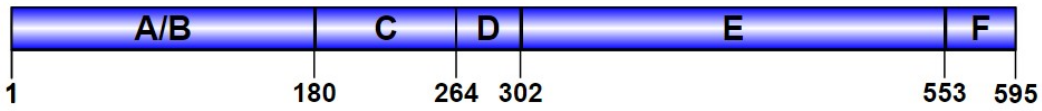


Figure 1.1. **Schematic representation of the Estrogen Receptor domain structure.**

Although ER localizes primarily to the nucleus independently of ligand, fractions of ER also observed to localize plasma membrane, cytoplasm and mitochondria from which ER in response to E2 could mediate the so-called the non-genomic signaling. The nuclearly localized ER on the other hand is a transcription factor. Upon binding to E2, ER regulates gene expressions through estrogen response element (ERE)-dependent or independent pathways. The ERE-dependent pathway involves the binding of the E2-ER complex to specific sequences on DNA called as estrogen response elements, EREs. EREs are 13-nucleotide-long sequences composed of two palindromic half-sites separated by three linker nucleotides. The consensus ERE is denoted as 5'- GGTCAnnnTGACC -3', although EREs could deviate from the consensus up to three-nucleotides [4].

Apart from the ERE-dependent signaling pathway, the E2-ER complex can also mediate transcription through interaction with transcription regulating proteins that are bound to their cognate response elements. This type of E2-ER signaling is referred to as the ERE-independent signaling pathway [5].

Estrogen signaling is shown to play critical roles in pathophysiology of many tissues including breast tissue. Breast cancer is characterized by extensive cellular proliferation and deregulation of integrated signaling pathways important in cell proliferation and cell death [6]. The control over the progression of the cell cycle is lost in cancers and aberrant expressions of cell cycle-related proteins such as Cyclin D1 that are responsive to E2 signal are thought to contribute to the progression of ER-positive breast cancer [7].

1.2 CXXC-Type Zinc Finger Protein 5 (CXXC5)

1.2.1 CXXC-Type Zinc Finger Family of Proteins

CXXC5 is a CXXC-type zinc finger family (ZF-CXXC) protein encoded by the *CXXC5* gene. The ZF-CXXC family consists of KDM2A, KDM2B (Lysine Specific Demethylases 2A and 2B), KMT2A, KMT2B (Lysine Methyltransferases 2A and 2B), DNMT1 (DNA Cytosine-5-Methyltrasnferase 1), TET1, TET3 (Tet Methylcytosine Deoxygenases 1 and 3), IDAX (Inhibition of the Dvl and Axin Complex) and CXXC5. This family of proteins contain a highly conserved ZF-CXXC domain. This domain is 50-70 amino-acid in length and contains 2 consecutive CXXCXXC consensus sequences that are separated by a sequence of amino acids that act as linkers and vary in length. Due to this domain, ZF-CXXC family proteins can bind to DNA [8]. It has been shown through high-resolution crystallography studies of CXXC domains of ZF-CXXC family proteins that these proteins bind to unmethylated CpG dinucleotide containing DNA sequences [9]. The ZF-CXXC family proteins affect DNA methylation through binding to unmethylated CpG dinucleotides and through mediating the demethylation of methylated CpG dinucleotides. Since the access of ZF-CXXC family proteins to major and minor grooves of DNA is required for these proteins to bind to DNA, it is thought that ZF-CXXC family proteins, such as KDM2A, bind to nucleosome-free regions of DNA and the linker regions between nucleosomes to prevent *de novo* DNA methylation [10]. Also, TET proteins have been shown to take part in the demethylation of Cytosines due to their deoxygenase activity [11]. Apart from affecting the methylation status of DNA, these proteins have been shown to modify the epigenetic environment and regulate gene expression through interaction with histone modifying proteins [12].

1.2.2 CXXC5 Protein

CXXC5 is a protein synthesized from the *CXXC5* gene located on the long arm of chromosome 5 in the human genome. CXXC5 protein consists of 322 amino acids and contains a highly conserved CXXC domain between its 256th and 297th amino acids [13]. Due to the nuclear localization signal located in the highly conserved CXXC domain between 257th and 262nd amino acids, CXXC5 is primarily localized in the nucleus (Figure 1.2).

Figure 1.2. **Schematic representation of CXXC5's protein structure.** The CXXC domain located between 256th and 297th amino acids is shown as a yellow rectangle and a red sphere between 257th and 262nd amino acids represents the nuclear localization signal (NLS).

1.2.3 The *CXXC5* Gene and It's Regulation

CXXC5 gene is located on the long arm of chromosome 5, in the 5q31.2 region. The gene locus is approximately 35 kilobase pair (kbp) long according to the *Homo sapiens* (human) genome assembly GRCh38. The gene contains 11 exons and 10 introns. *CXXC5* gene has 14 transcript variants, identified by the Expressed Sequence Tag method which is widely utilized to identify alternative splicing products of genes [14], [15]. The last two exons on the gene locus are found in all 14 transcript variants and contain the coding region. The coding region is 969 basepairs (bp) long and divided into Exons 10 and 11 by 924 bp and 45 bp respectively (Figure 1.3) [15]. All 14 transcript variants of *CXXC5* encode the same protein with an estimated molecular mass (MM) of 33 kDa.

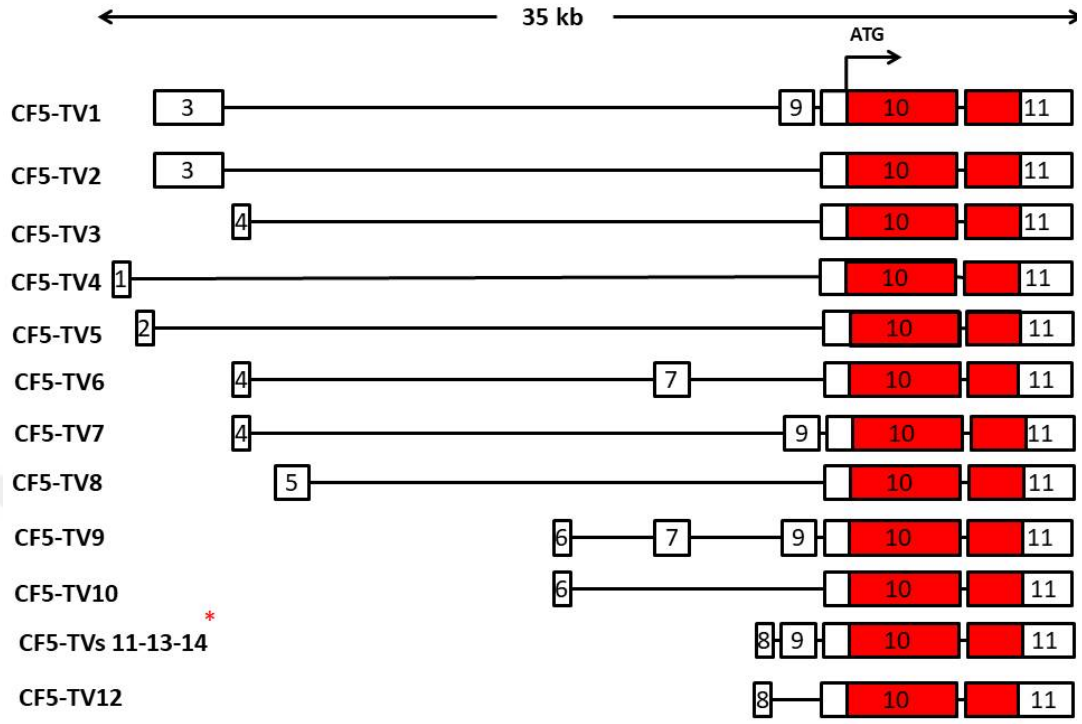


Figure 1.3. **Genomic representation of 14 transcript variants of the *CXXC5* gene.** *CXXC5* gene has 14 transcript variants encoded from the approximately 35 kbp long genomic locus. *CXXC5* gene has 11 exons which are represented by rectangles and 10 introns represented by horizontal lines. All transcript variants encode the same protein from their shared coding region found in Exons 10 and 11. *: *CXXC5* Transcript Variants 11, 13, and 14 differ in their noncoding exons by 6-12 base pairs.

Although studies on *CXXC5* gene and its protein product are limited, its expression is regulated by a number of signaling pathways. CXXC5 was initially discovered as a retinoid inducible factor (RINF) in NB4 human acute promyelocytic leukemia cell line. It was shown in NB4 cell line that *CXXC5* is a retinoic acid-responsive gene, whose expression is induced in response to all-*trans* retinoic acid treatment as early as 90 minutes. It has also been shown that the *CXXC5* gene is a target of Retinoic Acid Receptor alpha (RARα). A Retinoic Acid Response Element (RARE) of Direct Repeat 2 type (a motif of two direct repeats separated by two nucleotides) is located at 2123 bp upstream of the putative *CXXC5* gene promoter

(5’- GG**AGTTCA**TG**AGGTGA**GC -3’) that differ from the well-established Direct Repat 5 (a motif of two direct repeats separated by five nucleotides) type RARE located at the Retinoic Acid Receptor beta (RARβ2) gene promoter (5’- **GGTTCA**CCGAA**AGTTCA** -3’) by 1 nucleotide in its second direct repeat motif (Figure 1.4) [16], [17].

CXXC5 has also been shown to be regulated by Bone Morphogenic Protein 4 (BMP4) in neural stem cells (NSCs). BMP4 is a protein necessary for differentiation and regionalization of the telencephalic forebrain and exerts its effects through Smad proteins that act as transcription factors either directly by binding to DNA or indirectly by interacting with other transcription factors associated with their cognate response elements. It has been shown in telencephalic NSCs that BMP4 stimulation increases *CXXC5* expression [18].

Our previous microarray studies suggest that *CXXC5* is an E2 responsive gene [5]. In our laboratory, it has been shown through Quantitative Polymerase Chain Reaction that E2 augments *CXXC5* expression. This augmentation is mediated through interaction of ERα with an Estrogen Response Element (ERE) located at 242 bp upstream of Translation Start Site (TrSS) of *CXXC5* gene. This ERE differs from the consensus ERE by one nucleotide (5’- GGTCAggaTGAC<u>A</u> 3’) and it has been shown to have a regulatory role in gene expression through luciferase assay. The interaction of ERα with the aforementioned ERE has been shown by Electrophoretic Mobility Shift Assay and Chromatin Immunoprecipitation Figure 1.4) [2].

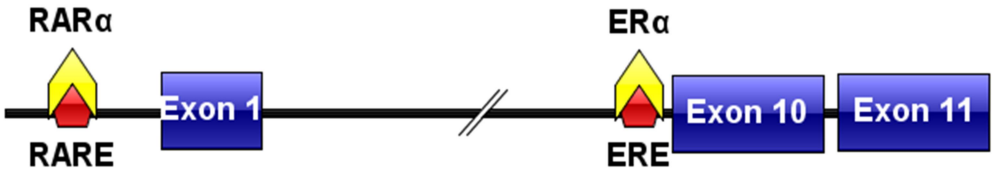


Figure 1.4. **Genomic representation of regulatory regions of the *CXXC5* gene.** *CXXC5* is a retinoic acid-responsive gene and a direct target of Retinoic Acid Receptor alpha (RARα) through a retinoic acid response element (RARE) located at 2123 bp upstream of the gene locus. *CXXC5* is an estrogen-responsive gene whose Expression is mediated by through interaction of Estrogen Receptor alpha (ERα) with an Estrogen Response Element (ERE) located at 242 bp upstream of Translation Start Site (TrSS) of *CXXC5* gene.

1.2.4 CXXC5 in Health and Disease

The *CXXC5* gene is located at the long arm of chromosome 5q31.2.. This region is commonly reported to undergo deletions and microduplications. The deletions in this region are implicated in various pathophysiologies such as Myelodysplastic Syndromes (MDS) and Acute Myeloid Leukemia (AML) as well as intellectual disabilities [19]–[21]. It has also been proposed that a 10.43 megabase pair (Mb) deletion in this region is associated with delayed psychomotor development [22]. All these findings make the genes found in this region potential candidates in the diagnosis and targeting of these pathologies.

Although the studies on the function of CXXC5 are limited, it has been proposed that CXXC5 is a transcription factor and/or epigenetic modulator contributing to cell proliferation, differentiation, and cell death.

CXXC5 is primarily localized to the nucleus in the cells and it can also be localized in the cytoplasm by interacting with Dishevelled proteins [23]. Dishevelled (Dvl) proteins are important scaffolds in the transmission of Wnt signal and it has been proposed that CXXC5 acts as a negative regulator of this signaling pathway through a negative feedback loop. As a result of the Wnt signaling pathway, β-catenin acts as a transcription factor to upregulate *CXXC5* expression and in return, CXXC5 interacts with the Dvl protein in the cytoplasm to inhibit the Wnt signal.

Since Wnt/β-catenin signaling pathway is important in wound healing and osteoblastogenesis, this suggests that CXXC5 is an important regulator in developmental processes and differentiation [23]–[25]. Dvl-CXXC5 interactions have recently been proposed to be a potential target for osteoporosis and small-molecule inhibition as a potential therapeutic approach [25]. Since Wnt/β-catenin signaling pathway is important in kidney development and disruptions of this pathway could be associated with tumor progression, CXXC5 could act as a tumor suppressor protein [26].

Tumor suppressor function of CXXC5 has also been implicated in hepatocellular carcinoma and tumors of peripheral nerve sheet. *CXXC5* is a target gene of the TGFβ signaling pathway and CXXC5 protein regulates this apoptotic pathway by positive feedback [27]–[29]. CXXC5 is also involved in the apoptosis of HEK293 cells in which it induces apoptotic pathways activated by TNFα [30].

CXXC5 has a role in the differentiation of neurons. The Wnt/β-catenin signaling pathway upregulates CXXC5, which activates myelin gene expression during the process of oligodendrocyte differentiation [31]. The Regulatory role of CXXC5 on myelin genes is not only implicated in brain development, but also skeletal myogenesis.

CXXC5 is involved in angiogenesis and endothelial cell differentiation by upregulating the expression of Vascular Endothelial Growth Factor Receptors (VEGFRs), as part of the Bone Morphogenic Factor Pathway (BMP) [32].

CXXC5 takes part in cellular metabolism. CXXC5 binds to an Oxygen Response Element (ORE) located in the promoter region of *COX4I2* gene which encodes a subunit of Cytochrome C Oxidase enzyme [33]. Cytochrome C Oxidase is an enzyme and a part of the Electron Transport Chain important in energy production through oxidative phosphorylation. Cytochrome C Oxidase accepts an electron from cytochrome c proteins and transfers it to oxygen (O_2) which is the final electron acceptor in the Chain. As a result of this reaction, water is formed [34]. It has been reported that in the presence of high oxygen levels, CXXC5 binds to the

ORE in the promoter of the *COX4I2* gene and acts as a repressor that downregulates the expression of the enzyme subunit. When the oxygen levels are low, CXXC5 dissociates and the COX4I2 expression is upregulated [34]. These findings suggest that CXXC5 is involved in the cellular response to varying levels of oxygen.

CXXC5 is also involved in immune responses. In plasmacytoid dendritic cells, CXXC5 and TET2 regulate the expression of Interferon Regulatory Factor 7 (IRF7) [35]. IRF7 is a transcription factor of the Interferon Regulatory Factor family proteins. In response to pathogenic nucleic acid sensing, Toll-like Receptors 7-9 activate IRF7, together with NFκB and AP-1 and these transcription factors take part in the production of IFNα/β [36]. CXXC5 and TET2 have been shown to keep the *IRF7* gene promoter in a hypomethylated state [35]. In addition to regulating *IRF7* expression, CXXC5 takes part in immune responses by inhibiting the expression of CD4+ helper T cell marker CD40 ligand (CD40L) in CD8+ cytotoxic T cells by inducing histone 3 lysine 9 methylation (H3K4me) in the promoter of *Cd40lg* gene through interaction with chromatin modifier SUV39H1 [37].

As previously mentioned, CXXC5 is suggested to have tumor suppressor roles. It is also involved in the activation of another well-known and widely studied tumor suppressor gene, p53, upon DNA damage [38]. CXXC5 is suggested to interact with ataxia-telangiectasia (A-T) mutated (ATM) protein which is the "master regulator" of DNA damage response. In response to DNA damage, ATM is recruited to the site of damage and phosphorylates p53 and Chk2 that induce cell cycle arrest [39]. CXXC5 has been suggested to interact with ATM in the process of p53 phosphorylation [38].

CXXC5 is also implicated in of the initiation and/or progression of various cancers. In Acute Myeloid Leukemia (AML), 5q31.2 chromosomal region that contains CXXC5 together with several tumor suppressor genes is frequently found to be deleted in AML patients. The retinoic acid receptor alpha (*RARα*) gene

translocation is used as a marker in Acute Promyelocytic Leukemia (APL), in which CXXC5 is essential for cellular maturation of promyelocytic leukemia cells [16]. In ERα+ breast cancer (breast cancer types that express estrogen receptor alpha), CXXC5 is identified as a poor prognostic factor since its overexpression is associated with lower survival rates of ERα+ breast cancer patients [40].

1.3 Regulation of Transcription from Promoters

The genetic code carried in and transmitted with DNA determines the functional and morphological features of cells. The development of complex organisms with different cell types from single cells and the homeostasis of living cells is dependent on strict control of gene expression in a spatiotemporal manner [41]. The genetic code comprises coding sequences with their genetic and epigenetic properties as well as the non-coding regulatory elements important in the spatiotemporal control of gene expression [42].

1.3.1 RNA Polymerase II Mediated Transcription from Promoters

In eukaryotic cells, RNA Polymerase II enzyme synthesizes mRNAs from protein-coding genes as well as long noncoding RNAs (lncRNAs), small nucleolar RNAs (snRNAs) and micro RNAs (miRNAs) [43]. Promoters are defined as DNA regions that contain the *cis*-acting elements which are important in the correct and efficient initiation of RNA Polymerase II-mediated transcription from Transcription Start Sites (TSSs) in response to various signaling pathways in a spatiotemporal manner [44]. Although there are reports of alternative promoter regions located distally from genes, typically, transcription is initiated from a defined region at which the general transcription factors (GTFs) bind. This region is called the core promoter and generally encompasses the 50 base pairs upstream and/or downstream of the TSS [41]. This makes TSSs important identifiers of core promoter regions. Core promoters by themselves have a basal level of transcriptional activity which can be

increased or decreased by proximal and/or distal regulatory regions such as enhancers, silencers, insulators, etc. [45]. These regions bind and recruit transcription factors to modulate gene expression.

1.3.2 Types of Core Promoters

Core promoters have different characteristics that enable them to be classified into different groups. These characteristics include their patterns of transcription initiation, the sequences and motifs that they contain, and chromatin configuration [46]. Generally, two major classes of promoters are defined: TATA-containing promoters and CpG Island (CGI) promoters [47].

1.3.2.1 TATA-Containing Promoters

TATA-box is a *cis*-acting element generally located at 25-30 bp upstream of TSS with a motif of 5'- TATAAA -3' [48]. This element is involved in the formation of stable transcription machinery in the gene promoter.

The recruitment of RNA Polymerase II to TATA-containing promoters is mediated by basal machinery called preinitiation complex (PIC) that is formed at the initiation of transcription. The initial component of this machinery is TATA-Binding Protein (TBP) which is a subunit of a multisubunit TFIID complex. TBP interacts with TATA-box and defines the promoter. The second member of PIC that is recruited to the promoter is TFIIA which stabilizes the binding of TFIID to DNA whose activity is not essential for the formation of the transcription machinery [49]. TFIIB interacts with TBP to assist DNA recognition and also it has a role in TSS selection. Binding of TFIIB is followed by an additional member of the PIC, which is TFIIF. TFIIF is the factor that recruits RNA Polymerase II to the

promoter and some evidence suggest that it also has a role in stabilizing TFIIB [50]. The last two elements of PIC are TFIIE and TFIIH. TFIIE recruits the multifunctional TFIIH. TFIIH is a complex with several enzymatic activities including helicase activity which facilitates the unwinding of DNA and melting of promoter at the initiation of transcription. Also, TFIIH has a kinase activity due to its Cdk7 component [51]. The kinase activity of Cdk7 is important in the activation of RNA Polymerase II.

RNA Polymerase II contains 12 subunits, the largest of which is the Rpb1 with a unique C-Terminal Domain (CTD). CTD has tandem heptad repeats of a consensus amino acid sequence which is Tyrosine-Serine-Proline-Threonine-Serine-Proline-Serine ($Y^1S^2P^3T^4S^5P^6S^7$). RNA Polymerase II is activated at the initiation of transcription by the kinase activity of TFIIH's Cdk7 component. Cdk7 phosphorylates the CTD of RNA Polymerase II at the Serine residue located at the 5^{th} position of the heptad repeat. The phosphorylated RNA Polymerase II (namely, Ser5-Phosphorylated RNA Pol II) marks the transcription initiation site, and antibodies specific for Ser5 –phosphorylation are used to locate RNA Polymerase II at the initiation of transcription. The end of transcription is marked by phosphorylation of Serine residue in the second position of the heptad repeat [52].

1.3.2.2 CpG Island (CGI) Containing Promoters

In eukaryotic cells, CpG dinucleotides can undergo methylation through the addition of a methyl group to cytosine. Regions of the increased amount of CpG dinucleotides and G + C base pairs are dispersed throughout the genome and they are identified as CpG islands [53]. CpG islands are generally hypomethylated and characterized by several features such as being larger than 500 bp long, having a G + C content of more than 55% with an observed-to-expected CpG dinucleotide frequency of 0.6 [54], [55].

Transcription initiation regions of 60-70% of the genes are associated with CpG islands [55]. These genes with CpG island containing promoters include housekeeping genes that are constitutively expressed in all cell lines as well as genes with tissue- and developmental stage-specific expression pattern [56], [57].

CpG islands generally lack the TATA-box and properties of TATA-containing promoters with distinct characteristics associated with CpG island containing promoters. Unlike the TATA-containing promoters that have a strict site of transcription initiation, CGI-containing promoters generally have a broad transcription initiation region with multiple transcription initiation sites characterized by genome-wide Cap Analysis of Gene Expression (CAGE) studies that identify TSSs by sequencing of 5' end of RNAs after cDNA conversion [58]. These generally constitutively active promoters often encode housekeeping genes and are in a hypomethylated state that permits transcription. Also, CGI promoters contain nucleosome-free regions able to bind transcription factors. Characterization of transcription factor binding patterns to CGI promoters revealed that they are enriched for certain factors such as Specificity Protein 1 (SP-1), Nuclear Respiratory Factor 1 (NRF-1), E2F family of transcription factors, ETS Transcription Factor family proteins [59]. Last but not least, Histone 3 Lysine 4 trimethylation (H3K4me3) has been proposed to be a histone mark associated with CGI promoters [60], [61].

1.4 Aim of This Study

As a member of CXXC-Type Zinc Finger Family of proteins, CXXC5 has been suggested to act as a transcription factor and/or epigenetic regulator implicated in various physiological and pathophysiological processes. The *CXXC5* gene is known to be regulated by a variety of distinct and/or integrated signaling pathways including E2-ERα, retinoic acid, BMP4, and Wnt/ β-catenin signaling but mechanisms underlying this regulation are largely unknown.

The understanding of mechanisms by which the expression of *CXXC5* is modulated in various cells in a spatiotemporal manner would be critical for the diagnosis and therapy of various pathologies to which CXXC5 might contribute. Since gene expressions are largely driven by distinct promoter elements responding to diverse signaling pathways, defining the promoter of the CXXC5 gene is the aim of my thesis study.

CXXC5 has 14 transcript variants all of which encode the same protein. Combinations of 11 exons scattered throughout the approximately 35 kb long gene locus make up these 14 transcript variants. We predicted that identifying the main transcript variant (the one that is mostly expressed) and its transcription start site(s) is crucial in predicting the promoter region of *CXXC5*. The objective of this study was therefore to determine and quantify the transcript variants expressed in ER-positive and E2 responsive MCF7 breast adenocarcinoma cell line as well as in ER-negative HL60 Acute Promyelocytic Leukemia cell line, in both of which CXXC5 is shown to contribute to phenotypic features.

CHAPTER 2

MATERIALS AND METHODS

2.1 Cell Culture

MCF7 human breast adenocarcinoma and HL60 human peripheral blood acute promyelocytic leukemia cell lines were grown in DMEM (Dulbecco's Modified Eagle's medium; high glucose) supplemented with 8% fetal bovine serum, 1% penicillin/streptomycin and 1% L-glutamine and 10 µg/ml Fungin (InvivoGen, San Diego, CA, USA). Cells were maintained in 5% CO_2 humidified atmosphere at 37^{o}C.

2.2 *In silico* Analyses

2.2.1 Transcript Databases

In order to identify the annotated transcript variants of *CXXC5*, we used different databases including Ensembl (https://www.ensembl.org/index.html), National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov) and University of California-Santa Cruz Genome Browser (UCSC, https://genome.ucsc.edu/).

2.2.2 Transcript Variant Expression Levels

We used GTEx Portal (https://www.gtexportal.org/home/) database that contains gene and transcript expression level data from different tissues.

2.3 RNA Isolation and cDNA conversion

RNA was isolated from 90-100% confluent MCF7 cells and $1x10^6$ cells/ml HL60 cells grown in T-75 tissue culture flasks. MCF7 cells were trypsinized and collected in maintenance medium and HL60 cells grown in suspension were directly collected. After centrifuging at 300 g for 6 minutes, cells were washed twice with 1 x Phosphate Buffered Saline (PBS) by centrifuging at 600 g for 7 minutes. RNA was isolated from pelleted cells using the Zymo Research Quick-RNA Miniprep Kit (ZymoResearch, Irvine, CA, USA) according to the manufacturer's instructions. RNA concentration and purity were assessed by BioDrop micro-volume measurement platform. A PCR with *GAPDH* gene-specific primers was carried out with the isolated RNA to see whether there was genomic DNA contamination. 500 ng of isolated total RNA devoid of genomic DNA contamination was subjected to reverse transcriptase reaction with the ThermoFisher Scientific RevertAid First Strand cDNA Synthesis Kit (ThermoFisher Scientific, Waltham, MA, USA) according to manufacturer's instructions. Briefly, first, annealing of 1 µL of 18-nucleotide long oligo d(T) primer provided by the kit was to 500 ng of RNA in RNAse/DNAse free water was carried out. To do this, 1 µL of oligo d(T) primer was added on 500 ng of RNA and the volume was completed to 12,5 µL with nuclease-free water provided by the manufacturer. The mixture was incubated at 65°C for 5 minutes. After that, to facilitate the annealing of primers, the mixture was immediately incubated on ice for 2 minutes. Subsequently, a master mix composed of 5X Reaction Buffer, 10 mM dNTP, 20U RiboLock (RNase inhibitor), and 200U RevertAid Reverse Transcriptase was prepared and the reaction volume was completed to 20 µL with this master mix. The reaction was carried out by incubating the reaction mixture at 42°C for 1 hour. The reaction was terminated by a 5-minute-incubation at 70°C. cDNA conversion was controlled with a PCR with *GAPDH* gene-specific primers.

PCRs with primers specific to *GAPDH* gene were carried out according to the following protocol: 20 µL reaction mixture was prepared with $MgCl_2$ buffer,10X KCl buffer, 0.5 µM forward and reverse primers and 5U Taq Polymerase (ThermoFisher Scientific, Waltham, MA, USA). All PCR reactions were carried out with Techne TC-5000 thermal cycler. Initial denaturation was carried out by incubating the PCR mix at 95^{o}C for 3 minutes. After that, denaturation, primer annealing, and extension steps were carried out for 30 cycles. Denaturation was carried out by incubating the PCR mix at 95^{o}C for 30 seconds. Primer annealing was carried out at 65^{o}C for 30 seconds and product extension was carried out at 72^{o}C for 30 seconds. After 30 cycles, the final extension was carried out by incubating the mixture at 72^{o}C for 10 minutes.

2.4 Amplification of Transcript Variants

cDNAs produced from total RNA isolates of each cell line were amplified by nested rounds of PCR using primer pairs specific for each transcript variant of *CXXC5*.

2.5 TA Cloning

CXXC5 transcript variants that were amplified from the cDNA pool were cloned in pGEM-T and pGEM-T-Easy vectors for sequencing (Promega, Madison, WI, USA) by TA cloning subsequent to gel extraction of PCR products with the Zymoclean Gel DNA Recovery Kit (ZymoResearch, Irvine, CA, USA). Gel extraction was carried out according to the manufacturer's instructions. Briefly, the gel slice containing DNA fragment of interest was dissolved in Agarose Dissolving Buffer at 55^{o}C and the DNA was cleaned with a wash buffer in a Zymo-Spin column. After wash steps, the DNA was eluted from the column with MilliQ water. TA cloning was carried out in 2X Rapid Ligation Buffer by T4 DNA Ligase enzyme which was provided by Promega. The inserts were ligated to 50 ng of

pGEM-T vector in a 6:1 molar ratio. The ligation mixture was prepared and incubated at room temperature for 1 hour. After the ligation, vectors were transformed into XL1 Blue competent cells. The transformation was carried out by adding 90 µL of XL1 blue competent cells on the ligation reaction. After addition, the cells were incubated on ice for 1 hour. Then, the transformation was carried out by heat shock. The cells were incubated at 42^{o}C for 45 seconds and transferred on ice for incubation for 2 minutes. After transformation, cells were added 450 µl of Luria Bertani (LB) Broth (Sigma-Aldrich) and incubated at 37^{o}C for 1 hour in a shaker at 180 rpm. After that, bacteria were centrifuged at 4000 g for 5 minutes to remove the broth. The resultant pellet was dissolved in 50 µl of LB. The cells were then plated on LB-agar plates containing 100 µg/ml ampicillin and spread with 1M IPTG (*lac* operon inducer) and 40 µg/ml X-gal (β-galactosidase substrate) for the blue-white screening of colonies. White colonies were selected and grown in 4 ml of LB medium containing 100 µg/ml ampicillin. Colony PCR was carried out with a forward primer specific to T7 promoter and a reverse primer specific to SP6 promoter to see whether the cells of selected colonies contain the insert carrying plasmid. Plasmid isolation from positive colonies was carried out by using QIAprep Spin Miniprep Kit (QIAGEN, Hilden, Germany). To control whether the isolated plasmids contain specific inserts, PCR was carried out with insert specific primers. Plasmids that contained specific inserts were sequenced subsequently.

2.6 cDNA Walk

cDNA walk is a method adapted from the genome walk method which is used to identify unknown sequences by using the DNA regions of known sequences [62]. cDNA walk was carried out by end-point PCR of MCF7 and HL60 cDNAs by using 7 forward primers for *CXXC5* Transcript Variant 2 and 6 forward primers for *CXXC5* Transcript Variant 3 that have overlapping sequences of 7-10 nucleotides (Figure 2.1) (Figure 2.2). Together with these overlapping primers, a reverse

primer from the Exon 10 and Exon 4 was used for the cDNA walk of TV2 and TV3 respectively.

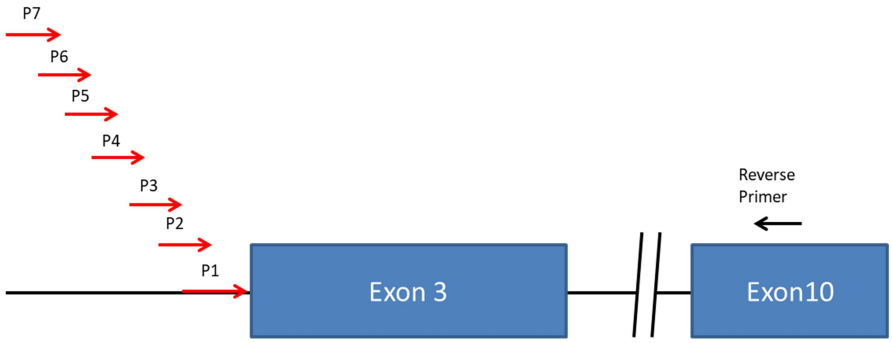


Figure 2.1. **Schematic representation of cDNA Walk of *CXXC5* Transcript Variant 2.** Blue rectangles depict the exons on genome which is represented as a black line. Red arrows represent the overlapping forward primers and black arrow represent the reverse primer.

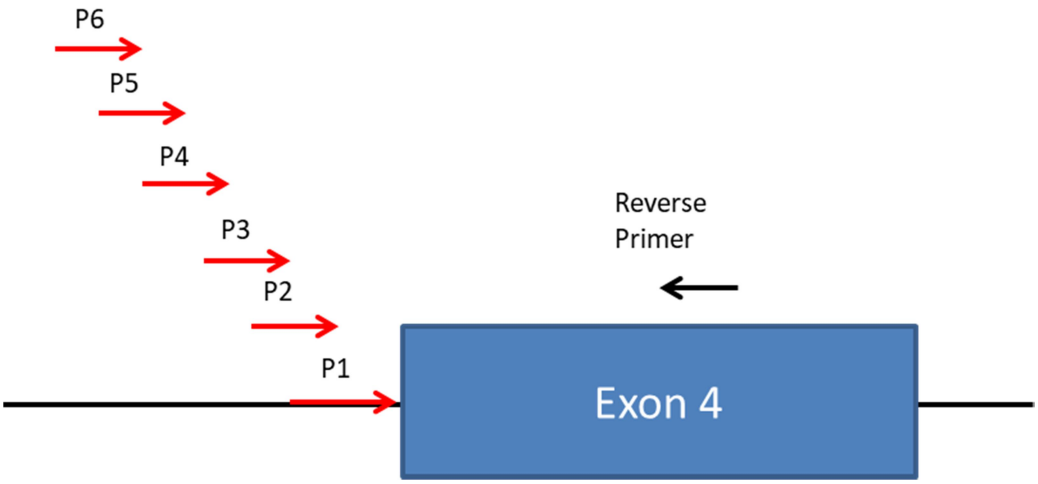


Figure 2.2. **Schematic representation of cDNA Walk of CXXC5 Transcript Variant 3.** Blue rectangles depict the exons on the genome which is represented as a black line. Red arrows represent the overlapping forward primers and the black arrow represents the reverse primer.

2.7 Quantitative PCR

Quantitative PCR (qPCR) was carried out by using the Sso Advanced™ Universal SYBR Green Supermix (BIO-RAD, Hercules, CA, USA) on CFX Connect™ Real-Time PCR Detection System (BIO-RAD, Hercules, CA, USA). Primer pairs amplifying detected transcript variants were used.

Expression levels were calculated by using an efficiency corrected form of comparative $2^{-\Delta C}{}_{T}$ method normalized by using *RPLP0* expression levels and the results are represented as two biological replicates [63].

2.8 Statistical Analyses

Graphs of qPCR analyses were plotted by using GraphPad Prism 7. Statistical analyses were carried out by using One-way ANOVA with Tukey's HSD post-hoc test on Minitab 18 software.

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Determination of *CXXC5* Transcript Variants Expressed in MCF7 and HL60 Cell Lines

Genes can have multiple transcript variants (TVs) encoding the same protein, encoding different variants of a protein, and/or non-coding transcript variants. The transcript variants are formed as a result of usage of alternative promoters, alternative exons and/or alternative splicing [64]. The expression and amount of TVs can vary in a tissue-specific manner.

TVs are identified by the Expressed Sequence Tag (EST) method that involves the generation of complementary DNA (cDNA) libraries from tissues and sequencing of these libraries [65]. The identified transcripts are deposited in different databases including Ensembl (https://www.ensembl.org/index.html), National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov) and University of California-Santa Cruz (UCSC, https://genome.ucsc.edu/).

According to NCBI, *CXXC5* has 14 transcripts all of which encode the same protein. Figure 3.1, taken from NCBI, shows the schematic representation of these variants that are enlisted according to the genomic locations of *CXXC5* exons from upstream to downstream (Figure 3.1).

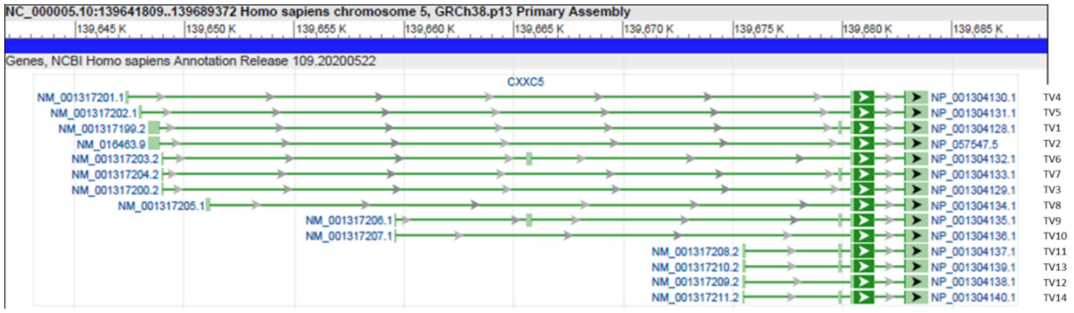


Figure 3.1. NCBI image of CXXC5 transcript variants enlisted according to the genomic location of the first exons of each variant, from upstream to downstream.

We have hypothesized that by determining the TSS(s) of the main transcript(s) could be used to identify the promoter region(s) of *CXXC5.* To determine the main transcript in predicting the promoter region, first, we explored the *CXXC5* TVs that are expressed in ER-positive and E2-responsive MCF7 breast adenocarcinoma cell line and ER-negative HL60 Acute Promyelocytic Leukemia cell line in which CXXC5 protein was initially identified [16].

To identify the *CXXC5* TVs expressed in MCF7 and HL60 cell lines, we isolated RNA from 90-100% confluent MCF7 cells and $1x10^6$ cells/ml HL60 cells grown in T-75 cell culture flasks. We converted the RNAs to cDNA and carried out end-point Polymerase Chain Reactions (PCRs) with variant-specific primer pairs. PCRs were carried out in multiple nested rounds PCR with nested primers when necessary until a DNA band of the expected size was observed on an agarose gel. Expected sized PCR products were gel extracted and sequenced following TA-cloning to pGEM-T Easy vector (Supplementary Figures). TVs 1-3, 5-8, and 10 were detected in MCF7 cells while TVs 1-8 and 10 were detected in HL60 cells (Table 3.1).

Table 3.1 Detected transcript variants in MCF7 and HL60 cells.

Transcript Variant (TV)	MCF7	HL60
TV1	+	+
TV2	+	+
TV3	+	+
TV4	-	+
TV5	+	+
TV6	+	+
TV7	+	+
TV8	+	+
TV9	-	-
TV10	+	+
TV11	-	-
TV12	-	-
TV13	-	-
TV14	-	-

The only difference in *CXXC5* TV expression between MCF7 and HL60 cell lines was the presence of TV4 in HL60 suggesting little to no difference in transcript variant expression in these two cell lines.

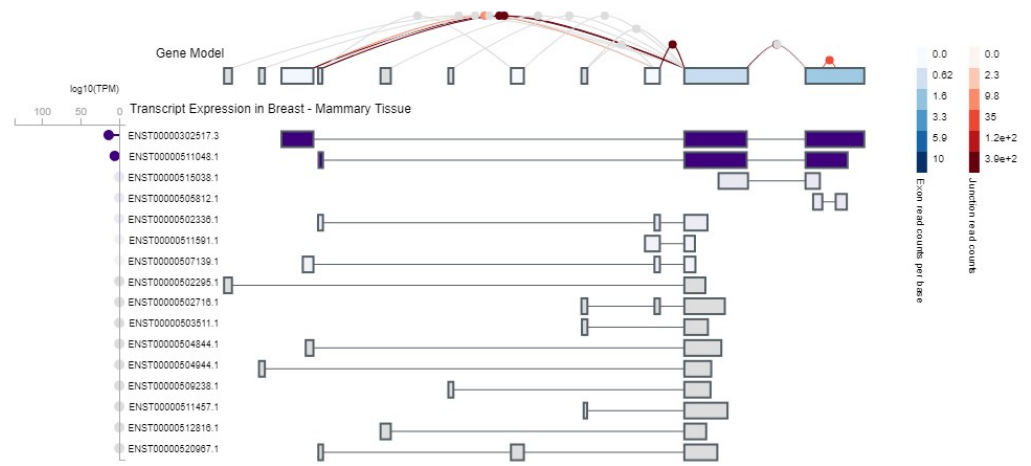
3.2 Quantification of *CXXC5* Transcript Variants in MCF7 and HL60 Cell Lines

3.2.1 *In silico* analyses of Expression Levels of *CXXC5* Transcript Variants

We initially carried out *in silico* analyses of *CXXC5* Transcript Variants expression levels. We used the GTEx Portal, which is a database of gene and transcript expression levels in different tissues [66]. The database is a collection of results of several molecular biology techniques to detect expression levels of transcripts, such as Whole Genome Sequencing (WGS), Whole Exome Sequencing (WES) and RNA-sequencing (RNA-seq). In the process of database formation, samples from 1000 individuals comprising 53 non-diseased tissues were analyzed.

We analyzed *CXXC5* TV expression in breast tissue and blood; since our cell lines of studies are cancers of these tissues. We observed in both breast tissue and blood that TV2 and TV3, containing Exons 3, 10, 11, and Exons 4, 10, 11 respectively, are the ones with the highest expression levels (Figure 3.2).

a



b

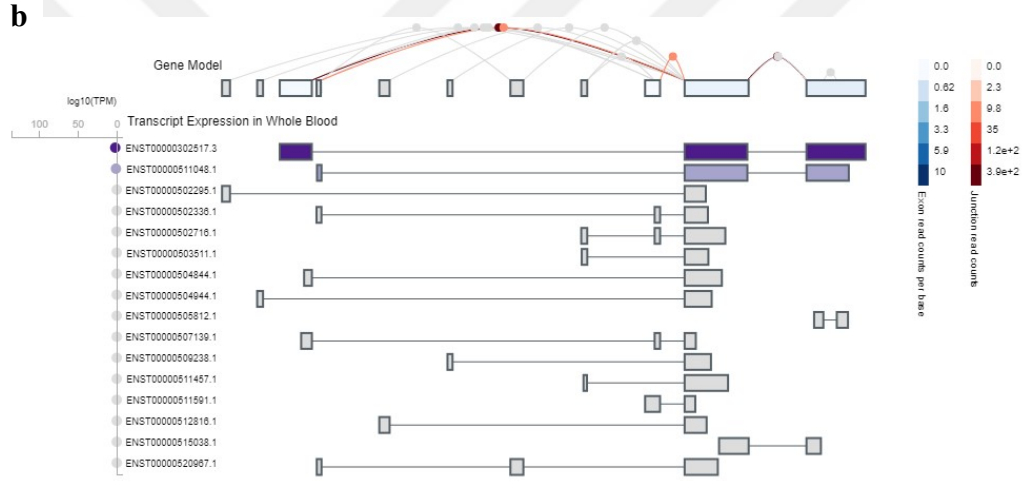


Figure 3.2. **GTEx Portal Analysis of *CXXC5* TV expression levels in a) breast tissue and b) blood tissue.** TV2 and TV3 are observed to have the highest expression levels in both tissues.

3.2.2 Quantification of *CXXC5* Transcript Variants

After *in silico* studies, we isolated RNA from 90-100% confluent MCF7 cells and $1x10^6$ cells/ml HL60 cells grown in T-75 cell culture flasks in order to be able to quantify the transcript variants of *CXXC5*. Following cDNA conversion, we carried out Quantitative PCR (qPCR) with TV-specific primer pairs (Figure 3.3).

We calculated the relative expressions of *CXXC5* TVs according to a commonly used normalization gene, which is the ribosomal protein, large, P0, *RPLP0*, gene. For the calculation of relative expressions of TVs, we used the efficiency corrected form of comparative $2^{-\Delta C}{}_{T}$ method which can be utilized to compare relative expressions of transcript pairs [63], [67], [68].

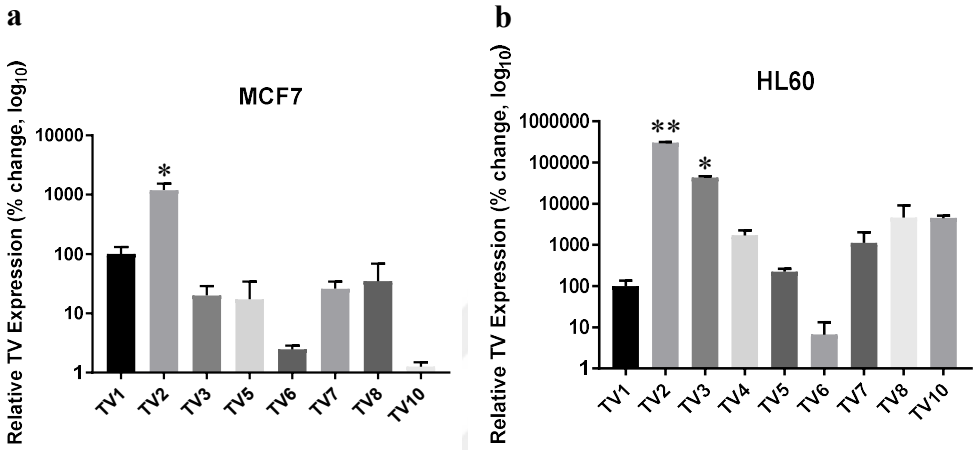


Figure 3.3. **Quantification of *CXXC5* TVs in a) MCF7 and b) HL60 cells.** RNA was isolated from cells which were grown in DMEM-FBS containing growth medium. RNA samples were converted to cDNA and cDNAs were subjected to qPCR. Expression levels were calculated by using the efficiency corrected form of comparative $2^{-\Delta C}{}_{T}$ method normalized by using *RPLP0* expression levels. [3] The expression level of Transcript Variant 1 was set to 100 and the expression levels of other transcripts were adjusted accordingly and shown in logarithmic scale. The results are the average of two biological replicates. Statistical analysis was performed with One-Way ANOVA test with Tukey's HSD test, 95% confidence interval, $p < 0.05$.

Our qPCR analyses revealed that *CXXC5* TV2, which contains Exons 3, 10, and 11 is the transcript that has the highest relative expression in both MCF7 and HL60 cells. This result is consistent with the GTEx Portal analyses, in which we observed *CXXC5* TV2 to have the highest expression level among *CXXC5* TVs. Thus, we identified *CXXC5* TV2 as the main transcript variant and we wanted to identify its

Transcription Start Site (TSS) to be able to predict the region that contains the putative promoter of *CXXC5*.

3.3 Determination of the Transcription Start Site of the Main Transcript Variant of *CXXC5*

To determine the TSS of the main transcript, TV2, a cDNA walk method was utilized. The cDNA walk is an approach that we adapted from the genome walking method in which a known sequence of DNA is used to identify unknown sequences through PCR amplification. In the cDNA walk, we used 7 forward primers were designed annealing sites starting from the beginning of Exon 3 as shown in databases such that, each one having 7-10 overlapping nucleotides with the previous forward primer. We named the forward primers (FPs) Walk FP1-7. To carry out cDNA walk, RNA was isolated from 90-100% confluent MCF7 cells and 1x10 6 cells/ml HL60 cells grown in T-75 cell culture flasks and converted to cDNA. PCRs were carried out with the 7 forward primers and a reverse primer from Exon 10. PCR results were visualized on 1% agarose gel and it was observed that expected sized PCR products were amplified with first 5 of the primer sets (Figure 3.4). The product of the fifth forward primer and the reverse primer from Exon 10 from both MCF7 and HL60 cDNAs was cloned into pGEM-T vector (Promega, Madison, WI, USA) by TA cloning and its specificity was confirmed by sequencing. It was concluded that the transcription start site of TV2 was within the 21 bp region to which the FP5 anneals (hg38 chr5: 139648232-139648252).

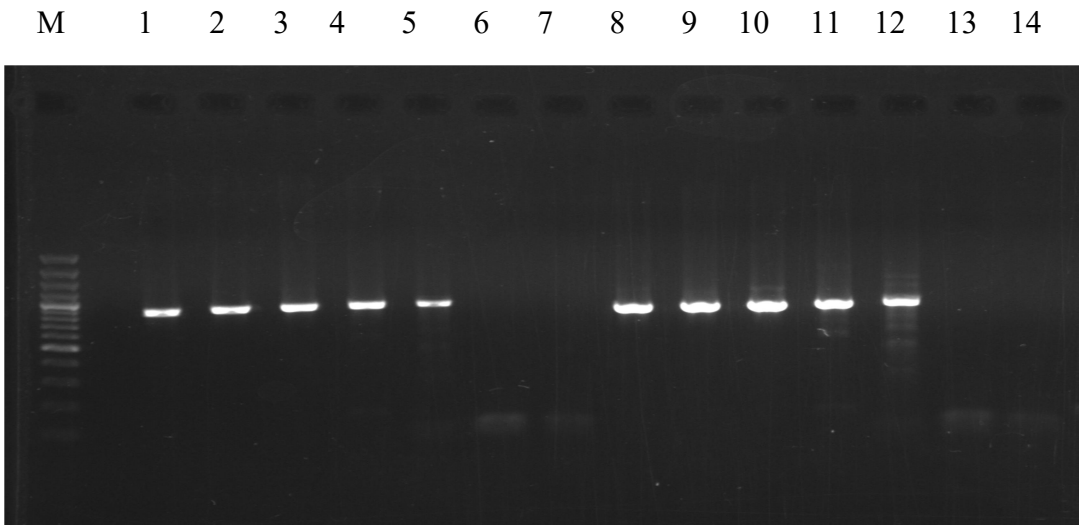


Figure 3.4. **Agarose gel image of the cDNA Walk PCR of *CXXC5* Transcript Variant 2 with overlapping forward primers (FP1-7) from upstream of Exon 3 and a reverse primer from Exon 10.** The products in lanes 1-7 are products of FP1-7 and a reverse primer from Exon 10 with MCF7 cDNA template and the products in lanes 1-7 are products of FP1-7 and a reverse primer from Exon 10 with HL60 cDNA template. The first 5 forward primers (FP1-5) gave products with both MCF7 and HL60 cDNA templates while the last two forward primers (FP6-7) did not, suggesting that the 5' end of the *CXXC5* Transcript Variant 2 resides in the region to which FP5 anneals (hg38 chr5: 139648232-139648252).

One of the recent approaches to identify 5' ends of transcripts is 5' Rapid Amplification of cDNA Ends (5' RACE) in which an RNA oligo with a known sequence called adaptor is used to detect unknown sequences at the 5' end of transcripts [69]. Although prone to biases introduced by various factors such as the secondary structure of RNA, G-C content, adaptor ligation efficiency, etc, this technique is quite useful.

In our laboratory, 5' RACE was utilized to detect the 5' end of the main TV of *CXXC5*, TV2. The sequencing of 5' RACE products revealed two different TSSs for TV2 which are 21 bp upstream and 11 bp downstream of known TSS of TV2 (Figure 3.5). This suggests that a broad transcription start region is used, instead of a single TSS, in the expression of TV2. The utilization of multiple TSSs in a broad

region is one of the characteristics of CGI-containing promoters. Our results are also reflected in our bioinformatics analyses showing that the transcription start region of TV2 is found in a GC rich region with a GC base composition greater than 50% and a CpG observed/expected ratio of more than 0.6. This is suggestive of a CG island. Altogether, these data suggest that the promoter of *CXXC5* resides within the transcription start region of TV2 and shows the characteristics of CGI promoters.

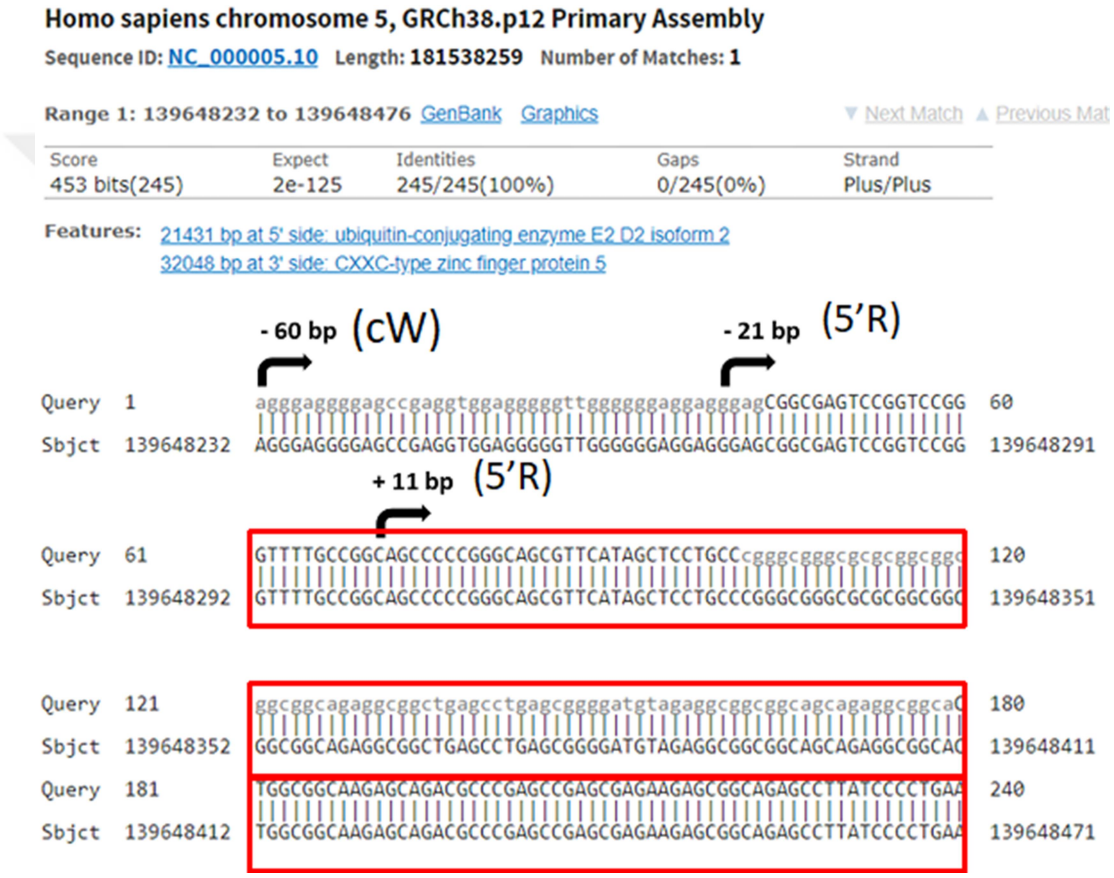


Figure 3.5. **NCBI Blast representation of TSSs of CXXC5.** Our cDNA walk (cW) and 5' RACE (5'R) approaches revealed several TSSs located at 60 bp upstream, 21 bp upstream, and 11 bp downstream of the known TSS of TV2. Exon 3 sequence starting from the known TSS of TV2 is marked with red rectangles.

3.4 Identification of a Novel Transcript Variant of *CXXC5*

3.4.1 Detection of a Novel Transcript Variant (NTV) of *CXXC5*

NCBI database indicates the presence of a *CXXC5* TV, referred to as TV3. TV3 contains an 85-bp-long exon together with Exons 10 and 11 which contain the coding sequence for CXXC5. This small exon, Exon 4, is located 80 bp downstream of the first exon of the main TV of *CXXC5*, TV2. Since we observed a broad region of transcription initiation for *CXXC5*, as is the case for many CGI-containing promoters, to obtain a better understanding of the range of transcription initiation region, we also carried out cDNA walk for TV3 with 6 forward primers (FP1-6) each one containing 7 overlapping nucleotides with the previous one and spanning the intron between Exons 3 and 4 by using MCF7 and HL60 cDNAs (Figure 3.6).

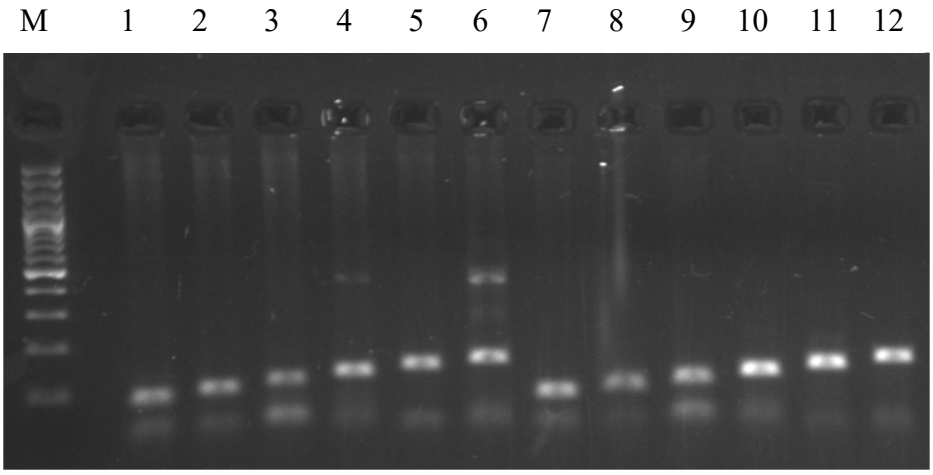


Figure 3.6. **Agarose gel image of the cDNA walk PCR of *CXXC5* Transcript Variant 3 with overlapping forward primers (FP1-6) from upstream of Exon 4 and a reverse primer from Exon 4.** The products in lanes 1-6 are products of the FP1-6 and a reverse primer from Exon 4 with MCF7 cDNA template and the products in the last 6 lanes are the products of the FP1-6 and a reverse primer from Exon 4 with HL60 cDNA template. All the primer sets encompassing the intron between Exons 3 and 4 gave products, suggesting the presence of a novel transcript variant that contains both Exons 3 and 4.

All the forward primers spanning the intron between Exons 3 and 4 gave specific products with their corresponding reverse primer and the last forward primer, FP6, which has 5 nucleotides annealing to Exon 3. This led to our prediction that there might be a novel transcript variant that contains both Exons 3 and 4. To test this prediction, we carried out PCR with a 21-bp-long forward primer together with a reverse primer from Exon 10. The forward primer had 18 nucleotides annealing to Exon 3 and 3 nucleotides annealing to Intron 3. We couldn't use a forward primer that completely anneals to Exon 3. This was because, primer pairs from Exon 3 and Exon 10 selectively amplified the main transcript variant, TV2. This consequently prevents the assessment of the presence of the novel transcript variant containing both Exons 3 and 4 in agarose gel electrophoresis. However, when we

cloned the and sequenced the generated PCR amplicon, we confirmed the presence of a novel transcript variant of *CXXC5* in both MCF7 and HL60 cells (Figure 3.7).

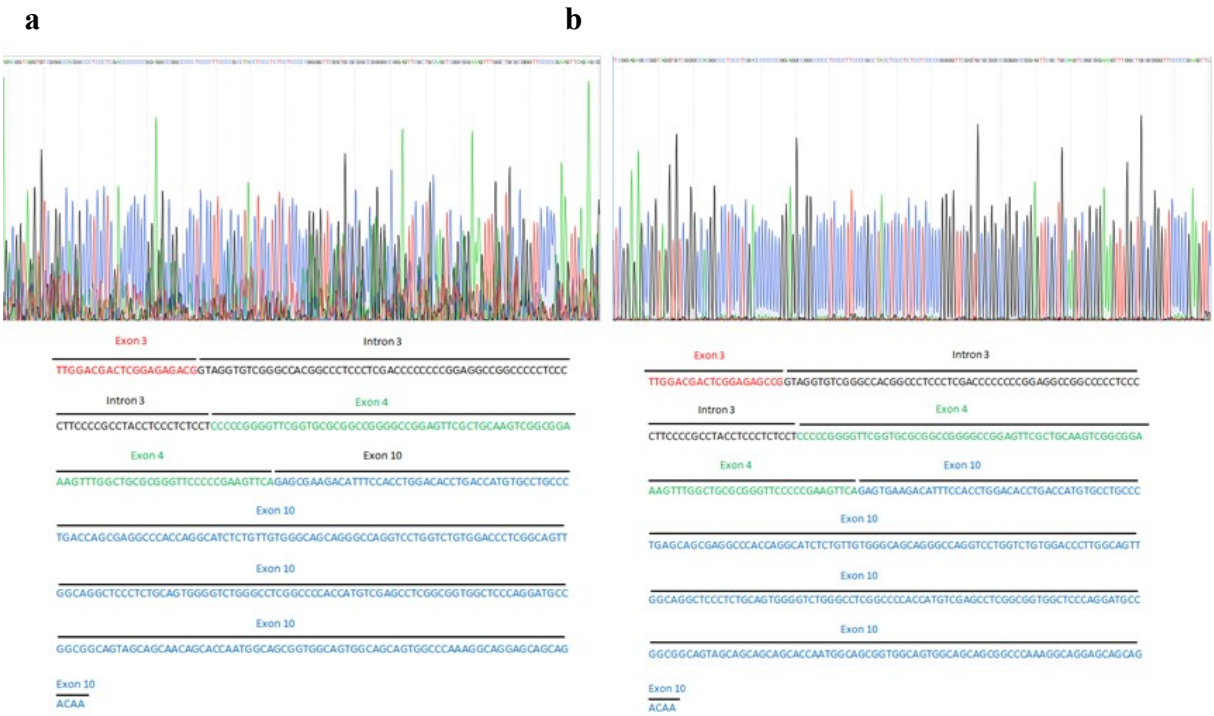


Figure 3.7. **Sequencing result of novel transcript variant of *CXXC5* in a) MCF7 b) HL60 cells.** Red nucleotides show Exon 3 sequences, black nucleotides indicate Intron 3 sequences, green nucleotides show Exon 4 sequences and blue nucleotides signifies Exon 10 sequences.

3.4.2 Quantification of NTV

We wanted to quantify the NTV in MCF7 and HL60 cells to compare its expression levels with the main TV, TV2, and the other transcript variant that contains Exon 4, TV3. We used the efficiency corrected form of comparative $2^{-\Delta C_T}$ method to compare the expression levels of the three transcript variants (Figure 3.8)

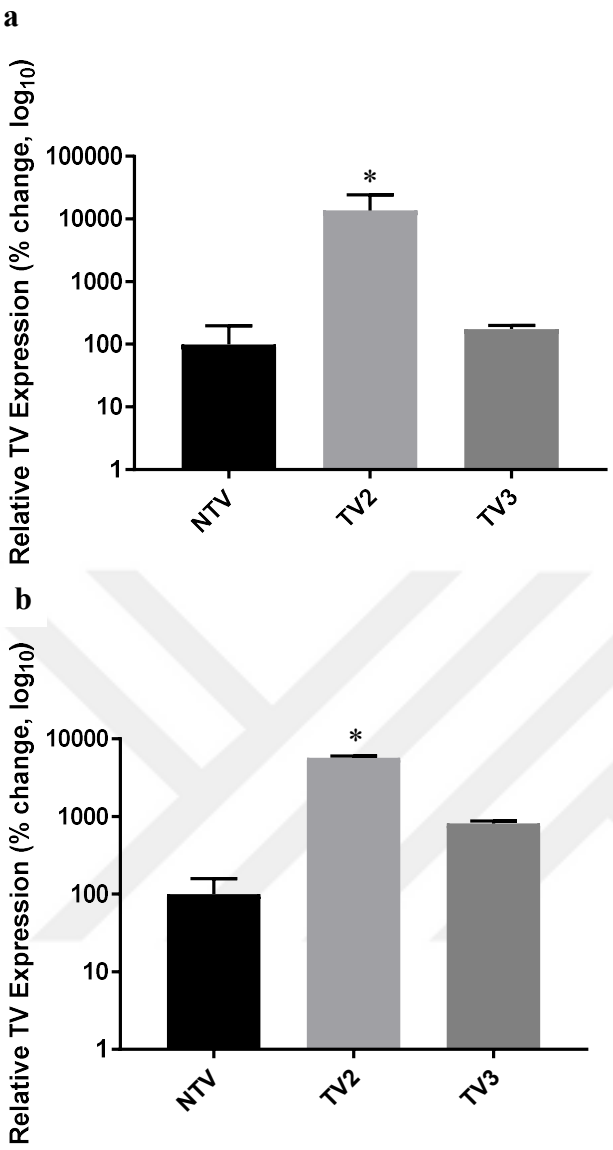


Figure 3.8. Quantification of the level of the Novel *CXXC5* Transcript Variant (NTV) containing Exons 3, 4, 10 and 11 and comparison with TV2 and TV3 in a) MCF7 and b) HL60 cells. Expression levels were calculated by using the efficiency corrected form of comparative $2^{-\Delta C}{}_{T}$ method normalized by using *RPLP0* expression levels. [3] The expression level of NTV was set to 100 and the expression levels of other transcripts were adjusted accordingly and shown in logarithmic scale. The results are the average of two biological replicates. Statistical analysis was performed with One-Way ANOVA test with Tukey's HSD test, 95% confidence interval, $p < 0.05$.

Our results revealed that the expression levels of the NTV are substantially lower than those of TV2 in both MCF7 and HL60 cell lines. In conclusion, the NTV is a newly identified transcript variant of *CXXC5* whose expression is a result of broad transcription initiation region which is a characteristic of CGI-containing promoters.

CHAPTER 4

CONCLUSION AND FUTURE DIRECTIONS

CXXC5 is a gene whose expression has previously shown to be regulated by various signaling pathways. The deregulations in the expression of *CXXC5* have been implicated in different disease states including various cancers. ERα + breast cancers are among the malignancies in which CXXC5 has been shown to be a potential prognostic factor and therapeutic target. In our laboratory, *CXXC5* has been shown to be an estrogen-responsive gene. This renders it crucial to elucidate the mechanisms through which *CXXC5* is expressed to be able to have a better understanding of the role of CXXC5 in various physiologies and pathophysiologies. Identification and characterization of the promoter region of the *CXXC5* gene is critical in understanding of the regulation of the *CXXC5* gene expression.

CXXC5 has 14 transcript variants all of which encode the same protein. This information led us to hypothesize that the promoter region could reside in close proximity to the transcription start site of the main transcript variant which shows the highest expression levels. We used MCF7 cell line, which is an ERα + breast adenocarcinoma model and HL60, acute promyelocytic leukemia cell line which does not express ERα and thus, is not responsive to E2.

Our studies revealed several findings that could be summarized as follows:

1. 8 of the 14 previously annotated *CXXC5* transcript variants are expressed in MCF7 cells and 9 of the 14 previously annotated *CXXC5* transcript variants are expressed in HL60 cells suggesting that the expression patterns of *CXXC5* transcript variants are similar in these cell lines.
2. *CXXC5* Transcript Variant 2, which contains Exons 3, 10 and 11, is identified as the main transcript variant of *CXXC5,* since it is the one with

the highest expression level in both cell lines. This finding is consistent with the previously obtained high-throughput data that is deposited as a database called GTEx Portal. The GTEx Portal analysis shows that *CXXC5* Transcript Variant 2 has the highest expression levels in breast tissue and blood.

3. Our studies to identify the transcription start site of the main transcript variant of *CXXC5* revealed multiple transcription start sites scattered around the 100 bp region encompassing the previously annotated transcription start site of *CXXC5* Transcript Variant 2. This suggests that the variant has a transcription initiation region, rather than a defined site. This is a characteristic of CGI-containing promoters and the GC-content characteristic of this region suggests that a CGI-containing promoter might be driving the expression of *CXXC5*.
4. Our studies to have a grasp of the ranges of transcription initiation region of *CXXC5* revealed the presence of a novel transcript variant that contains Exons 3, 4, 10 and 11 in both MCF7 and HL60 cell lines. This variant's expression levels are substantially lower than that of the main transcript variant. Exons 3 and 4 are located in close proximity on the genome and the presence of such a transcript can be a result of the broad transcription initiation region from which the expressions of transcript variants that contain Exons 3 and 4 are driven.

Our studies to identify the location of the promoter region of *CXXC5* focused on the region in the genome that contains Exons 3 and 4. Luciferase studies carried out in our laboratory identified a regulatory region located at the beginning of Exon 3. The epigenetic characterization of this region, carried out in our laboratory, revealed a hypomethylated state and histone modifications that are markers of active transcription. This region has also been shown in our studies to have a nucleosome-free region. All in all, these epigenetic characteristics also suggest this region to be a CGI-containing promoter that drives the expression of *CXXC5*. The transcription factor binding characterization was also carried out in our laboratory by promoter-pulldown assays and Chromatin Immunoprecipitation (ChIP). Several CGI-promoter associated proteins were shown to bind to this region, further supporting the CGI promoter characteristic of this region.

CRISPR-Cas9 is an emerging method to induce various changes to the genome, including insertions and deletions as well as epigenetic changes by the recruitment of chromatin modifying proteins. This technique can be quite useful in the verification of the effect of the *CXXC5* promoter region identified in our laboratory on the gene's expression.

As previously indicated, *CXXC5* is regulated by a number of signaling pathways. Most of these pathways involve the direct or indirect interaction of transcription factors with the DNA. These transcription factors can either act by binding to gene promoters or to other regulatory regions such as enhancers. One of these regulatory regions is an Estrogen Response Element (ERE) identified in our laboratory. It is located at 242 bp upstream of translation initiation site and the linear distance between this ERE and the *CXXC5* promoter is approximately 30-kb. The interaction of such distal regulatory sites with promoters is mediated by chromatin loops. Investigating the regions that interact with *CXXC5* promoter can be crucial to address the regulatory mechanisms of *CXXC5* in response to signaling pathways including estrogens.

CHAPTER 5

SUPPLEMENTARY DATA

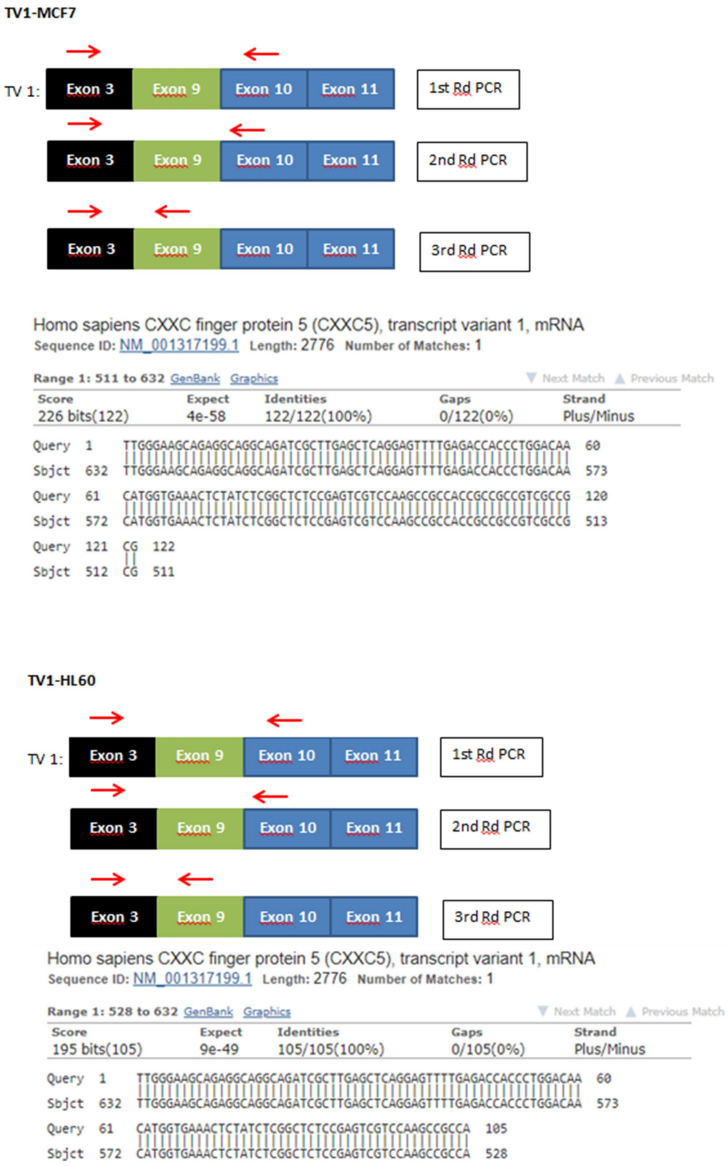


Figure 5.1. Schematic representation of the locations of primers used to amplify *CXXC5* TV1 from MCF7 and HL60 cDNAs and the Nucleotide BLAST image of the sequencing result.

Figure 5.2. Schematic representation of the locations of primers used to amplify *CXXC5* TV2 from MCF7 and HL60 cDNAs and the Nucleotide BLAST image of the sequencing result.

Figure 5.3. Schematic representation of the locations of primers used to amplify *CXXC5* TV3 from MCF7 and HL60 cDNAs and the Nucleotide BLAST image of the sequencing result.

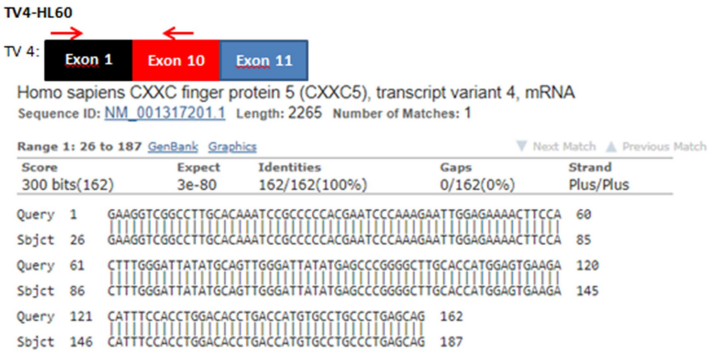


Figure 5.4. Schematic representation of the locations of primers used to amplify *CXXC5* TV4 from HL60 cDNA and the Nucleotide BLAST image of the sequencing result.

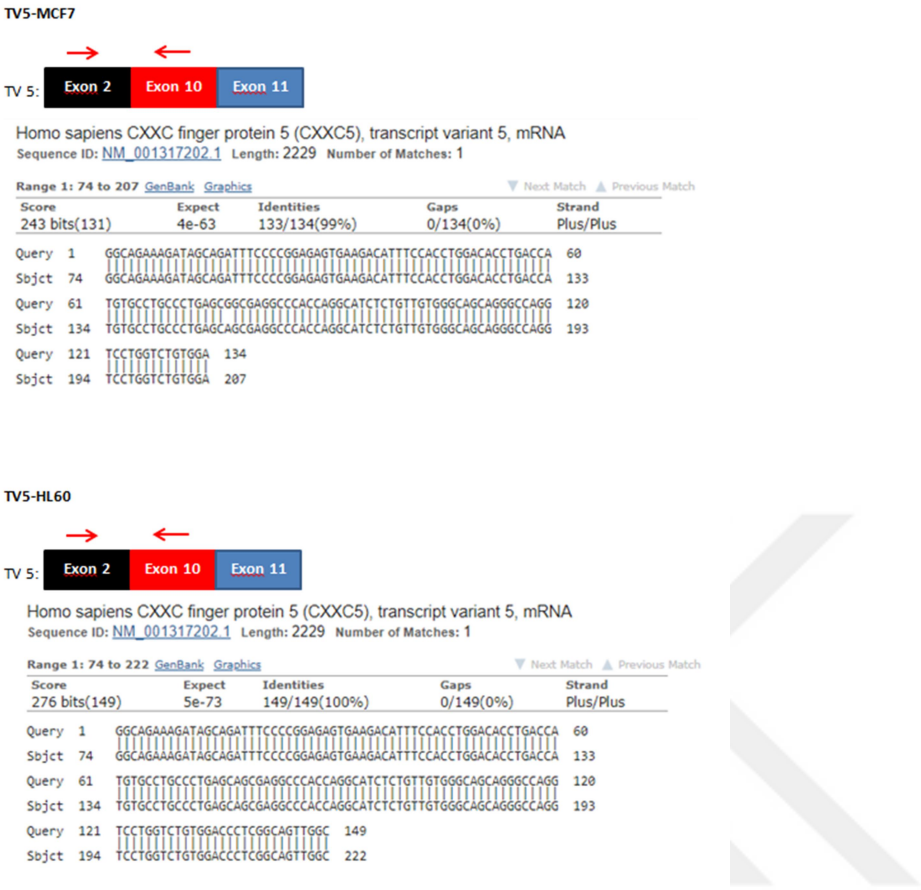


Figure 5.5. Schematic representation of the locations of primers used to amplify *CXXC5* TV5 from MCF7 and HL60 cDNAs and the Nucleotide BLAST image of the sequencing result.

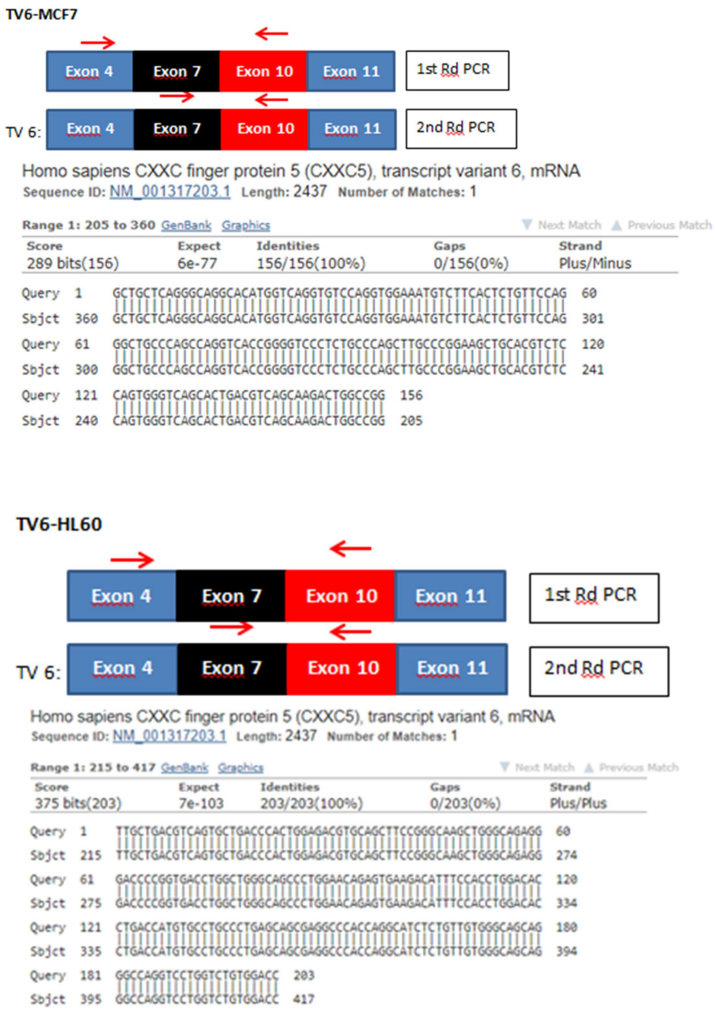


Figure 5.6. Schematic representation of the locations of primers used to amplify *CXXC5* TV6 from MCF7 and HL60 cDNAs and the Nucleotide BLAST image of the sequencing result.

Figure 5.7. Schematic representation of the locations of primers used to amplify *CXXC5* TV7 from MCF7 and HL60 cDNAs and the Nucleotide BLAST image of the sequencing result.

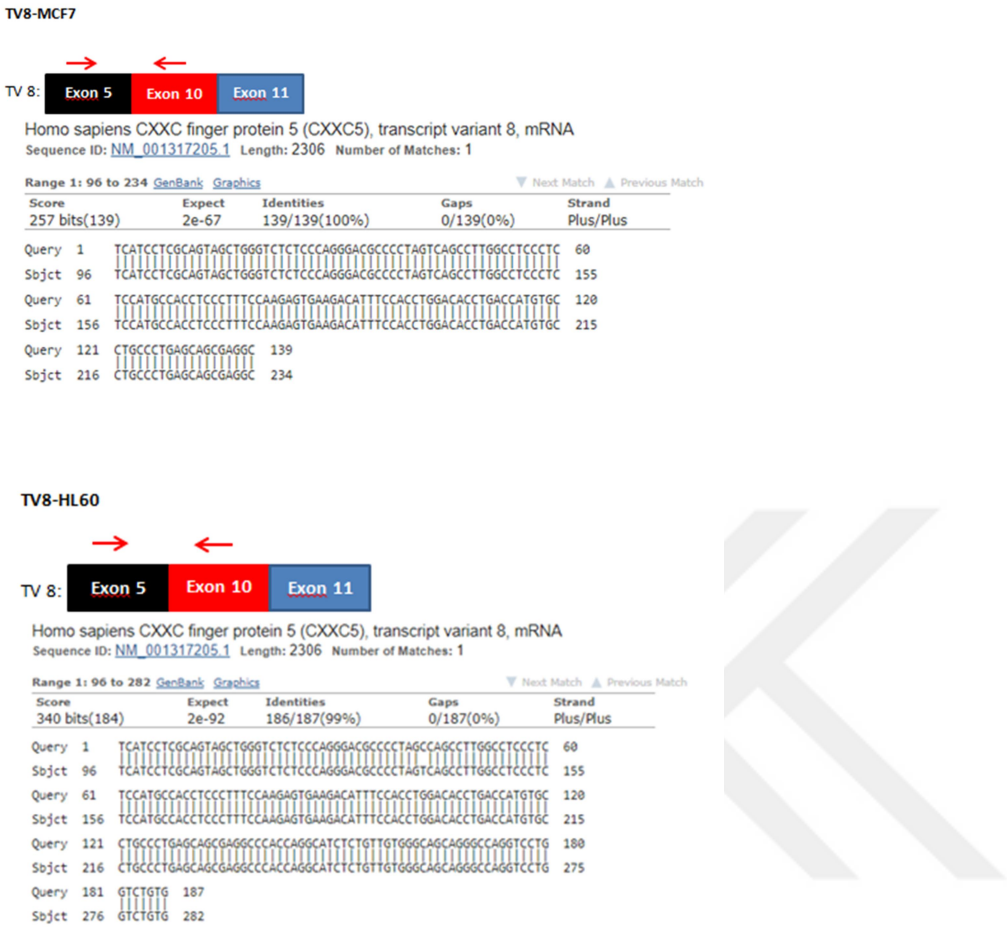


Figure 5.8. Schematic representation of the locations of primers used to amplify *CXXC5* TV8 from MCF7 and HL60 cDNAs and the Nucleotide BLAST image of the sequencing result.

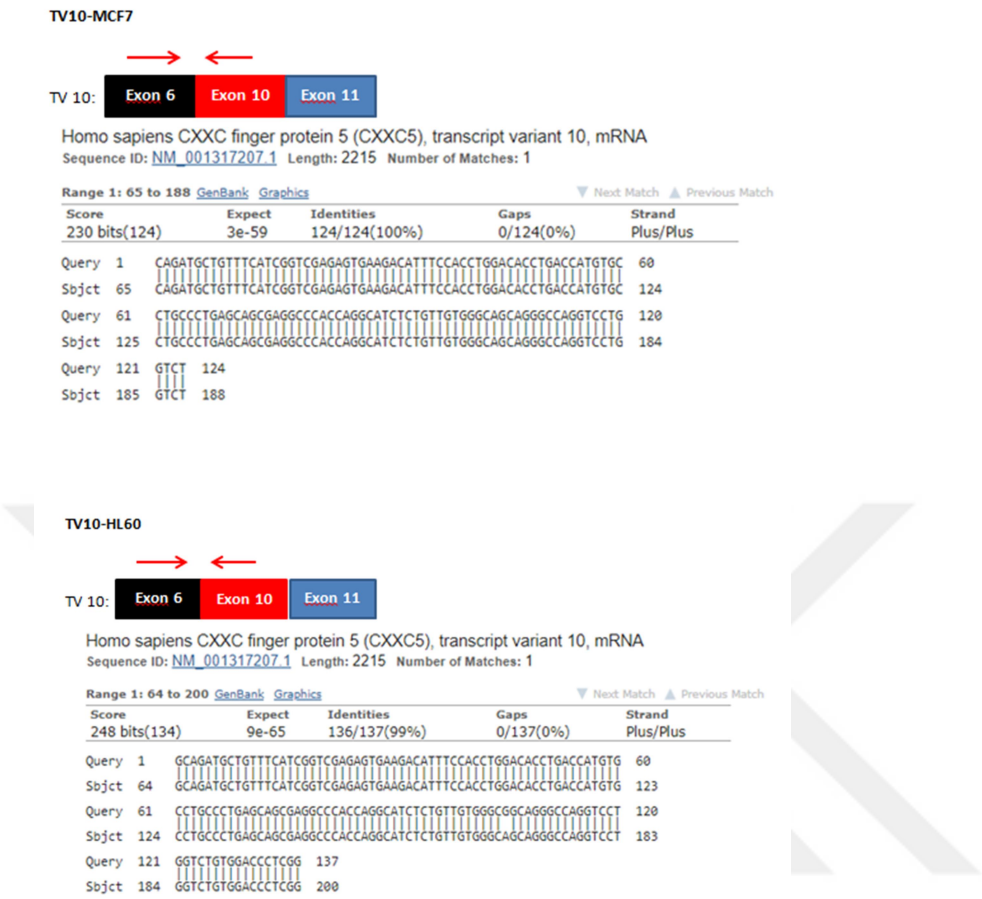


Figure 5.9. Schematic representation of the locations of primers used to amplify *CXXC5* TV10 from MCF7 and HL60 cDNAs and the Nucleotide BLAST image of the sequencing result.

Table 5.1 *GAPDH* Primer Sequences.

Primer Name	**Primer Sequence (5' to 3')**
GAPDH Forward Primer	GGGAGCCAAAAGGGTCATCA
GAPDH Reverse Primer	TTTCTAGACGGCAGGTCAGGT

Table 5.2 Sequences of primers used for detection of transcript variants.

Transcript Variant	Forward Primer Sequence (5' - 3') (1st Round)	Forward Primer Sequence (5' - 3') (2nd Round)	Reverse Primer Sequence (5' – 3') (1st Round)	Reverse Primer Sequence (5' – 3') (2nd Round)	Reverse Primer Sequence (5' – 3') (3rd Round)
1	ATGTAGAGGCGGCG GCAGCA	-	TTGTCTGCTGCTC CTGCCTTT	TTGGGAAGCAGA GGCAGGCAGA	
2	ATGTAGAGGCGGCG GCAGCA	-	ACCACCACTGCT GCCAAAAGA	TTGTCTGCTGCTC CTGCCTTT	CATTGGTGCTG CTGCTGCTA
3	TGCAAGTCGGCGGA AAGTTT	-	TTGTCTGCTGCTC CTGCCTTT		
4	ACTGCGTTTTCTTCG CCTCG	GAAGGTCGGCC TTGCACAAA	ACCACCACTGCT GCCAAAAGA	TTGTCTGCTGCTC CTGCCTTT	
5	GGCAGAAAGATAGC AGATTTCCCC	-	ACCACCACTGCT GCCAAAAGA	TTGTCTGCTGCTC CTGCCTTT	
6	TGCAAGTCGGCGGA AAGTTT	-	TTGTCTGCTGCTC CTGCCTTT	-	-
7	TGCAAGTCGGCGGA AAGTTT	-	TTGTCTGCTGCTC CTGCCTTT	-	-
8	GAGATGGGAACAGC TGAAGG	CTCATCCTCGC AGTAGCTGG	ACCACCACTGCT GCCAAAAGA	TTGTCTGCTGCTC CTGCCTTT	TTGGGAAGCA GAGGCAGG
9	GCAGATGCTGTTTC ATCGGTC	GGGCTGATGGA GACAGATGT	TTGTCTGCTGCTC CTGCCTTT	TTGGGAAGCAGA GGCAGG	
10	GCAGATGCTGTTTC ATCGGTC		TTGTCTGCTGCTC CTGCCTTT	CATTGGTGCTGCT GCTGCTA	
11	GCTTAGCATTCCAG GCCCTG	CAACCACGGAG AGGTGACA	ACCACCACTGCT GCCAAAAGA	TTGTCTGCTGCTC CTGCCTTT	TTGGGAAGCA GAGGCAGG
12	GCTTAGCATTCCAG GCCCTG	CAACCACGGAG AGGTGACA	ACCACCACTGCT GCCAAAAGA	TTGTCTGCTGCTC CTGCCTTT	TTGGGAAGCA GAGGCAGG
13	GCTTAGCATTCCAG GCCCTG	CAACCACGGAG AGGTGACA	ACCACCACTGCT GCCAAAAGA	TTGTCTGCTGCTC CTGCCTTT	TTGGGAAGCA GAGGCAGG
14	ATCTCCAAGTGCCC CAGGTGG		ACCACCACTGCT GCCAAAAGA		
NTV	GGACGACTCGGAG AGCCGGTA		AAAGGCAGGAG CAGCAGACAA		

Table 5.3 *CXXC5* Transcript Variant 2 cDNA Walk Primer Sequences

Primer Name	**Primer Sequence (5' to 3')**	**T_m (^oC)**	**DMSO (%)**
TV2 P1	GTTTTGCCGGCAGCCCCCGGG	55	5
TV2 P2	GGAGCGGCGAGTCCGGTCCGG	55	5
TV2 P3	CGCATCTCGAGGTTGGGGGGAGGAGGGAGCGG	65	5
TV2 P4	CGCATCTCGAGAGGTGGAGGGGGTTGGGGGGA	65	5
TV2 P5	CGCATCTCGAGAGGGAGGGGAGCCGAGGTGGA	65	5
TV2 P6	CGCATCTCGAGGAGGCGGCGGGAGGAGGGAGG	55	5
TV2 P7	CGCATCTCGAGGCTGCGCTGGGAGGGAGGCGG	55	5
TV2 Reverse Primer	CGCATAAGCTT TTGTCTGCTGCTCCTGCCTTT	Same with whichever forward primer was used	5

Table 5.4 *CXXC5* Transcript Variant 3 cDNA Walk Primer Sequences

Primer Name	**Primer Sequence (5' to 3')**	**T_m (^oC)**	**DMSO (%)**
TV3 P1	CGCCTACCTCCCTCTCCTC	60	5
TV3 P2	CTCCCCTTCCCCGCCTA	60	5
TV3 P3	GGAGGCCGGCCCCCTCCCCTT	60	5
TV3 P4	CCTCGACCCCCCCCGGAGGCC	60	5
TV3 P5	GGCCACGGCCCTCCCTCGAC	60	5
TV3 P6	AGCCGGTAGGTGTCGGGCCAC	60	5
TV3 Reverse Primer	AACCCGCGCAGCCAAACTT	60	5

Table 5.5 qPCR Primer Sequences

Transcript Variant	**Forward Primer Sequence (5' to 3')**	**Reverse Primer Sequence (5' to 3')**	**T_m (^oC)**	**DMSO**
1	CAGCTCGGGGGTGATTAGTTG	ATCGCTTGAGCTCAGGAGTTT	65	5
2	CAGCTCGGGGGTGATTAGTTG	AACAGAGATGCCTGGTGGGCCT	65	5
3	AAGTTTGGCTGCGCGGGTT	AACAGAGATGCCTGGTGGGCCT	55	0
4	ACTGCGTTTTCTTCGCCTCG	TCCCAACTGCATATAATCCCA	60	0
5	CGCGGCCAAATTCTCCCTCCC	CTCCGGGGAAATCTGCTATCT	60	5
6	TGCAAGTCGGCGGAAAGTTT	ACATCTGTCTCCATCAGCCC	55	5
7	TGCAAGTCGGCGGAAAGTTT	ATCGCTTGAGCTCAGGAGTTT	65	0
8	AACAGCTGAAGGAGATTGGCC	CCAGCTACTGCGAGGATGAGG	65	5
10	AGCCTTCCCCAGGGGCCTTGT	AACAGAGATGCCTGGTGGGCCT	55	0
NTV	CTCCCCTTCCCCGCCTA	AAAGGCAGGAGCAGCAGACAA	60	5

REFERENCES

[1] N. Fuentes and P. Silveyra, "Estrogen receptor signaling mechanisms," in *Advances in Protein Chemistry and Structural Biology*, vol. 116, 2019, pp. 135–170.

[2] P. Yaşar, G. Ayaz, and M. Muyan, "Estradiol-Estrogen Receptor α Mediates the Expression of the CXXC5 Gene through the Estrogen Response Element-Dependent Signaling Pathway," *Sci. Rep.*, vol. 6, 2016, doi: 10.1038/srep37808.

[3] P. Ascenzi, A. Bocedi, and M. Marino, "Structure-function relationship of estrogen receptor α and β: Impact on human health," *Molecular Aspects of Medicine*, vol. 27, no. 4. pp. 299–402, 2006, doi: 10.1016/j.mam.2006.07.001.

[4] C. E. Mason *et al.*, "Location analysis for the estrogen receptor-α reveals binding to diverse ERE sequences and widespread binding within repetitive DNA elements," *Nucleic Acids Res.*, vol. 38, no. 7, pp. 2355–2368, 2010, doi: 10.1093/nar/gkp1188.

[5] S. L. Nott *et al.*, "Genomic responses from the estrogen-responsive element-dependent signaling pathway mediated by estrogen receptor α are required to elicit cellular alterations," *J. Biol. Chem.*, vol. 284, no. 22, pp. 15277–15288, 2009, doi: 10.1074/jbc.M900365200.

[6] J. Huang, X. Li, R. Hilf, R. A. Bambara, and M. Muyan, "Molecular basis of therapeutic strategies for breast cancer," *Current Drug Targets: Immune, Endocrine and Metabolic Disorders*, vol. 5, no. 4. pp. 379–396, 2005, doi: 10.2174/156800805774912944.

[7] E. Castro-Rivera, I. Samudio, and S. Safe, "Estrogen Regulation of Cyclin D1 Gene Expression in ZR-75 Breast Cancer Cells Involves Multiple

Enhancer Elements," *J. Biol. Chem.*, vol. 276, no. 33, pp. 30853–30861, 2001, doi: 10.1074/jbc.M103339200.

[8] H. K. Long, N. P. Blackledge, and R. J. Klose, "ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection," in *Biochemical Society Transactions*, 2013, vol. 41, no. 3, pp. 727–740, doi: 10.1042/BST20130028.

[9] C. Xu *et al.*, "DNA Sequence Recognition of Human CXXC Domains and Their Structural Determinants," *Structure*, vol. 26, no. 1, p. 85–95.e3, 2018, doi: 10.1016/j.str.2017.11.022.

[10] J. C. Zhou, N. P. Blackledge, A. M. Farcas, and R. J. Klose, "Recognition of CpG Island Chromatin by KDM2A Requires Direct and Specific Interaction with Linker DNA," *Mol. Cell. Biol.*, vol. 32, no. 2, pp. 479–489, 2012, doi: 10.1128/mcb.06332-11.

[11] M. Ko *et al.*, "Modulation of TET2 expression and 5-methylcytosine oxidation by the CXXC domain protein IDAX," *Nature*, vol. 497, no. 7447, pp. 122–126, 2013, doi: 10.1038/nature12052.

[12] N. P. Blackledge, J. C. Zhou, M. Y. Tolstorukov, A. M. Farcas, P. J. Park, and R. J. Klose, "CpG Islands Recruit a Histone H3 Lysine 36 Demethylase," *Mol. Cell*, vol. 38, no. 2, pp. 179–190, 2010, doi: 10.1016/j.molcel.2010.04.009.

[13] J. M. Hancock, M. J. Zvelebil, and M. J. Zvelebil, "UniProt," in *Dictionary of Bioinformatics and Computational Biology*, 2004.

[14] R. Sorek, R. Shemesh, Y. Cohen, O. Basechess, G. Ast, and R. Shamir, "A non-EST-based method for exon-skipping prediction," *Genome Res.*, vol. 14, no. 8, pp. 1617–1623, 2004, doi: 10.1101/gr.2572604.

[15] National Center for Biotechnology Information, "RefSeq: NCBI Reference Sequence Database," *U.S. National Library of Medicine*, 2018.

https://www.ncbi.nlm.nih.gov/refseq/.

[16] F. Pendino *et al.*, "Functional involvement of RINF, retinoid-inducible nuclear factor (CXXC5), in normal and tumoral human myelopoiesis," *Blood*, vol. 113, no. 14, pp. 3172–3181, 2009, doi: 10.1182/blood-2008-07-170035.

[17] R. Zhang, Y. Wang, R. Li, and G. Chen, "Transcriptional factors mediating retinoic acid signals in the control of energy metabolism," *International Journal of Molecular Sciences*, vol. 16, no. 6. pp. 14210–14244, 2015, doi: 10.3390/ijms160614210.

[18] T. Andersson *et al.*, "CXXC5 is a novel BMP4-regulated modulator of Wnt signaling in neural stem cells," *J. Biol. Chem.*, vol. 284, no. 6, pp. 3672–3681, 2009, doi: 10.1074/jbc.M808119200.

[19] R. N. MacKinnon, G. Kannourakis, M. Wall, and L. J. Campbell, "A cryptic deletion in 5q31.2 provides further evidence for a minimally deleted region in myelodysplastic syndromes," *Cancer Genet.*, vol. 204, no. 4, pp. 187–194, 2011, doi: 10.1016/j.cancergen.2011.02.001.

[20] T. A. Graubert *et al.*, "Integrated genomic analysis implicates haploinsufficiency of multiple chromosome 5q31.2 genes in de novo myelodysplastic syndromes pathogenesis," *PLoS One*, vol. 4, no. 2, 2009, doi: 10.1371/journal.pone.0004583.

[21] W. Kleffmann *et al.*, "5q31 microdeletions: Definition of a critical region and analysis of LRRTM2, a candidate gene for intellectual disability," *Mol. Syndromol.*, vol. 3, no. 2, pp. 68–75, 2012, doi: 10.1159/000341252.

[22] E. Santoro, H. Pandolfo, J. Rosenfeld, and S. Raskin, "A 10.43 Mb duplication of chromosome region 5q31.2-q32 associated with a general delay in psychomotor development," *Meta Gene*, vol. 11, pp. 169–171, 2017, doi: 10.1016/j.mgene.2016.09.001.

[23] D. Kim and L. A. Garza, “The Negative Regulator CXXC5: Making WNT Look a Little Less Dishevelled,” *Journal of Investigative Dermatology*, vol. 137, no. 11. pp. 2248–2250, 2017, doi: 10.1016/j.jid.2017.07.826.

[24] S. H. Lee *et al.*, “The Dishevelled-binding protein CXXC5 negatively regulates cutaneous wound healing,” *J. Exp. Med.*, vol. 212, no. 7, pp. 1061–1080, 2015, doi: 10.1084/jem.20141601.

[25] S. Ma *et al.*, “Discovery of a small-molecule inhibitor of Dvl–CXXC5 interaction by computational approaches,” *J. Comput. Aided. Mol. Des.*, vol. 32, no. 5, pp. 643–655, 2018, doi: 10.1007/s10822-018-0118-x.

[26] K. Pulkkinen, S. Murugan, and S. Vainio, “Wnt signaling in kidney development and disease,” *Organogenesis*, vol. 4, no. 2. pp. 55–59, 2008, doi: 10.4161/org.4.2.5849.

[27] X. Peng *et al.*, “CXXC5 is required for cardiac looping relating to TGFβ signaling pathway in zebrafish,” *Int. J. Cardiol.*, vol. 214, pp. 246–253, 2016, doi: 10.1016/j.ijcard.2016.03.201.

[28] X. Yan, J. Wu, Q. Jiang, H. Cheng, J. D. J. Han, and Y. G. Chen, “CXXC5 suppresses hepatocellular carcinoma by promoting TGF-β-induced cell cycle arrest and apoptosis,” *J. Mol. Cell Biol.*, vol. 10, no. 1, pp. 48–59, 2018, doi: 10.1093/jmcb/mjx042.

[29] X. Xiong, S. Tu, J. Wang, S. Luo, and X. Yan, “CXXC5: A novel regulator and coordinator of TGF-β, BMP and Wnt signaling,” *Journal of Cellular and Molecular Medicine*, vol. 23, no. 2. pp. 740–749, 2019, doi: 10.1111/jcmm.14046.

[30] Z. Cui, Y. Shen, K. H. Chen, S. K. Mittal, J. Y. Yang, and G. J. Zhang, “KANK1 inhibits cell growth by inducing apoptosis though regulating CXXC5 in human malignant peripheral nerve sheath tumors,” *Sci. Rep.*, vol. 7, 2017, doi: 10.1038/srep40325.

[31] M. Y. Kim *et al.*, “CXXC5 plays a role as a transcription activator for myelin genes on oligodendrocyte differentiation,” *Glia*, vol. 64, no. 3, pp. 350–362, 2016, doi: 10.1002/glia.22932.

[32] H. Y. Kim *et al.*, “CXXC5 is a transcriptional activator of Flk-1 and mediates bone morphogenic protein-induced endothelial cell differentiation and vessel formation,” *FASEB J.*, vol. 28, no. 2, pp. 615–626, 2014, doi: 10.1096/fj.13-236216.

[33] F. Gagné, “Cellular Energy Allocation,” in *Biochemical Ecotoxicology: Principles and Methods*, 2014, pp. 131–144.

[34] S. Aras *et al.*, “Oxygen-dependent expression of cytochrome c oxidase subunit 4-2 gene expression is mediated by transcription factors RBPJ, CXXC5 and CHCHD2,” *Nucleic Acids Res.*, vol. 41, no. 4, pp. 2255–2266, 2013, doi: 10.1093/nar/gks1454.

[35] S. Ma *et al.*, “Epigenetic regulator CXXC5 recruits DNA demethylase Tet2 to regulate TLR7/9-elicited IFN response in pDCs,” *J. Exp. Med.*, vol. 214, no. 5, pp. 1471–1491, 2017, doi: 10.1084/jem.20161149.

[36] T. Kawai *et al.*, “Interferon-α induction through Toll-like receptors involves a direct interaction of IRF7 with MyD88 and TRAF6,” *Nat. Immunol.*, vol. 5, no. 10, pp. 1061–1068, 2004, doi: 10.1038/ni1118.

[37] Y. Tsuchiya *et al.*, “ ThPOK represses CXXC5, which induces methylation of histone H3 lysine 9 in Cd40lg promoter by association with SUV39H1: implications in repression of CD40L expression in CD8 + cytotoxic T cells ,” *J. Leukoc. Biol.*, vol. 100, no. 2, pp. 327–338, 2016, doi: 10.1189/jlb.1a0915-396rr.

[38] A. Kühnl *et al.*, “Downregulation of the Wnt inhibitor CXXC5 predicts a better prognosis in acute myeloid leukemia,” *Blood*, vol. 125, no. 19, pp. 2985–2994, 2015, doi: 10.1182/blood-2014-12-613703.

[39] M. Zhang *et al.*, “The CXXC finger 5 protein is required for DNA damage-induced p53 activation,” *Sci. China, Ser. C Life Sci.*, vol. 52, no. 6, pp. 528–538, 2009, doi: 10.1007/s11427-009-0083-7.

[40] G. Ayaz *et al.*, “CXXC5 as an unmethylated CpG dinucleotide binding protein contributes to estrogen-mediated cellular proliferation,” *Sci. Rep.*, vol. 10, no. 1, 2020, doi: 10.1038/s41598-020-62912-0.

[41] F. Spitz and E. E. M. Furlong, “Transcription factors: From enhancer binding to developmental control,” *Nature Reviews Genetics*, vol. 13, no. 9. pp. 613–626, 2012, doi: 10.1038/nrg3207.

[42] V. Haberle and A. Stark, “Eukaryotic core promoters and the functional basis of transcription initiation,” *Nature Reviews Molecular Cell Biology*, vol. 19, no. 10. pp. 621–637, 2018, doi: 10.1038/s41580-018-0028-8.

[43] R. J. Sims, S. S. Mandal, and D. Reinberg, “Recent highlights of RNA-polymerase-II-mediated transcription,” *Current Opinion in Cell Biology*, vol. 16, no. 3. pp. 263–271, 2004, doi: 10.1016/j.ceb.2004.04.004.

[44] A. J. M. Walhout, “Gene-Centered Regulatory Network Mapping,” in *Methods in Cell Biology*, vol. 106, 2011, pp. 271–288.

[45] D. Shlyueva, G. Stampfel, and A. Stark, “Transcriptional enhancers: From properties to genome-wide predictions,” *Nature Reviews Genetics*, vol. 15, no. 4. pp. 272–286, 2014, doi: 10.1038/nrg3682.

[46] B. Lenhard, A. Sandelin, and P. Carninci, “Metazoan promoters: Emerging characteristics and insights into transcriptional regulation,” *Nature Reviews Genetics*, vol. 13, no. 4. pp. 233–245, 2012, doi: 10.1038/nrg3163.

[47] P. Gagniuc and C. Ionescu-Tirgoviste, “Eukaryotic genomes may exhibit up to 10 generic classes of gene promoters,” *BMC Genomics*, vol. 13, no. 1, 2012, doi: 10.1186/1471-2164-13-512.

[48] S. T. Smale and J. T. Kadonaga, “The RNA Polymerase II Core Promoter,”

Annu. Rev. Biochem., vol. 72, no. 1, pp. 449–479, 2003, doi: 10.1146/annurev.biochem.72.121801.161520.

[49] J. Dejong, R. Bernstein, and R. G. Roeder, "Human general transcription factor TFIIA: Characterization of a cDNA encoding the small subunit and requirement for basal and activated transcription," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 8, pp. 3313–3317, 1995, doi: 10.1073/pnas.92.8.3313.

[50] J. Eichner, H. T. Chen, L. Warfield, and S. Hahn, "Position of the general transcription factor TFIIF within the RNA polymerase II transcription preinitiation complex," *EMBO J.*, vol. 29, no. 4, pp. 706–716, 2010, doi: 10.1038/emboj.2009.386.

[51] D. S. Luse, "The RNA polymerase II preinitiation complex," *Transcription*, vol. 5, no. 1, p. e27050, 2014, doi: 10.4161/trns.27050.

[52] J.-P. Hsin, K. Xiang, and J. L. Manley, "Function and Control of RNA Polymerase II C-Terminal Domain Phosphorylation in Vertebrate Transcription and RNA Processing," *Mol. Cell. Biol.*, vol. 34, no. 13, pp. 2488–2498, 2014, doi: 10.1128/mcb.00181-14.

[53] R. S. Illingworth and A. P. Bird, "CpG islands - 'A rough guide,'" *FEBS Letters*, vol. 583, no. 11. pp. 1713–1720, 2009, doi: 10.1016/j.febslet.2009.04.012.

[54] S. Saxonov, P. Berg, and D. L. Brutlag, "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 5, pp. 1412–1417, 2006, doi: 10.1073/pnas.0510310103.

[55] F. Asmar, A. Søgaard, and K. Grønbæk, "DNA Methylation and Hydroxymethylation in Cancer," in *Epigenetic Cancer Therapy*, 2015, pp. 9–30.

[56] T. Vavouri and B. Lehner, "Human genes with CpG island promoters have a

distinct transcription-associated chromatin organization," *Genome Biol.*, vol. 13, no. 11, 2012, doi: 10.1186/gb-2012-13-11-r110.

[57] J. Zhu, F. He, S. Hu, and J. Yu, "On the nature of human housekeeping genes," *Trends in Genetics*, vol. 24, no. 10. pp. 481–484, 2008, doi: 10.1016/j.tig.2008.08.004.

[58] N. Cvetesic *et al.*, "SLIC-CAGE: high-resolution transcription start site mapping using nanogram-levels of total RNA," *SLIC-CAGE High-resolution Transcr. start site Mapp. using nanogram-levels Total RNA*, p. 368795, 2018, doi: 10.1101/368795.

[59] A. M. Deaton and A. Bird, "CpG islands and the regulation of transcription," *Genes Dev.*, vol. 25, no. 10, pp. 1010–1022, 2011, doi: 10.1101/gad.2037511.

[60] Z. Wang *et al.*, "Combinatorial patterns of histone acetylations and methylations in the human genome," *Nat. Genet.*, vol. 40, no. 7, pp. 897–903, 2008, doi: 10.1038/ng.154.

[61] E. Birney *et al.*, "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project," *Nature*, vol. 447, no. 7146, pp. 799–816, 2007, doi: 10.1038/nature05874.

[62] F. M. Shapter and D. L. E. Waters, "Genome walking," *Methods Mol. Biol.*, vol. 1099, pp. 133–146, 2014, doi: 10.1007/978-1-62703-715-0_12.

[63] K. J. Livak and T. D. Schmittgen, "Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method," *Methods*, vol. 25, no. 4, pp. 402–408, 2001, doi: 10.1006/meth.2001.1262.

[64] J. Y. Ryu, H. U. Kim, and S. Y. Lee, "Human genes with a greater number of transcript variants tend to show biological features of housekeeping and essential genes," *Mol. Biosyst.*, vol. 11, no. 10, pp. 2798–2807, 2015, doi: 10.1039/c5mb00322a.

[65] A. G. Joshi and A. R. Pathak, “EST (Expressed Sequence Tag): A Technique for Identification of Plant Secondary Metabolite Genes,” in *Plant and Human Health, Volume 2*, 2019, pp. 207–225.

[66] GTEx Project, “GTEx portal,” *GTEx Analysis Release V6p (dbGaP Accession phs000424.v6.p1)*. 2017, doi: 10.1038/86850.

[67] N. Silver, S. Best, J. Jiang, and S. L. Thein, “Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR,” *BMC Mol. Biol.*, vol. 7, 2006, doi: 10.1186/1471-2199-7-33.

[68] M. W. Pfaffl, “Relative quantification,” in *Real-time PCR*, 2007, pp. 64–82.

[69] O. Yeku and M. A. Frohman, “Rapid amplification of cDNA ends (RACE).,” *Methods Mol. Biol.*, vol. 703, pp. 107–122, 2011, doi: 10.1007/978-1-59745-248-9_8.

APPENDICES

A. Genomic DNA (gDNA) Contamination Control PCR of MCF7 and HL60 RNAs

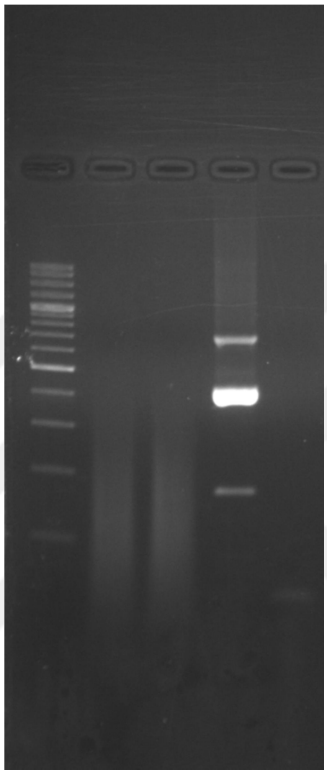


Figure 5.10. gDNA Contamination control PCR of MCF7 and HL60 RNAs was carried out with primers specific for *GAPDH* gene locus. Only positive control reaction with genomic DNA gave products (Third well after size marker).

B. MCF7 and HL60 cDNA PCR

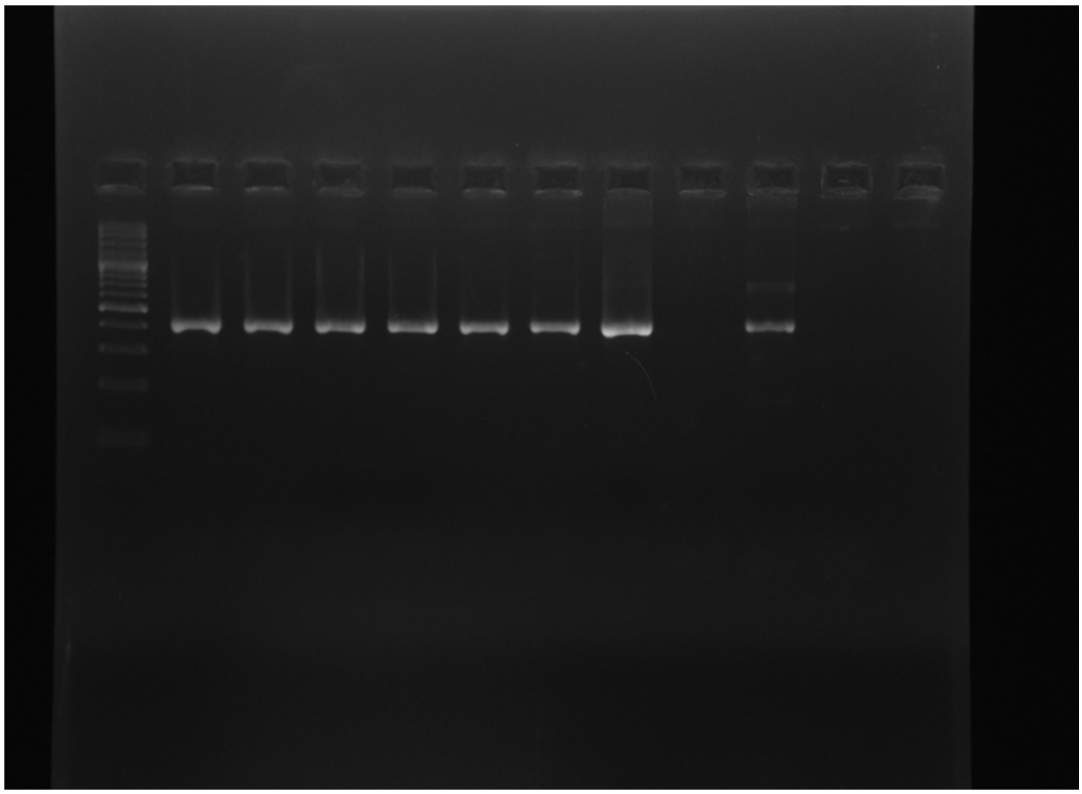


Figure 5.11. MCF7 and HL60 cDNA PCR with *GAPDH* specific primers.

C. MIQE Checklist

Table 5.6 MIQE Checklist

ITEM TO CHECK	IMPORTANCE	CHECKLIST
EXPERIMENTAL DESIGN		
Definition of experimental and control groups	E	YES
Number within each group	E	YES
Assay carried out by core lab or investigator's lab?	D	YES
Acknowledgement of authors' contributions	D	N/A
SAMPLE		
Description	E	N/A
Volume/mass of sample processed	D	N/A
Microdissection or macrodissection	E	N/A
Processing procedure	E	N/A
If frozen - how and how quickly?	E	N/A
If fixed - with what, how quickly?	E	N/A
Sample storage conditions and duration (especially for FFPE samples)	E	N/A
NUCLEIC ACID EXTRACTION		
Procedure and/or instrumentation	E	YES
Name of kit and details of any modifications	E	YES
Source of additional reagents used	D	N/A
Details of DNase or RNAse treatment	E	YES
Contamination assessment (DNA or RNA)	E	YES

Table 5.6 (cont'd) MIQE Checklist

Nucleic acid quantification	E	YES
Instrument and method	E	YES
Purity (A260/A280)	D	YES
Yield	D	NO
RNA integrity method/instrument	E	YES
RIN/RQI or Cq of 3' and 5' transcripts	E	YES
Electrophoresis traces	D	YES
Inhibition testing (Cq dilutions, spike or other)	E	YES
REVERSE TRANSCRIPTION		
Complete reaction conditions	E	YES
Amount of RNA and reaction volume	E	YES
Priming oligonucleotide (if using GSP) and concentration	E	YES
Reverse transcriptase and concentration	E	YES
Temperature and time	E	YES
Manufacturer of reagents and catalogue numbers	D	YES
Cqs with and without RT	D*	N/A
Storage conditions of cDNA	D	YES
qPCR TARGET INFORMATION		
If multiplex, efficiency and LOD of each assay.	E	N/A
Sequence accession number	E	YES
Location of amplicon	D	YES
Amplicon length	E	NO
In silico specificity screen (BLAST, etc)	E	NO
Pseudogenes, retropseudogenes or other homologs?	D	YES

Table 5.6 (cont'd) MIQE Checklist

Sequence alignment	D	YES
Secondary structure analysis of amplicon	D	NO
Location of each primer by exon or intron (if applicable)	E	YES
What splice variants are targeted?	E	YES
qPCR OLIGONUCLEOTIDES		
Primer sequences	E	YES
RTPrimerDB Identification Number	D	N/A
Probe sequences	D**	N/A
Location and identity of any modifications	E	N/A
Manufacturer of oligonucleotides	D	NO
Purification method	D	NO
qPCR PROTOCOL		
Complete reaction conditions	E	YES
Reaction volume and amount of cDNA/DNA	E	YES
Primer, (probe), Mg++ and dNTP concentrations	E	N/A
Polymerase identity and concentration	E	N/A
Buffer/kit identity and manufacturer	E	YES
Exact chemical constitution of the buffer	D	N/A
Additives (SYBR Green I, DMSO, etc.)	E	YES
Manufacturer of plates/tubes and catalog number	D	NO
Complete thermocycling parameters	E	YES
Reaction setup (manual/robotic)	D	YES
Manufacturer of qPCR instrument	E	YES

Table 5.6 (cont'd) MIQE Checklist

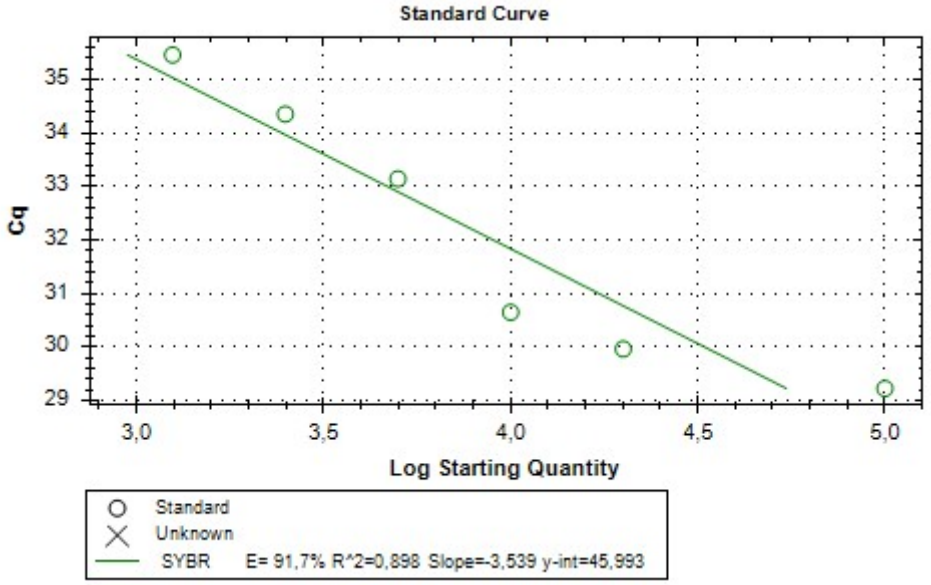
qPCR VALIDATION		
Evidence of optimisation (from gradients)	D	NO
Specificity (gel, sequence, melt, or digest)	E	YES
For SYBR Green I, Cq of the NTC	E	YES
Standard curves with slope and y-intercept	E	YES
PCR efficiency calculated from slope	E	YES
Confidence interval for PCR efficiency or standard error	D	NO
r2 of standard curve	E	YES
Linear dynamic range	E	YES
Cq variation at lower limit	E	YES
Confidence intervals throughout range	D	N/A
Evidence for limit of detection	E	NO
If multiplex, efficiency and LOD of each assay.	E	N/A
DATA ANALYSIS		
qPCR analysis program (source, version)	E	YES
Cq method determination	E	YES
Outlier identification and disposition	E	N/A
Results of NTCs	E	YES
Justification of number and choice of reference genes	E	YES
Description of normalisation method	E	YES
Number and concordance of biological replicates	D	YES
Number and stage (RT or qPCR) of technical replicates	E	YES
Repeatability (intra-assay variation)	E	YES

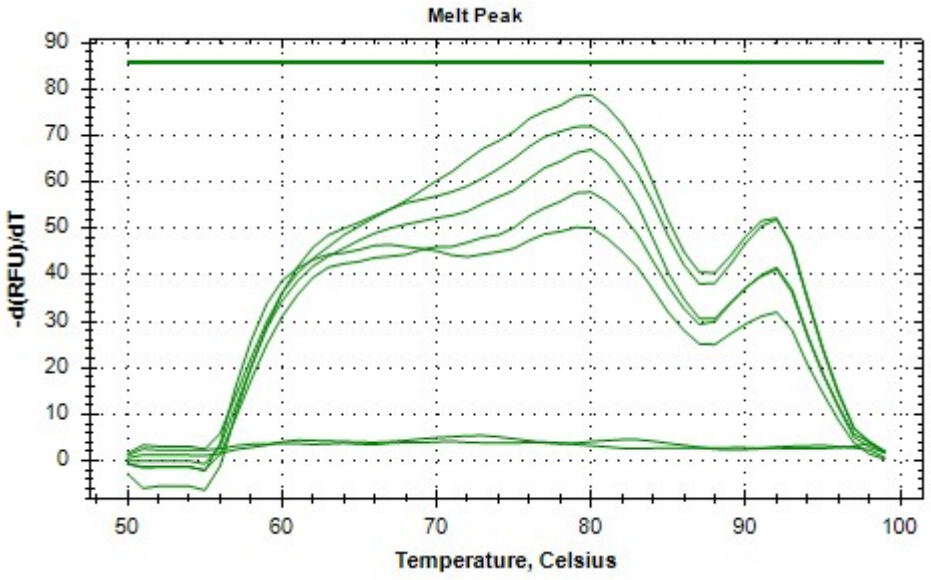
Table 5.6 (cont'd) MIQE Checklist

Reproducibility (inter-assay variation, %CV)	D	YES
Power analysis	D	NO
Statistical methods for result significance	E	YES
Software (source, version)	E	YES
Cq or raw data submission using RDML	D	N/A

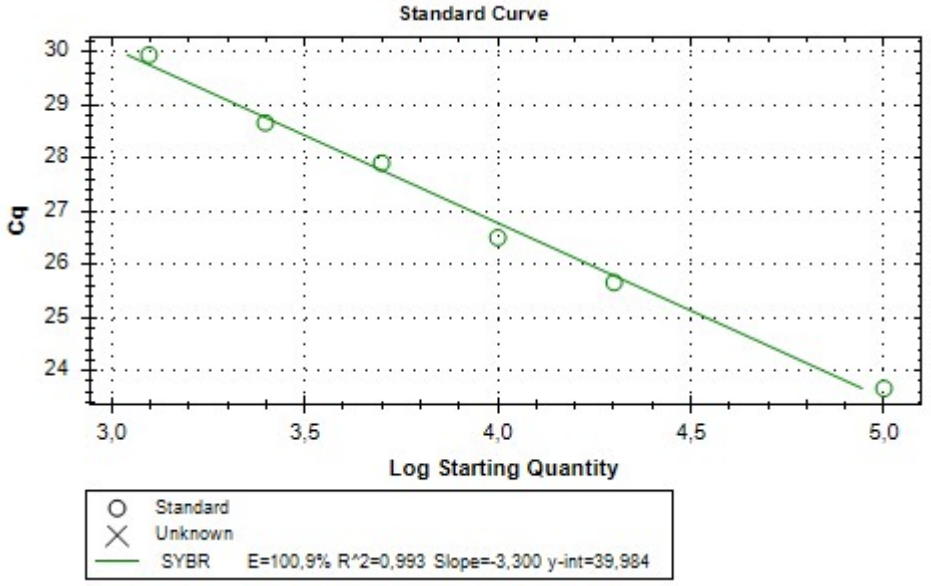
D. Standard Curve and Melt Curve of qPCR Reactions

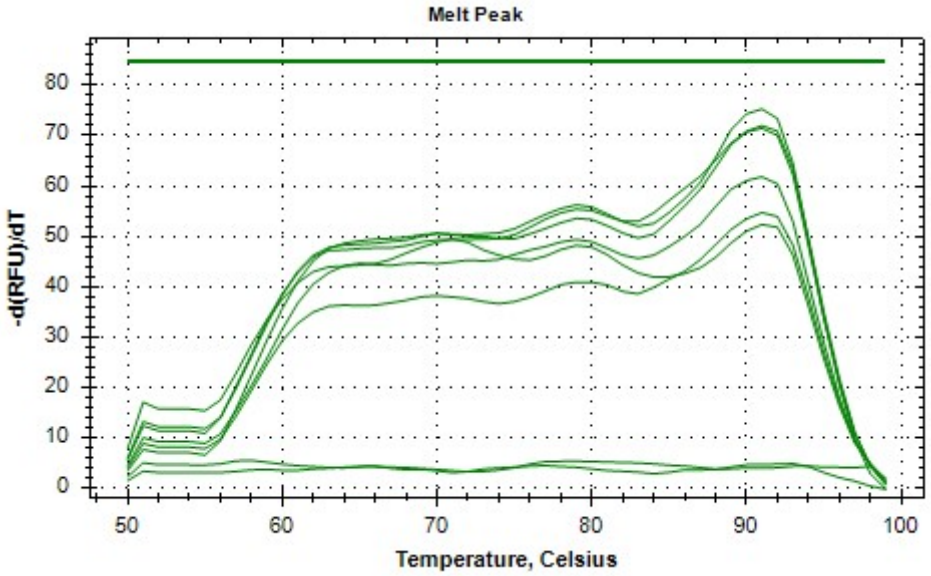
Transcript Variant 1 Standard Curve and Melt Curve



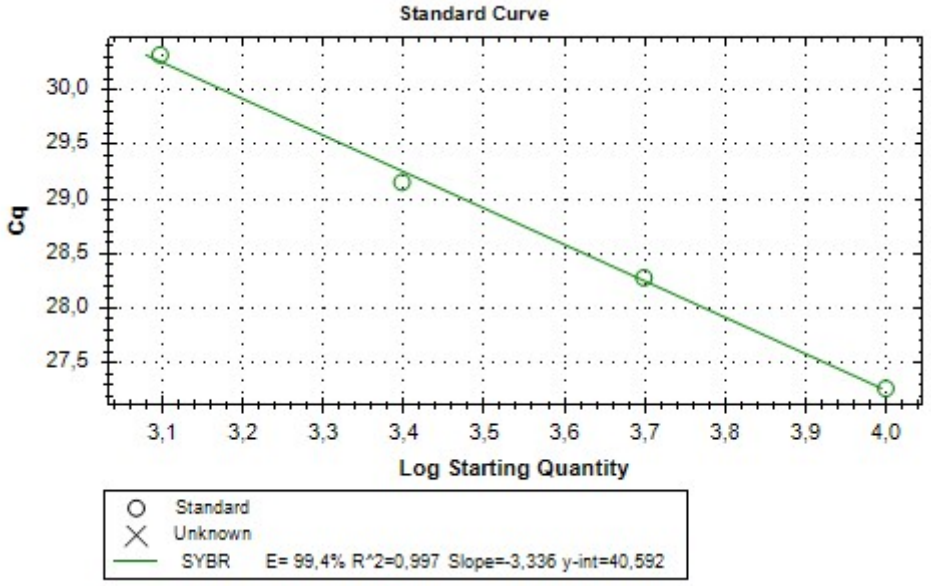


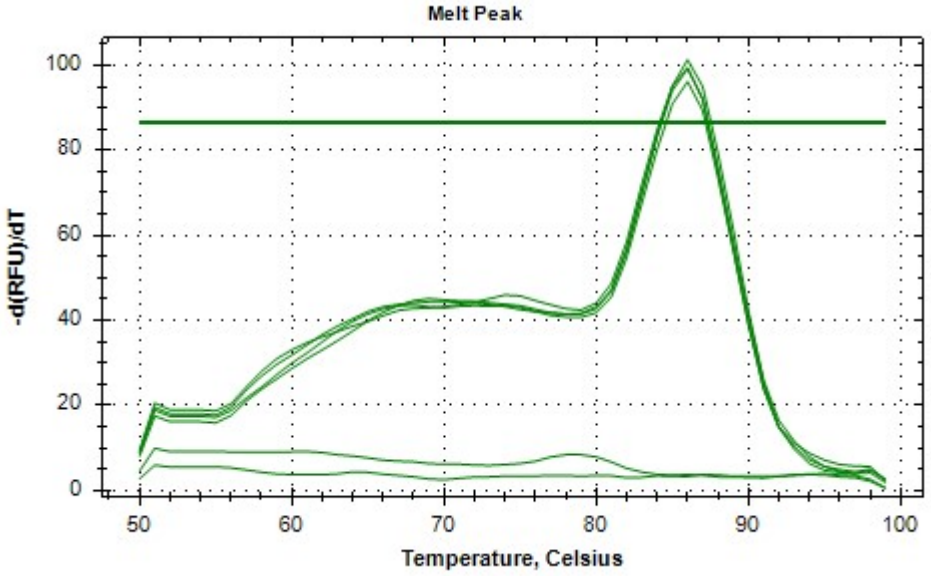
Transcript Variant 2 Standard Curve and Melt Curve



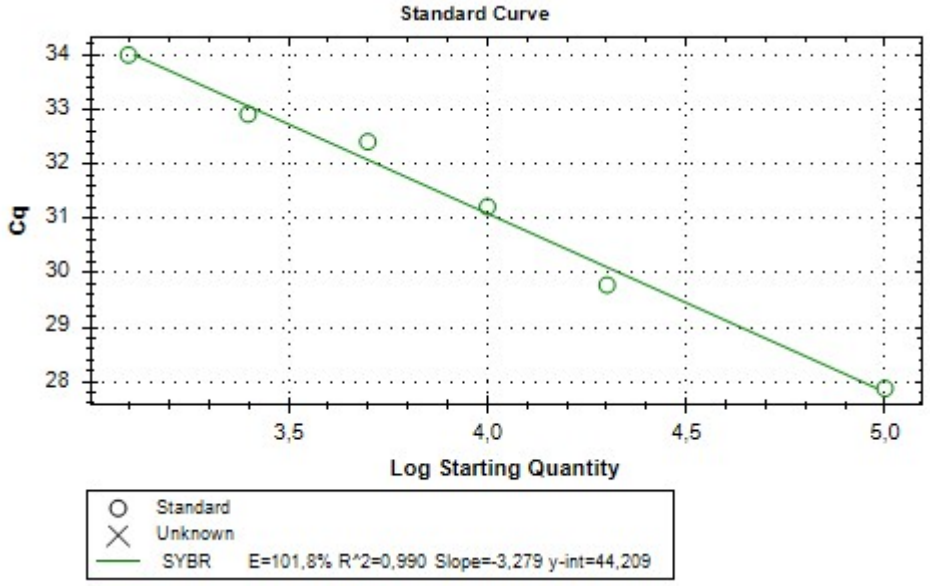


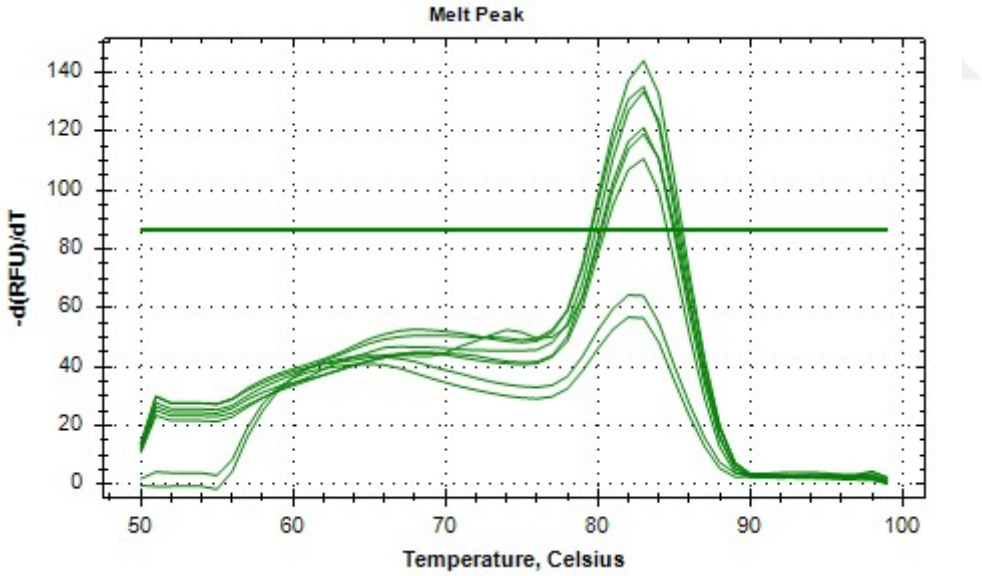
Transcript Variant 3 Standard Curve and Melt Curve



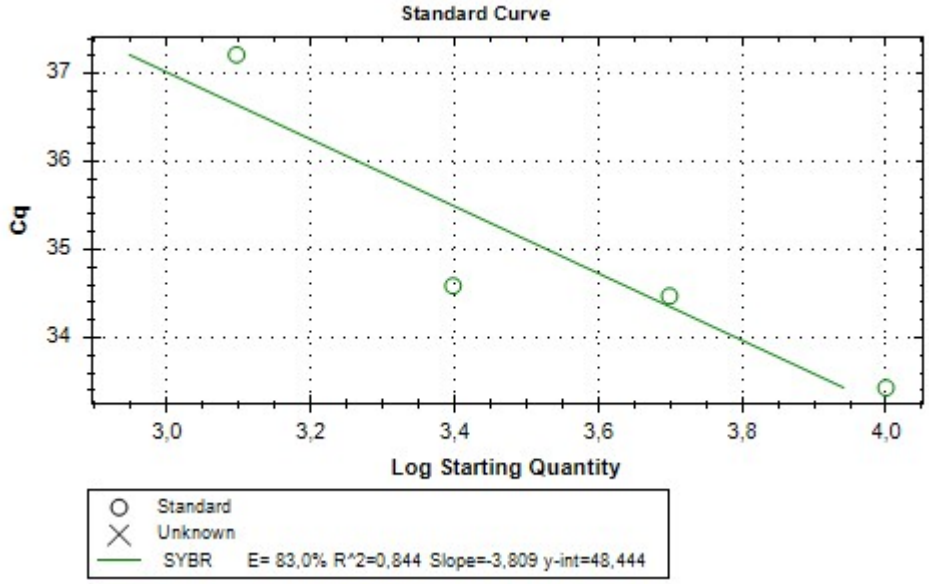


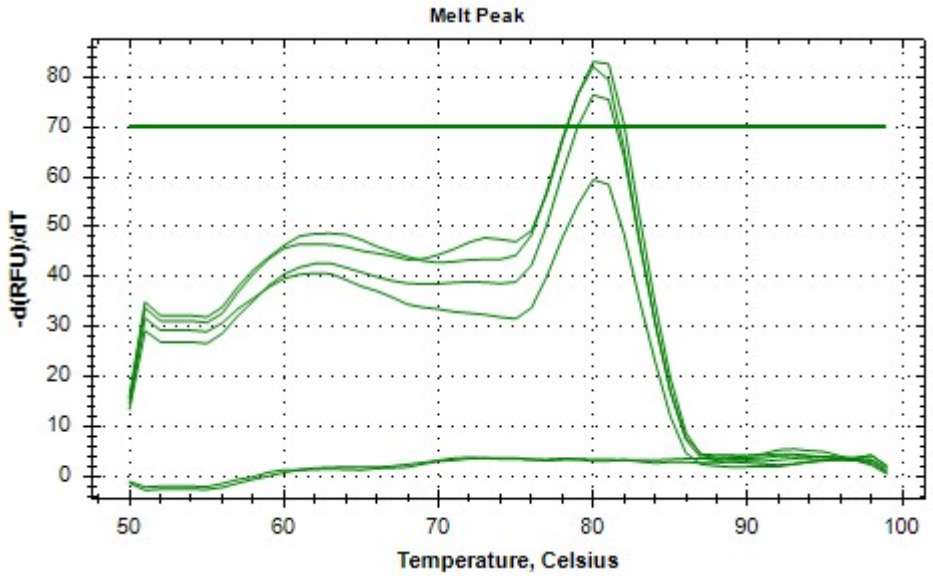
Transcript Variant 4 Standard Curve and Melt Curve



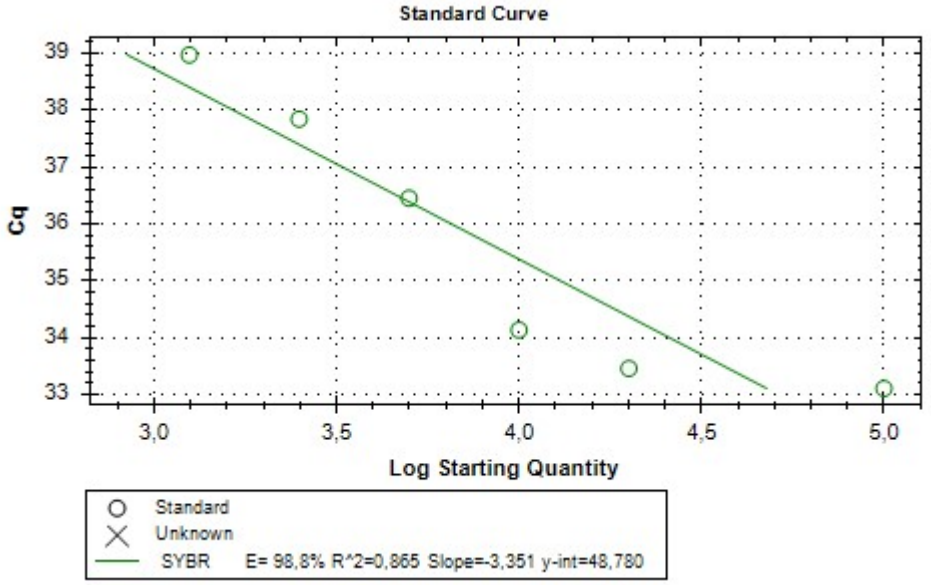


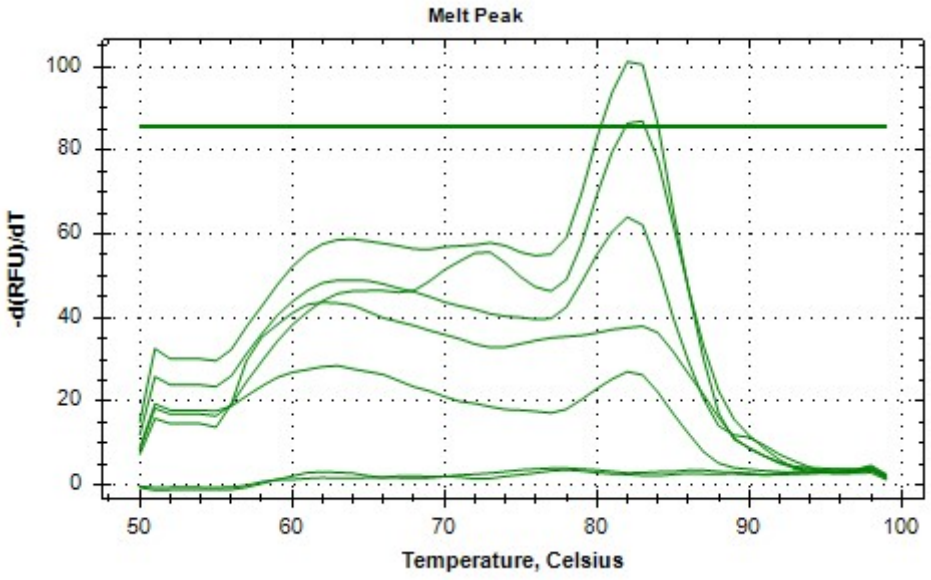
Transcript Variant 5 Standard Curve and Melt Curve



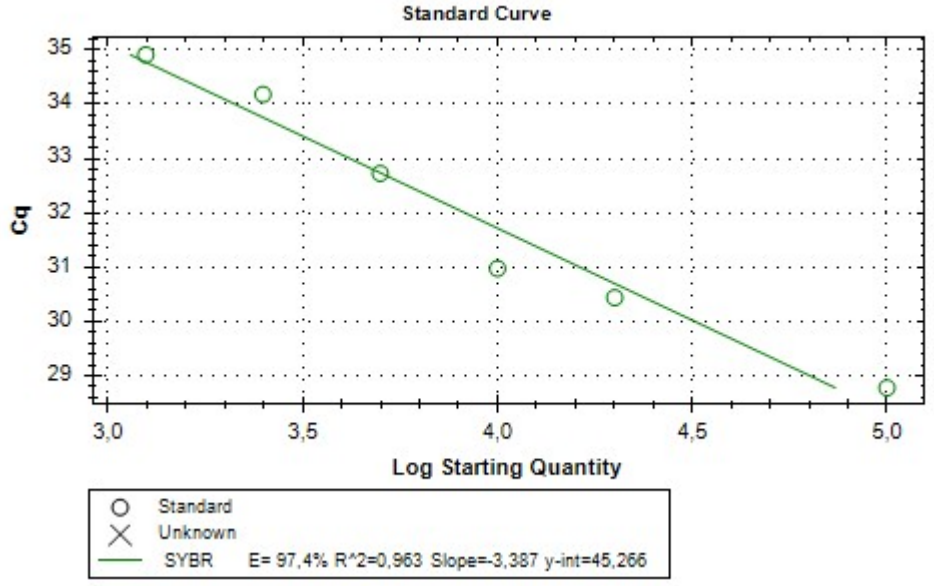


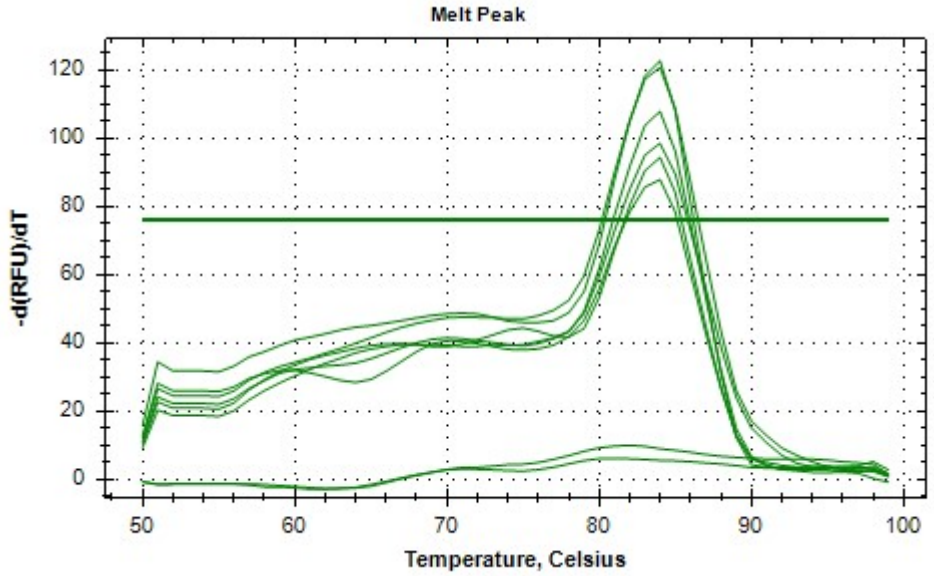
Transcript Variant 6 Standard Curve and Melt Curve



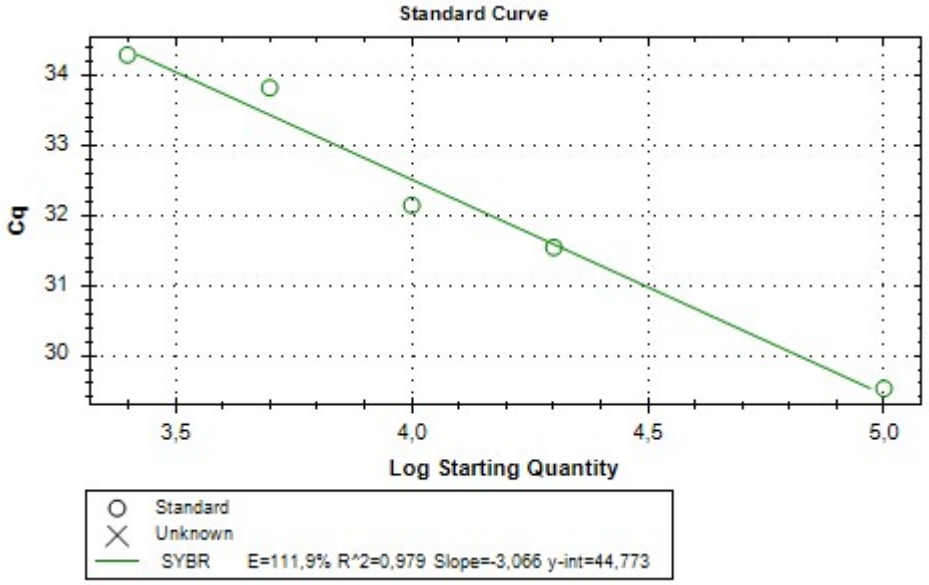


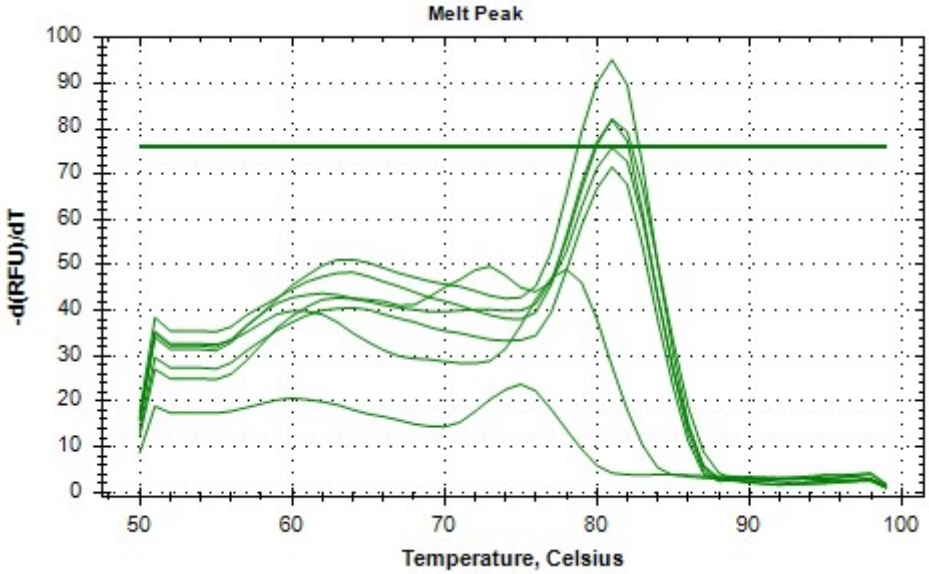
Transcript Variant 7 Standard Curve and Melt Curve





Transcript Variant 8 Standard Curve and Melt Curve





Transcript Variant 10 Standard Curve and Melt Curve