THE REGULATION OF THE CXXC5 GENE EXPRESSION

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

BY

PELİN YAŞAR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
MOLECULER BIOLOGY AND GENETICS

JANUARY 2021

Approval of the thesis:

THE REGULATION OF THE CXXC5 GENE EXPRESSION

submitted by **PELİN YAŞAR** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** 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:

Prof. Dr. Ayşe Elif Erson Bensan
Biological Sciences, METU _____

Prof. Dr. Mesut Muyan
Biological Sciences, METU _____

Prof. Dr. Tolga Can
Computer Engineering, METU _____

Prof. Dr. Yusuf Çetin Kocaefe
Medical Biology, Hacettepe Uni. _____

Assoc. Prof. Dr. Bala Gür Dedeoğlu
Biotechnology, Ankara Uni. _____

Date: 19.01.2021

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 : Pelin, Yaşar

Signature :

ABSTRACT

THE REGULATION OF THE CXXC5 GENE EXPRESSION

Yaşar, Pelin
Doctor of Philosophy, Molecular Biology and Genetics
Supervisor : Prof. Dr. Mesut Muyan

January 2021, 86 pages

17β-estradiol (E2) is the main circulating estrogen hormone in the body and is involved in the physiological and pathophysiological regulation of various tissue notably mammary tissue functions. E2 is responsible for cellular proliferation, differentiation, and/or death in target tissues. Our previous microarray studies suggested that expression of *CXXC5* is regulated by E2-ERα through ERE-dependent signaling pathway and I verified that the *CXXC5* transcript levels are augmented in response to E2. As a member of the ZF-CXXC domain protein family, CXXC5 harbors a highly conserved CXXC domain and nuclear localization signal. It is known that functionally characterized ZF-CXXC domain protein family members bind to preferentially non-methylated CpG dinucleotides in CpG islands of transcriptionally active DNA regions via their CXXC domains, and we showed that CXXC5 also binds to non-methylated CpG dinucleotides. Cytosine methylation is prevented due to this binding, and a nucleation site formation is induced for the direct/indirect recruitment of transcription co-regulators, histone-modifying proteins, which leads to the regulation of transcription. According to the limited studies on CXXC5, it appears to be that CXXC5 participates in the cellular events as an epigenetic regulator and/or co-modulator in response to various

signaling pathways. However, the mechanisms of how the CXXC5 gene expression is mediated remain unknown.

In my doctoral studies, I showed the expression and the synthesis of CXXC5 is E2- and ERα-dependent. I found that an intronic ERE sequence in the *CXXC5* locus, and I showed that this *de novo* ERE is bound by ERα *in vitro* and *in cellula*. In addition, I showed that the binding of the E2-ERα complex is functional and resulted in the transcriptional activation using reporter enzyme assays. To understand how this distally located ERE participates in the regulation of *CXXC5* expression regulation; first I wanted to identify the *CXXC5* promoter. Since there are 14 annotated transcript variants for *CXXC5*, I foresaw that the identification of the main *CXXC5* variant(s) expressed at the highest amount in MCF7 cells as our cell model is necessary for the promoter analyses studies. Therefore, after the identification of the primary *CXXC5* transcript (CXXC5-TV2), I conducted 5'Rapid Amplification of cDNA Ends (5'RACE) studies to uncover the transcription start site(s) (TSSs) of the CXXC5-TV2 to characterize the core promoter elements which are generally found in proximity to the TSSs. Next, I performed luciferase reporter assays to locate the core promoter elements of the *CXXC5*. I showed that the *CXXC5* promoter region resides within the first exon of the CXXC5-TV2, and the expression of *CXXC5* is driven by a CGI promoter (CpG Island Promoter). After identifying the promoter region, I continued with promoter pull-down studies to characterize the putative *CXXC5* promoter interactor proteins. Of the identified proteins by Liquid chromatography-tandem mass spectrometry (LC-MS/MS), I validated the binding of the ELF1 (E74 Like ETS Transcription Factor 1), MAZ (Myc-associated zinc finger protein), and RB1 (Retinoblastoma-associated protein) to the *CXXC5* promoter. I also found the sequence motifs for DNA binding of ELF1 and MAZ in the *CXXC5* promoter. I verified that ELF1 and MAZ contribute to the regulation of *CXXC5* expression using endogenous and heterologous gene expression approaches.

In summary, I found that the expression, and consequently, the synthesis of CXXC5 are regulated by E2-ERα signaling through an intronic ERE. I located the *CXXC5* promoter region and identified the several transcription factors engaged with the promoter, and verify that these transcription factors are involved in the gene expression regulation of *CXXC5*. The findings presented in this dissertation could provide a basis for understanding the regulation of an E2-responsive gene CXXC5, therefore, could provide a better understanding of the E2-ER signaling actions in physiological and/or pathophysiological conditions.

Keywords: Estrogen Signaling, *CXXC5*, Promoter, CpG island, Expression Regulation

ÖZ

CXXC5 GEN EKSRESYONUNUN DÜZENLENMESİ

Yaşar, Pelin
Doktora, Moleküler Biyoloji ve Genetik
Tez Yöneticisi: Prof. Dr. Mesut Muyan

Ocak 2021, 86 sayfa

Vücutta dolaşımda bulunan ana östojen hormonu 17β-estradiol (E2)'dur ve başta meme dokusu fonksiyonları olmak üzere, farklı dokuların fizyolojik ya da patofizyolojik durumlardaki düzenlenmesinde görev alır. E2 hedef dokularındaki, hücresel çoğalma, farklılaşma ve/veya ölüm görevlerinden sorumludur. Önceki mikrodizin çalışmalarımız, *CXXC5* ifadesinin E2-ERα ile düzenlendiğini ve bunun ERE-bağımlı sinyal yolağı üzerinden gerçekleştiğini önermekteydi ve biz de *CXXC5* transkript seviyelerinin E2'ye yanıtla arttığını doğruladık. CXXC5 bir ZF-CXXC5 domain protein ailesi üyesi olarak yüksekçe korunmuş olan CXXC domaini ve nükleer lokalizasyon sinyali içermektedir. Fonksiyonel olarak karakterize edilmiş olan ZF-CXXC protein ailesi üyeleri, CXXC domainleri aracılığıyla; tercihen transkripsiyonel olarak aktif olan DNA bölgelerindeki CpG adalarının, metile-olmamış CpG dinükleotitlerine bağlanmaktadırlar; biz de CXXC5'in de metile-olmamış CpG dinükleotitlere bağlanan bir protein olduğunu gösterdik. Bu bağlanma sayesinde sitozin metilasyonu engellenerek, tranksripsiyon düzenlenmesinde görev alan histon modifiye edici proteinlerin getirilmesi için, doğrudan ya da dolaylı olarak kullanılan bir çekirdeklenme alanı oluşturulur. CXXC5 ile ilgili yapılmış olan kısıtlı çalışmaya göre CXXC5; farklı sinyallere

yanıtla, epigenetik düzenleyici ve/veya eş-düzenleyici olarak görev almaktadır. Ancak CXXC5 geninin ekspresyonunun nasıl sağlandığı ile ilgili mekanizmalar hala bilinmemektedir.

Doktora çalışmalarımda, CXXC5 proteininin sentezlenmesinin tıpkı *CXXC5* ifadesinde olduğu gibi E2-yanıtlı ve ERα bağımlı olduğunu gösterdim. CXXC5 lokusunda yer alan bir *de novo* ERE buldum ve ERα'nın hem *in vitro* hem de *in cellula* olarak bu ERE'ye bağlandığını gösterdim. Buna ek olarak, reportör enzim tahlilleri kullanarak, E2-ERα kompleksinin bu ERE'ye bağlanmasının fonksiyonel olduğunu ve trankripsiyonel bir aktivasyon sağladığını bulguladım. Sonrasında bu distal olarak konumlanmış olan ERE'nin *CXXC5* ifadesinin düzenlenmesine nasıl katkıda bulunduğunu anlayabilmek için *CXXC5* promotörünü bulgulamak üzere çalışmalarıma devam ettim. *CXXC5*'in tanımlanmış olan 14 transkript varyantı olduğundan, model hücre hattımız olan MCF7'lerde hangi *CXXC5* varyant(lar)ının daha çok bulunduğunun tespit edilmesinin *CXXC5* promoter çalışmaları için gerekli olduğuna karar verdim. Bu sebeple ana CXXC5 varyatını (CXXC5-TV2) tespit ettikten sonra, öz promotör elemanları genellikle transkripsiyon başlangıç noktaları (TSSs) civarında yer almakta olduğu için; bu varyanta yönelik olarak 5'Rapid Amplification of cDNA Ends (5'RACE, cDNA sonlarının hızlı olarak çoğaltılması) yaklaşımını kullanarak transkripsiyon başlangıç noktalarını belirledim. Sonrasında lusiferaz raportör tahlilleri kullanarak *CXXC5* öz promotör elemanlarının yerini saptadım. *CXXC5* promotörünün CXXC5-TV2'nin ilk eksonu içinde yer aldığını ve *CXXC5*'in bir CGI promotör (CpG Island Promoter, CpG Adacık Promotör) tarafından düzenlendiğini bulguladım. Promotör bölgesini tespit ettikten sonra, *CXXC5* promotörü ile etkileşime geçen proteinleri tespit edebilmek amacıyla promotör çöktürme çalışmaları gerçekleştirdim. Sıvı kromatografisi-ardışık kütle spektrometresi (LC-MS/MS) ile tespit edilen proteinlerden, ELF1 (E74 Like ETS Transcription Factor 1), MAZ (Myc-associated zinc finger protein), and RB1 (Retinoblastoma-associated protein) proteinlerinin *CXXC5* promotörüne *in vitro* ve *in cellula* olarak bağlandığını gösterdim. Ayrıca *CXXC5* promotörü

içerisinde; ELF1 ve MAZ transkripsiyon faktörlerinin bağlanma motif sekanslarının yerini tespit ettim. Buna ek olarak, ELF1 ve MAZ proteinlerinin *CXXC5* ekspresyonunun düzenlenmesine dahil olduklarını arttırma (overexpression) ve azaltma (knocking-down) çalışmaları ile gösterdim.

CXXC5 ekspresyonunun ve buna bağlı olarak protein sentezinin, E2-ERα sinyali tarafından, intronik bir bölgede yer alan ERE üzerinden düzenlendiğini gösterdim. *CXXC5* promotör bölgesini bulguladım ve promotör ile etkileşime geçen birkaç transkripsiyon faktörünü göstererek, bunların *CXXC5* ifadesinin düzenlenmesine dahil olduklarını tespit ettim. Tezimde gösterdiğim bulgular, E2-yanıtlı olan CXXC5 geninin nasıl düzenlendiğinin anlaşılabilmesi için bir temel oluşturacaktır. Buna bağlı olarak da E2-ER sinyal yolağının fizyolojik ve patofizyolojik koşullardaki aksiyonlarının daha iyi anlaşılmasını sağlayabilecektir.

Anahtar Kelimeler: Östrojen Sinyali, *CXXC5*, Promotör, CpG Adaları, Ekspresyon Düzenlenmesi

To the charm of chasing dreams and their realizations

ACKNOWLEDGEMENTS

First of all I would like to thank my advisor Prof. Dr. Mesut Muyan for his endless patience, and supports through my journey in the scientific community from the initial steps to today. The things that I've learned from his scientific vision and life perspective took me forward and provided a self-confidency and spirited resolution in me for present and also for the future. I am heartily thankful for his invaluable guidance and efforts for all the times.

I wanted to express my sincere gratitude to Prof. Dr. A. Elif Erson Bensan for her frank supports and educatory guidance. I also wanted to thank Prof. Dr. Tolga Can for his guidance in bioinformatics and his kind efforts throuhout my thesis studies. I would like thank to my thesis committee members Prof. Dr. Yusuf Çetin Kocaefe, Assoc. Prof. Dr. Bala Gür Dedeoğlu.

I owe a great many thanks to my lab mates for their kindness and friendship. I am very grateful to Gamze Ayaz, Kerim Yavuz, Gizem Kars, Çağla Ece Olgun, Gizem Turan, and Öykü Deniz Demiralay for their all efforts and supports. I also want to express my thanks Hazal Ayten her tireless works and kindness and Edanur Şen for friendly supports. In addition, I want to thank all the members of Erson-Bensan Lab, Banerjee Lab and Gürsel Lab for their support and help.

My deep thanks to my dear friends Safiye, Gözde, and Süleyman for their support everytime that I need. Finally, I would like to thank to my family to be there for me all the times, my lovely sister Aylin and parents Füsun and Hüseyin.

I would like to thank TÜBİTAK for supporting this study through 114Z243 and 118Z957 grants.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS

E2: 17β-Estradiol

ICI 182,780, ICI: Imperial Chemical Industries 182,780

CXXC5: Zinc-Finger CXXC5 domain protein 5

CHAPTER 1

INTRODUCTION

1.1 E2 Signaling

17β-estradiol, E2, is the essential circulating estrogen hormone and is involed in the regulation of the physiological and pathophysiological functions in the target tissue and organs, notably breast tissue [1]. Breast cancer is a complex disease comprising different subtypes and the genetic background, hormonal and/or environmental elements contribute to the incidences [2]. However, one critical factor is the E2 signaling that causes uncontrolled cellular growth in the breast epithelial cells leading to the initiation/progression of breast cancer [3], [4].

E2 exerts its effects through ligand-dependent nuclear hormone receptors, estrogen receptor α and β (ERα and β). ERα and β are transcription factors and products of distinct genes, *ESR1* and *ESR2*, which display a high degree of homology in structural and functional features. While ERs are synthesized at various levels in different tissues, ERα is the main ER expressed in breast tissue [1], [3]. Upon binding of E2, ERα undergoes major conformational changes, which result in the activation of ERα and led to the formation of the interaction surfaces for the recruitment of the co-regulatory proteins and basal transcription machinery [5], [6].

Activated ERα modulates the regulation of the target gene expressions through two distinct genomic signaling pathways: the estrogen response element (ERE)-dependent and ERE-independent signaling pathways. EREs arise from the derivatives of the DNA sequence 5’- GGTCA nnn TGACC -3' which is a palindromic sequence separated by three non-specific nucleotides [5]. In the ERE-dependent signaling pathway, the E2-ER complex directly binds to EREs on the DNA and regulates the target gene expressions. On the other hand, in the ERE-

independent signaling pathway, the E2-ER complex binds to the other transcription factors that are already bound to their cognate response elements and participates in gene expression regulation [3], [7], [8].

In a microarray study previously performed in our laboratory, *CXXC5* was suggested to be an E2-responsive gene [8]. In my master thesis studies, I showed that the expression levels of *CXXC5* are augmented in the presence of E2 at 3, 6, and 24 hours which indicates that *CXXC5* is indeed an E2-regulated gene (Pelin Yaşar, Initial Characterization of CXXC5 as a Putative DNA Binding Protein, Master Thesis, METU, 2015). We also carried out a multiplex gene expression analysis which provides direct measurements of cellular levels of mRNA. Knocking down of the CXXC5 in the presence of E2 revealed that CXXC5 participates in the expression of a set of E2-responsive genes antagonistically or additively [9]. The functional significance of these alterations in gene expressions is reflected as alterations in cell growth in ER-positive cell lines. Cell growth decreased in the E2-treated and CXXC5 knocked down cells suggests that CXXC5 is involved in the cellular proliferation mediated by the E2-ERα signaling [9]. This decrease in cell growth was further analyzed through the cell-cycle phases. The results indicated that knocking down of the CXXC5 reduces the E2 mediated augmentation of the cell population in the S-phase and increased the population in the G2-phase. These results suggest that CXXC5 contributes to events in the E2-driven S-phase and the G2/G1 transitions [9].

In addition, we also investigate whether there is a correlation between the *CXXC5* and *ERα* expressions in ER-positive breast cancer patients using publicly available breast cancer patient datasets METABRIC to address the question of whether *CXXC5* expression has a prognostic value. We found that the *CXXC5* expression is higher in ERα positive luminal breast tumors compared to that observed with normal-like and ERα negative breast cancer types. Importantly higher levels of *CXXC5* expression predict a poor overall survival only in luminal patients, further supporting its clinical relevance to ER-positive breast cancer [9]. All these findings

suggest that deregulated *CXXC5* expression could contribute to the initiation and/or progression of ER-positive breast cancer.

1.2 CXXC Type Zinc Finger Protein 5 (CXXC5)

In 2000, the genome-wide cloning studies in hematopoietic stem and progenitor cell line CD34+ led to the identification of the CXXC5 gene [10]. *CXXC5* is located on the 5q31.2 chromosomal region, oriented on the forward strand, and encompasses approximately 35 kb long DNA. CXXC5 is a 322 amino-acid long protein and has a 33 kDa calculated molecular mass; it contains a nuclear localization signal juxtaposed to the highly conserved CXXC type zinc-finger domain at the C-terminus. CXXC5 is a member of ZF-CXXC family proteins in which the characterized members of the family have diverse functions in the cell [11]–[13]. There are 12 proteins in this family including CFP1 (CXXC protein finger 1, CXXC1), MBD1 (Methyl-CpG Binding Domain Protein 1, CXXC3), KDM2A & 2B (Lysine (K)-Specific Demethylase 2A & 2B; CXXC8 & CXXC2), DNMT1 (DNA (Cytosine-5-)-Methyltransferase 1, CXXC9) and TET1, 2, and 3 (Tet Methylcytosine Dioxygenase 1, 2, 3) proteins [11]. The protein family members preferentially recognize and bind to non-methylated CG dinucleotide containing DNA, in the regions known as CpG islands [11]. ZF-CXXC family members participate in regulating gene expression by preventing DNA methylation, leading to a state permissive to transcription activation [11]. Our Isothermal Titration Calorimetry (ITC) and Electrophoretic Mobility Shift Assay (EMSA) studies indicated that CXXC5 also preferentially binds to non-methylated CG dinucleotides [9]. The binding of CXXC5 to the non-methylated CGs raised the question of whether it has a transcriptional regulatory function or not. To address this question we used a mammalian one hybrid approach and found that CXXC5-alone does not affect the transcription regulation [9]. Collectively our studies revealed that CXXC5 has no intrinsic transcription regulation function and

therefore suggest that CXXC5 may act as a nucleation factor for the establishment of a transcription state.

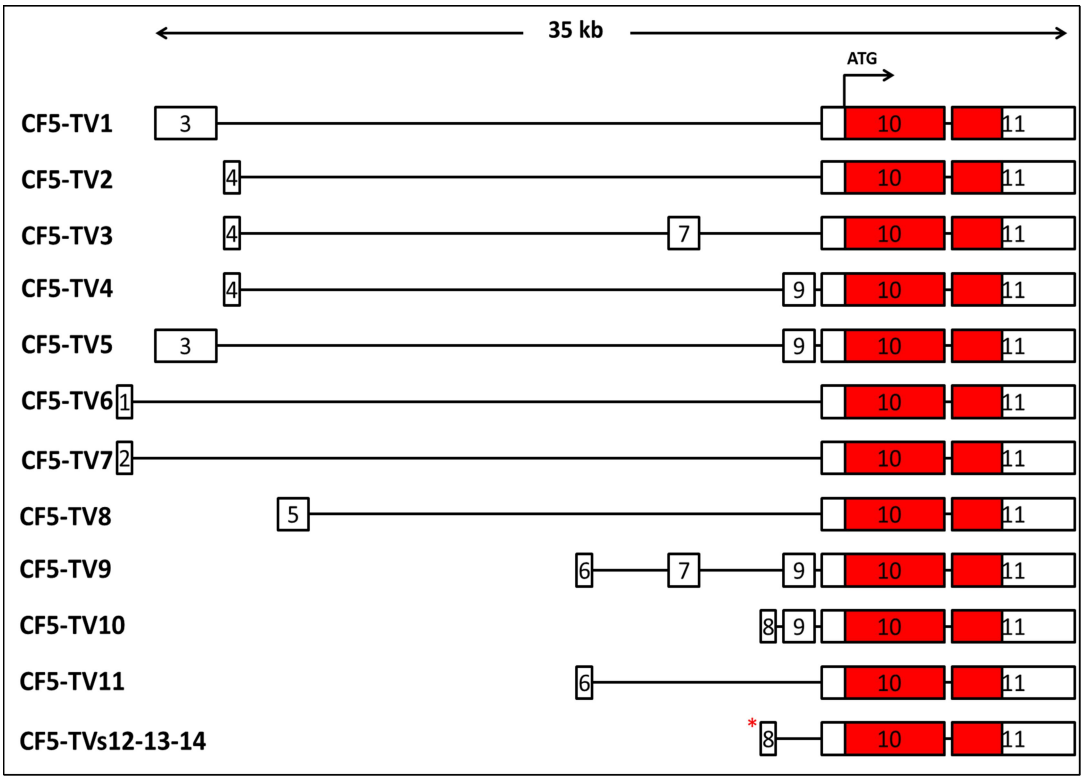

Figure 1-1 Exon usage in the *CXXC5* transcript variants (TVs). Schematics of TVs of CXXC5. Red box indicates the protein coding sequences for the CXXC5 protein. * indicates the same exon usage for three different variants with very small nucleotide changes in the exon 8. Representative image was draw according to the Ensembl (https://www.ensembl.org/index.html) CXXC5 transcript variant annotations.

There are 14 protein-coding transcript variants annotated in different databases such as NCBI, Ensembl (https://www.ncbi.nlm.nih.gov, https://www.ensembl.org/index.html). In the *CXXC5* locus, there are 11 exons and 10 introns, the exon usage of the different variants are depicted in Figure 1-1. Although some of the transcript variants display small nucleotide changes at their 3'UTR (untranslated region), the transcript diversity is mostly caused by the 5' ends of the variants. The 5' end heterogeneity could be the result of alternative promoter usage and/or alternative splicing. The last two exons which are Exon 10

and 11 are harboring the protein-encoding regions which are represented in red color (Figure 1-1).

Emerging evidence indicates that morphogenic retinoic acid [14], multifunctional cytokine family member transforming growth factor-β [15], bone morphogenetic protein 4 [16], [17], the Wnt family of secreted glycolipoprotein Wnt3a [18]–[20] or estrogen [8], [21], [22] alter the *CXXC5* expression as the primary response gene, which subsequently leads to changes in cell type-specific secondary gene expressions [16], [20], [23]–[27]. These changes are manifested as the modulation of cellular energy metabolism, proliferation, differentiation, or death in developmental processes and tissue maintenance [9], [14], [30], [15], [16], [18], [23], [25], [26], [28], [29]. Consistent with the functional importance of *CXXC5* in physiology, de-regulated expressions of *CXXC5* have been reported to correlate with the development of various pathologies, including acute myeloid leukemia (AML), gastric, prostate, and breast cancer [14], [31]–[37].

Although it has been shown that various signaling pathways are involved in the regulation of *CXXC5* expression and/or the involvement of CXXC5 protein in different cellular events through distinct signaling pathways; the mechanism by which the expression of the CXXC5 gene is regulated remains largely unknown. The fundamental component of spatial-temporal control of gene expression is achieved at the transcriptional level. This requires the integrated effects of sequence-specific trans-factors, general transcription regulators, and cis-acting DNA regulatory elements including promoters, promoter-proximal elements, distance-independent elements, locus control regions, and insulator within a highly dynamic chromatin environment. Nevertheless, promoters as diverse and complex architectural DNA segments primarily located adjacent to the transcriptional start sites (TSSs) of genes constitute the key platform for the assembly of pre-initiation complexes to mediate transcription. Delineation of promoter features of the CXXC5 gene is essential for understanding the mechanisms of the CXXC5 gene regulation in a signal pathway- and cell type-dependent manner that could underlie its role in physiology and pathophysiology.

1.3 The Aim of the Study

E2-ER complex-mediated genomic events participate in the regulation of the expressions of primary target genes, which contribute to cellular proliferation, differentiation, motility, and death. Therefore, identifying the E2 target genes is crucial for a better understanding of the complex mechanisms that underlie ER actions, which holds a critical trajectory for the development of novel prognostic and therapeutic approaches with substantial impacts on the systemic management of the target tissue diseases.

Our previous studies showed that *CXXC5* is an E2-responsive gene. CXXC5 is a non-methylated CpG dinucleotide binding protein and, it participates in the regulation of the gene expression and accordingly modulates E2-ERα mediated cellular proliferation. Here we aim to investigate how the CXXC5 gene expression is regulated and characterize the regulatory *cis*- and *trans*-acting elements of the *CXXC5*.

CHAPTER 2

MATERIALS AND METHODS

2.1 Biochemicals

Restriction and DNA modifying enzymes were obtained from New England Bio-Labs (Beverly, MA, USA) or ThermoFisher (ThermoFisher, Waltham, MA, USA). Chemicals were obtained from Sigma-Aldrich (Germany) or ThermoFisher. The antibodies for β-actin (ab8227), HDAC1 (ab19845), and CXXC5 (ab106533) were purchased from Abcam Inc. (Cambridge, MA, USA). Pageruler Prestained Protein Ladder (ThermoFisher; 26616) or Pageruler Plus Prestained Protein Ladder (ThermoFisher; 26620) was used as the molecular mass (MM) marker.

2.2 Cell Culture and Transfections

MCF7 and MDAMB231 cells were grown in phenol red-free, high glucose (4.5 g/L) containing Dulbecco's Modified Eagle's medium (DMEM, Lonza, Belgium, BE12-917F) supplemented with 10% fetal bovine serum (FBS, Lonza), 1% L-Glutamine (Lonza, BE17-605E) and 1% Penicillin/Streptomycin (Lonza, Belgium) as described previously [9], [21], [38]. HL60 cells derived from acute promyelocytic leukemia were grown in phenol red-free, low glucose (1 g/L) containing DMEM supplemented with 10% fetal bovine serum, 1% L-Glutamine, and 1% Penicillin/Streptomycin. For protein isolation to monitor the E2-responsiveness of CXXC5, MCF7 cells in 6-well tissue culture plates were maintained for 48 h in a medium containing 10% charcoal dextran-stripped fetal bovine serum (CD-FBS). Cells were then treated without (Ethanol, EtOH, 0.01%) or with 10^{-9} M E2 and maintained for 24 hours. At the termination, cells were subjected to protein extraction (NE-PER protein extraction kit, Thermo-Fisher).

Cells were transiently transfected with Turbofect transfection reagent (R0533; ThermoFisher) for 48h if not otherwise specified. Protein contents were assessed with Bradford Protein Assay (Bio-Rad Life Sciences Inc., Hercules, CA, USA).

2.2.1 siRNA Transfections

MCF7 cells grown in 12-well tissue culture plates for RT-qPCR or six-well tissue culture plates for western blot analysis. The HiPerfect transfection reagent (Qiagen) was used with a Fast-Forward transfection protocol, using 10 nM of a *CXXC5* siRNA-#9, #10, and #2 (FlexiTube GeneSolution GS51523, Qiagen), *ELF1* siRNA pool (sc-37837, Santa Cruz, CA, USA), or *MAZ* siRNA pool (sc-38035, Santa Cruz, CA, USA). 48h after transfection, cells were subjected to total RNA isolation (ZymoResearch) or protein extraction (Thermo-Fisher).

2.3 Engineering of Reporter Vectors

For reporter assays, we generated luciferase reporter vectors (pGL3-Basic) that bear the *Firefly Luciferase* cDNA as the reporter enzyme (Promega Corp., Madison, WI, USA). The reporter plasmids bearing the estrogen-responsive region of *Oxytocin* (*OXT*) and *TFF1* were described previously [39] were used as controls. For the engineering of the reporter vector bearing the estrogen-responsive 305 bp long *CXXC5* region, which is found upstream of the first ATG (-305 to +1, +1 denotes the A residue of the first ATG), PCR was performed using MCF7 genomic DNA and cloned into the pGL3-Basic vector (Figure 2-1). An overlapping PCR approach was used for the generation of the mutant-CXXC5-ERE within the *CXXC5* region context. With this mutant, the CXXC5-ERE 5'-GGTCAggaTGACA-3′ was converted into a non-ERE sequence, which is 5′-TTTGATCCCTCAA-3′. The non-ERE was inserted into pGL3-Basic vector and sequenced.

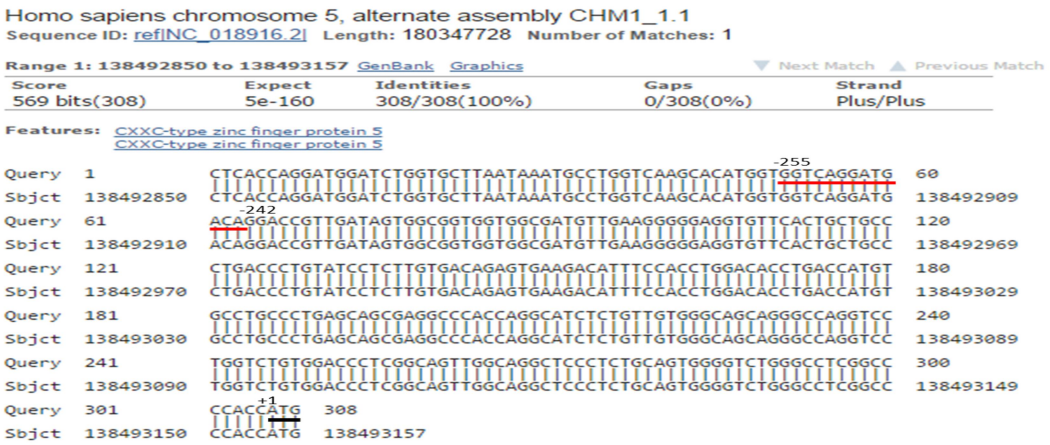

Figure 2-1 Estrogen responsive region of *CXXC5*. The 305 base-pair long *CXXC5* region containing an ERE sequence was obtained by PCR using the genomic DNA of MCF7 cell and sequenced. The sequence (query) was then compared to sequences in human genome using Basic Local Alignment Search Tool (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The ERE sequence (underlined in red) resides between residues -242 through -255 from the first encoding ATG (A being +1) codon of CXXC5, underlined in black. The PCR generated DNA fragment was inserted into the pGL3-Basic vector for the construction of the pGL3-CXXC5-ERE reporter plasmid.

For the promoter analyses studies, putative promoter elements of the CXXC5 gene were inserted into the pGL3-Basic vector using appropriate restriction enzymes. For the reporter vector engineering bearing the putative *CXXC5* promoter containing genomic region, a DNA fragment of 1548 bp of the *CXXC5* locus (GRCh38.p12 Primary Assembly, chromosome 5: 139647220-139649194) generated by PCR using the genomic DNA of MCF7 cells as a template. To increase the putative *CXXC5* promoter region's resolution, we carried out deletions from both the 5' and 3' ends of the region by PCR (Figure 2-2). I inserted them into the pGL3-Basic vector with appropriate restriction enzymes. All constructs were sequenced for the PCR fidelity. A reporter vector carrying a *CMV* promoter that drives the expression of *Renilla Luciferase* cDNA (pCMV-RL, Promega), which we described earlier [40], [41], was monitored for transfection efficacy. For luciferase studies, MCF7 cells, 4 x 10^4 cells/well, were seeded in 48-well plates for 48h. Cells were then transiently transfected with 125 ng reporter vector, and if

necessary 75 ng expression vector, together with 0.5 ng pCMV-RL. Turbofect (ThermoFisher Scientific) was used as the transfection reagent. Luciferase assays were performed with a Dual-Luciferase Assay Kit (Promega, Corp., Madison, WI, USA) as described previously [21].

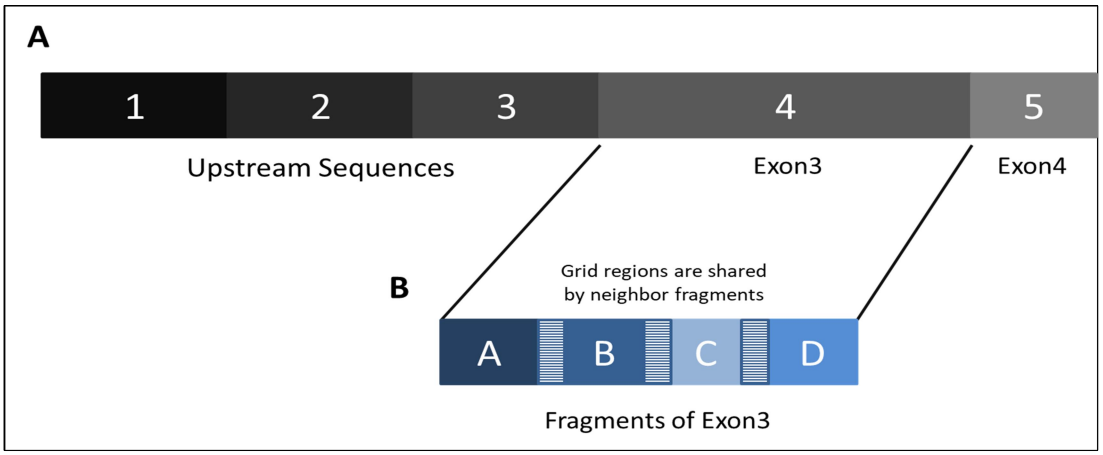

Figure 2-2 Schematic representation of the *CXXC5* promoter cloning into reporter luciferase vector. *CXXC5* putative promoter regions were cloned into pGL3-Basic vector and to increase the resolution of the regulatory regions of *CXXC5* promoter further 5' and 3' deletions were introduced.

2.4 PCR and RT-qPCR

We used *Taq* Polymerase (ThermoScientific) and SYBR® Green Mastermix (Roche Applied Science, Indianapolis, IN, USA) and product-specific primer sets for conventional PCR or qPCR, respectively, according to the Manufacturer's Instructions together with DMSO at appropriate concentrations if necessary. During the RT-qPCR experiments, MIQE Guidelines were followed [42].

2.4.1 Western Blotting (WB)

MCF7 cells grown in 6-well tissue culture plates in medium supplemented with CD-FBS for 48h were treated without (EtOH, 0.01%) or with 10^{-9} M E2 and/or 10^{-7} ICI for 24h. At the termination, cells were collected, and protein isolation was

performed using NE-PER protein extraction kit (Thermo-Fisher). Protein content in extracts was measured with Bradford Protein Assay (Bio-Rad). Nuclear extracts (50 μg) were then subjected to SDS 10%-PAGE. Proteins were probed with an antibody specific to CXXC5 (ab106533, Abcam), followed by a secondary antibody conjugated with the horseradish peroxidase (Santa Cruz). Protein images were developed using the ECL-Plus Western Blotting kit (GE Healthcare Bio-Sciences, Pittsburgh, PA, USA) and captured with ChemiDoc™ Imaging System (Bio-Rad). The quantification of images was carried out using the ImageJ image processing program (https://imagej.nih.gov/ij/).

2.5 *In silico* Analysis for ERE Sequence

For the prediction of putative ERE sequences in the *CXXC5* locus, the dragon ERE finder version 3, System for Identification and Interactive Analyses of Estrogen Response Elements in DNA Sequences [43] was used.

2.6 Electrophoretic Mobility Shift Assay (EMSA)

In EMSA, we used none or 5' end biotin-labeled oligomers. The oligomer sequences are represented in Appendix B. Biotin-labeled oligomers were purchased from Integrated DNA Technologies (IDT; Coralville, IA, USA) and annealed. For the investigation of Flag-tagged ERα binding to consensus and CXXC-ERE sequences; double-stranded DNA fragments were incubated in the presence or absence of extracts (10 μg) of MDAMB231 cells transfected with expression vectors bearing none (control) or the Flag-ERα cDNA for 15 minutes and after the Flag-M2 (Sigma-Aldrich, Germany) antibody addition to the antibody control samples another 15 minutes incubation was performed; in total reactions were waited for 30 minutes to obtain the protein-DNA complex.

On the other hand, for the investigation of the proteins of ELF1 and MAZ binding to their binding motif sequence which is found within the *CXXC5* promoter; 45 μg

of endogenous MCF7 cell extracts were used together with the double-stranded corresponding binding motif containing sequence, cold competitor or binding motif mutated DNAs. ELF1 (B-9, sc133210, Santa Cruz, CA, USA) or MAZ (133.7, sc-130915, Santa Cruz, CA, USA) antibodies were added to the antibody-control reactions for the verification of the observed shifts are due to specific protein of interests.

Reactions were run on 5% non-denaturing polyacrylamide gel in 0.5X Tris/Borate/EDTA (TBE) up to lower dye migrates 3/2 of the gel, and samples were transferred onto a positively charged nylon membrane (Thermo Scientific, MA, USA), which is soaked with 0.5X TBE for 10 minutes. After the transfer, the DNAs were crosslinked to the membrane via UV (312 nm) for 15 minutes. For the detection of the DNAs on the membrane, Chemiluminescent Nucleic Acid Detection Module (Thermo Scientific, CA, USA) was used according to the Manufacturer's Instructions. The developed blot was visualized with ChemiDoc-MP system (Bio-Rad, MO, USA).

2.7 Chromatin Immunoprecipitation Assay (ChIP)

We used ChIP specific ERα antibody (HC-20x; Santa Cruz Biotechnology Inc., Santa Cruz, CA, USA) in the CXXC5-ERE ChIPs. MCF7 cells, grown in medium supplemented with CD-FBS in T75 tissue culture plates for 72h, were treated without or with 10^{-9} M E2 for one hour. Cells were then fixed with 0,75% paraformaldehyde at room temperature for 10 min followed by the addition of 0,125 M Glycine (BioShop, #GLN002) for 10 min to quench and lysed with Nuclei Lysis Buffer 1% SDS, 10mM EDTA, 50mM Tris-HCl pH=8.0, 0.5mM PMSF, 1X PIC (Roche) and actively sonicated for 40 minutes. Cell debris was pelleted, and the supernatant was collected. After pre-clearing, the supernatant was incubated with the ERα antibody or rabbit IgG (Santa Cruz) overnight.

For the *CXXC5* promoter ChIPs, MCF7 cells, grown in medium supplemented with 10% FBS in T75 tissue culture plates for 48h, were fixed with 1% formaldehyde at RT for 15 min followed by the addition of 0,125 M Glycine for 10 min to quench and lysed with Nuclei Lysis Buffer and actively sonicated for 20 min. Cell debris was pelleted, and the supernatant was collected. After blocking of the nuclear extracts, the supernatant was incubated with a species-specific (Mouse or Rabbit) IgG (Santa Cruz), RNAP II antibody (POLR2A, CTD4H8; Santa Cruz Biotechnology Inc., Santa Cruz, CA, USA) for the precipitation of hypo- and hyper-phosphorylated RNAP II or Ser5-RNAP II (POLR2A, Phospho-Rpb1 CTD, D9N5I, Cell Signaling Technology, Beverly, MA, USA) for the precipitation of specifically Ser5 phosphorylated RNAP II overnight. Nuclear extracts were also incubated with an antibody specific to CREB1 (X-12, sc-240, Santa Cruz Biotechnology Inc., Santa Cruz, CA, USA), ELF1 (B-9, sc-133210, Santa Cruz), MAZ (133.7, Santa Cruz, c-130915), or RB1 (Retinoblastoma, 4H1, #9309, Cell Signaling Technology, Beverly, MA, USA) overnight.

All samples were then subjected to immunoprecipitation with Protein G-coupled magnetic beads (New England BioLabs) for anti-mouse antibodies, or Protein A/G coupled magnetic beads (New England BioLabs) for anti-rabbit antibodies. After washes, de-crosslinking, RNA, and protein digestion, DNA was recovered with phenol:chloroform: isoamylalcohol (25:24:1) and ethanol precipitation. Samples (2 µl of 60 µl elution) were subjected to qPCR using ChIP primers (Appendix A) specific to the CXXC5-ERE region, *TFF1* promoter region containing ERE sequence, the *CXXC5* promoter, the *GAPDH* promoter (positive control, for RNAP II and Ser5P RNAP II occupancy), *C-MYC* promoter (positive control, for RB1 and MAZ occupancy), *OAS1* promoter (positive control, for ELF1 occupancy), *CCNA2* promoter (positive control, for CREB1 occupancy) or *MB_Exon2* (Myoglobin gene as a negative control) were used. qPCR results were normalized using percent (%) of input approach [44] and represented in fold change of the two biological replicates.

2.8 5' Rapid Amplification of cDNA Ends

2.8.1 RNA Isolation from HL60 and MCF7 Cells

HL-60 and MCF7 cells grown in two T-75 flasks at 90% confluency or one flask in total approximately $15x10^6$ cells, respectively, were used for total RNA isolation with TRIzol (#15596026, Thermo Scientific, USA). Cells were trypsinized, spun-down to pellet in 15 ml falcons. Pellets were not washed with PBS to prevent the possibility of mRNA degradation. 8 ml TRIzol was directly added onto the pellet to suspend cells, which then incubated 5 minutes at RT to allow complete lysis. The suspension was then transferred into a sterile 15ml falcon which is treated with phase-lock (#2302830, Quantabio, USA). Chloroform was added into the tube as 0,2ml per 1ml TRIzol and shook vigorously for 15 seconds. Samples were incubated for 2-3 minutes at RT and centrifuged at 3040 G for 25 minutes at 4^oC. The aqueous phase contains the nucleic acids (the volume of the aqueous phase was about 60% of the original TRIzol volume) was transferred to a new sterile 15 ml falcon tube and was mixed with isopropyl alcohol as 0,5 ml isopropyl alcohol per 1ml TRIzol and incubated for 10 minutes at RT to precipitate the nucleic acids. Samples were then centrifuged at 3040 G for 20 minutes at 4^oC. Precipitated nucleic acids may not be visible, but it forms a jelly-like structure at the bottom of the falcon. After removal of the supernatant, nucleic acid pellet was washed with 8 ml 75% ethanol, followed by another wash with 70% ethanol and centrifuged at 3040 G for 7 minutes at 4^oC. The RNA pellet was then incubated at RT to air-dry for 15-25 min to prevent the complete drying of nucleic acid pellet, which reduces the proper reconstitution of RNA. Then 60 µl of nuclease-free water was used for the resuspension of nucleic acids. RNA quantification was carried out with and Nanodrop (BioDrop μLITE). The general yield of the isolated nucleic acid concentration was 2,2 – 2,8 µg/µl for one T-75 flask of MCF7 as well as for two T-75 flasks of HL-60 cells.

2.8.2 DNase I Treatment for the Isolated Nucleic Acids

In solution DNase I (1 U/µl, # EN0525, Thermo Scientific, USA) treatment was performed for the RNA purification. The reaction was prepared, as shown in Table 2.1.

Table 2-1 DNase I Reaction

Components	Final Concentration
RNA	X µg
10X DNase I Reaction Buffer	1X
DNase I Enzyme (1U/µl)	X µl (1U for 1µg RNA)
Nuclease-free water	Up to 100 µl

The mixture was incubated at 37^{o}C for one hour. The reaction was then stopped with phenol:chloroform:isoamylalcohol (25:24:1 v/v, # 15593031, Thermo Scientific, USA) by using the same volume of the Dnase I reaction (100 µl). The mixture was vortexed 30 second and then transferred into a new tube treated with phase-lock for incubation on ice for 10 minutes and centrifuged for 20 minutes at 4^{o}C, 14000 G. After the centrifuge, the upper phase was taken into a clean tube, and 1/10 V 3M sodium-acetate (NaAc) and 3 V 100% ethanol were added. The samples were incubated at -20^{o}C for overnight or longer. Samples were then centrifuged for 30 minutes at 4^{o}C 14000 G. Supernatant was removed, the pellet can be seen as jelly-like, and washed with 600 µl 70% ethanol. Then samples were centrifuged for 15 minutes at 4^{o}C 14000 G. Finally, ethanol was removed, and RNA pellet was air-dried for a maximum of 15-20 minutes. RNA was eluted to 20 µl of nuclease-free water and quantified with Nanodrop (BioDrop μLITE).

2.8.3 Removal of the Ribosomal-RNA (rRNA) from the Total RNA

Depending on the source and quality, RNA yield may vary, but in general, 1-5% of the total RNA represents messenger-RNA (mRNA). When we used total RNA isolation in our analyses, we failed to obtain sufficient quantity and quality of *CXXC5* transcript, likely due to low transcript levels of *CXXC5*. To circumvent this problem, we decided to remove rRNAs, which constitute about 80% of the RNA population. To accomplish this, we used RiboMinus™ Human/Mouse Transcriptome Isolation Kit (#K155001, Thermo Scientific, USA). We performed rRNA removal according to the Manufacturer's Instructions. The method is based on the selective depletion of the rRNAs by the hybridization of rRNA to Locked Nucleic Acid (LNA) probes conjugated to magnetic beads. LNA, referred to as inaccessible RNA, is a modified RNA nucleotide that significantly increases the hybridization properties to DNA or RNA. After the hybridization, LNA probe bound rRNAs were captured with the help of a magnetic stand, and the supernatant contains mRNAs, which is depleted from rRNAs, can be recovered. Phenol:chloroform: isoamylalcohol was then used for the purification of mRNAs. After the removal of rRNAs from the population, the remaining RNA concentration would be low as expected; therefore, the amount and the quality of the starting material are significant. However, a maximum of 10 µg of total RNA can be applied to one reaction due to magnetic bead probes' limitations. For 10 µg total RNA, 8 µl RiboMinus probe (100 pmol/µl) was used in 300 µl hybridization buffer (B5). The mixture was incubated at 75^{o}C for 5 minutes to denature RNA and then placed at 37^{o}C to allow the reaction to cool for 30 minutes. Meanwhile, I started to prepare the magnetic beads; 500 µl of resuspended RiboMinus magnetic beads were separated with a magnetic stand, and beads were washed with 500 µl of nuclease-free water and again separated. After washing four times, beads were resuspended in 200 µl Hybridization Buffer (B5) and kept at 37^{o}C until use. The cooled and hybridized ~328 µl sample was then mixed with Magnetic beads and incubated at 37^{o}C for 15 minutes. Every 5 minutes mixture was mixed gently. After

the incubation, the mixture was placed on the magnetic stand and waited for 1 minute to separate magnetic bead-bound rRNAs. Supernatant, which contains rRNA depleted RNA population, was transferred into a tube capable of holding 2X volume of the supernatant. Magnetic beads were resuspended in 50 µl Hybridization Buffer (B5), and after the separation of magnetic beads again, the supernatant was added to the previous supernatant. Since the total volume of the supernatant was ~575 µl, the same amount of phenol:chloroform:isoamylalcohol (25:24:1, v/v, # 15593031, Thermo Scientific, USA) was added for mRNA purification. Finally, the concentration of the RNA content was measured with Nanodrop (BioDrop μLITE).

2.9 Northern Blot (NB)

2.9.1 Designing of the Probes for NB

Vector-specific biotin-conjugated primers were used for the amplification of the probe sequences of Exon10-Exon11 boundary, *GAPDH,* and the first exon of the CXXC5-TV2 (Exon3) (Appendix A). Initially, I cloned the probes' target sequences into the pGL3-basic vector and sequenced them for the fidelity control of the PCR amplification. PCR purified probes are used for the hybridization in northern blotting.

2.9.2 Northern Blotting

Northern blot is used to detect the presence of the *CXXC5* transcript. We used NorthernMax™ Kit (#AM1940, Thermo Scientific, USA) according to the Manufacturer's Instructions. 0,5 g of agarose-LE is melted in nuclease-free water for 50 ml of gel. After the melted agarose is cooled down to 50-60^{o}C, 5 ml of 10X denaturating gel buffer was added and mixed well and then poured to about 0,6 cm in thickness with a 120 µl holding capacity comb. After the gel hardens, it is placed

into a running tank filled with 1X MOPS gel running buffer, prepared with nuclease-free water from 10X MOPS. Meanwhile, 10 µg RNA samples were mixed with three volumes formaldehyde load dye, 3 µl of RNA ladder (RiboRuler High Range RNA Ladder, # SM1821, Thermo Scientific, USA), and the samples incubated at 65^{o}C for 15 minutes. After a brief spin, samples and the ladder were loaded into wells. The gel was run at 5V/cm (the distance between the electrodes of the electrophoresis chamber) for approximately two hours while bromophenol dye (lower dye) was monitored until it migrates to the bottom of the gel. Since the pre-stained agarose gel (i.e., ethidium bromide) causes significant changes in RNA motility and reduces probe hybridization compared with unstained gel [45], we used Sybr Gold (#S11494, Thermo Scientific, USA) staining for the assessment of the transfer efficiency after the running. The last well was loaded with the same amount of RiboMinus-treated RNA sample to ensure that RNA degradation did not occur during the RNA isolation or RiboMinus treatment. After gel running, the last well of the gel was cut with a razor and stained with Sybr Green for 10 minutes to visualize with UV transilluminator. Observation of distinct bands under the UV is indicative of intact mRNA isolation; in contrast, a smear-like view of staining suggests mRNA is degraded. Ensuring that mRNAs are not degraded after visualization, the gel was then placed onto a positively charged membrane (#11209272001, Roche, Sigma-Aldrich), and a sandwich complex was prepared by using pre-wet Whatman papers. To maintain the continuous moisture, the Whatman filter paper that is found at the top of the sandwich was used as a bridge for the sandwich and the transfer buffer container. Approximately 150 g weight was placed to the top. The transfer takes approximately two hours (15-20 min per mm of the gel thickness). Transferred mRNAs were then crosslinked to the membrane via UV transilluminator which has UV lambs with 312 nm wavelength for 10 minutes. After crosslinking, membranes that would be subjected to different corresponding biotin-tagged probes were cut and placed into different sterile 15 ml falcon tubes. 6 ml of pre-heated to 42^{o}C ULTRAhyb buffer was added into each falcon tube and placed onto vertical rotator (ISOLAB) at 42^{o}C oven for 40 minutes

to prehybridize. Meanwhile, biotin-tagged probes that are double-stranded PCR products were prepared for the hybridization. Biotin-tagged probes were diluted as 10-fold with 10mM EDTA containing TE buffer to a final volume of 100 µl and incubated at 90^{o}C for 10 minutes on a heat-block (ISOLAB) for denaturation to allow the antisense probe hybridize to the target mRNA. Immediately after the incubation, probes were mixed with 500 µl of ULRTAhyb buffer from the prehybridized samples to prevent the self-annealing of the DNAs. The mixture was immediately added to the falcons and placed in the oven for hybridization overnight. The next day, the hybridization buffer mixture was removed, and membranes were washed with low stringency buffer at RT for 5 minutes, twice. Later membranes were washed with high stringency buffer for 15 minutes at 42^{o}C, twice. After washing steps, membranes were washed with MilliQ water for 2 minutes at RT to remove the wash buffer and continue with detection of the biotin-labeled probes that were expected to hybridize with RNA by chemiluminescence. We used Chemiluminescence Nucleic Acid Detection Module (89880, Thermo Scientific, US) for visualization of the northern blot membranes. Each membrane was blocked with 5 ml of pre-heated to 37^{o}C of Blocking buffer for 15 minutes at RT on the vertical rotator. Meanwhile, 5 ml of Blocking buffer was mixed with 16,7 µl Stabilized Streptavidin-Horseradish Peroxidase Conjugate (substrate) as 1:300 dilution. After blocking, the membranes were incubated at RT for 15 minutes with the substrate. Meanwhile, 4X Wash buffer was diluted to 1X with MilliQ water, and membranes were washed with 5 ml of wash buffer for 5 minutes at RT, for four times. After washing, membranes were treated with 5 ml of Equilibrium buffer for 5 minutes at RT. Finally, membranes were incubated with Substrate working solution (prepared by addition of 500 µl Luminol/Enhancer Solution to 500 µl Stable Peroxide Solution) for 5 minutes at dark and RT. At the end of the incubation, the excess amount of the Substrate solution was removed by paper towels, and membranes were placed into a sheet protector and visualized with ChemidocTM MP system (Bio-Rad, US).

2.10 Pull-Down Assay

2.10.1 Nuclear Protein Extraction

MCF7 cells grown in T75 flasks were trypsinized and collected by centrifugation at 1000xG for 5 min at 4°C. Pellet was washed with ice-cold PBS twice, and packed cell volume (PCV) was determined. Cells were resuspended in 1xPCV of Buffer A [ice-cold Swelling Buffer: 10 mM HEPES pH 7.9, 1,5 mM $MgCl_2$, 10 mM KCl, 0,5% NP-40, freshly added; 0,5 mM Phenylmethylsulfonyl Floride (PMSF), 0,5 mM DTT, and 1xProtease Inhibitor (PI)] and rested on ice to allow cells to swell. Cells were then lysed by passaging 25 times through a 25-gauge needle. Lysed cells were centrifuged to pellet crude nuclei at 12000xG for 20 sec at 4°C. The crude nuclear pellet was washed twice with 1xPCV of WB [ice-cold Wash Buffer: 10 mM HEPES pH 7.9, 1,5 mM $MgCl_2$, 10 mM KCl, freshly added; 0,5 mM Phenylmethylsulfonyl Floride (PMSF), 0,5 mM DTT, and 1xProtease Inhibitor (PI)]. After centrifugation, pellet was resuspended in 2/3 PCV of Buffer B [ice-cold-20 mM HEPES pH 7.9, 1,5 mM $MgCl_2$, 420 mM KCl, 0,2 mM EDTA, 2,5% glycerol, freshly added; 0,5 mM Phenylmethylsulfonyl Floride (PMSF), 0,5 mM DTT, and 1xProtease Inhibitor (PI)] and rested on ice for 30 min with occasional agitation. Samples were then centrifuged for 5 min at 4°C. Supernatant which is the nuclear extract was diluted isovolumetrically to decrease the salt concentration to 125 mM with Buffer D [ice-cold-20 mM HEPES pH 7.9, 1,5 mM $MgCl_2$, 100 mM KCl, 0,2 mM EDTA, 10% glycerol, freshly added; 0,5 mM Phenylmethylsulfonyl Floride (PMSF), 0,5 mM DTT, and 1xProtease Inhibitor (1xPI)].

2.10.2 Promoter Pull-Down Assay

Based on luciferase reporter results, we amplified a 232 bp in length DNA fragment by PCR from the genomic DNA of MCF7 cells that includes the *CXXC5* promoter (Segment A; -72 to +152, +1 being the first nucleotide of the annotated

Exon3) and inserted it into pGL3-Basic reporter vector with appropriate restriction enzyme cut sites. Similarly, a 232 bp in length DNA fragment within Exon10 of the *CXXC5* gene was cloned into the vector as the control pulldown DNA. 5' end-biotinylated forward and reverse primers specific to the expression vector were then used for the amplification of Segment A and the control Exon10 DNA sequences by PCR. Biotinylated double-stranded PCR amplicons were recovered from agarose gels with Zymoclean Gel DNA Recovery Kit (Zymo Research).

Streptavidin magnetic beads (SMB, New Biolabs England) were blocked using 2% BSA in PBS for two hours at 4^{o}C, followed by washes with 1xPBS twice. Half of the blocked SMBs were resuspended in 200 µl PBS containing 0,5 mM Phenylmethylsulfonyl Floride (PMSF), 0,5 mM DTT, and 1xPI and incubated with biotinylated-DNAs. The other half of blocked SMB (12 µl) was mixed with one ml of nuclear extracts for pre-clearing for one hour at 4^{o}C in 300 µl PBS. Subsequently, the pre-cleared nuclear extract was divided into three 400 µl aliquots. Aliquotes of the extract was then mixed with beads together with 10 µg biotinylated double-stranded Segment A, control DNA or beads alone in the presence of 10 µg of Poly[d(I-C)] to form the protein-DNA complexes overnight at 4^{o}C on a rotator. The SMB-DNA-protein mixtures were subsequently washed with 1xPBS three times for 5 min each and resuspended in 200 µl 1xPBS for Mass Spectrometry (MS) analyses.

2.10.3 Protein Identification by Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS)

MS analyses of two biological replicates were carried out at the Koç University Proteomic Facility (Istanbul, Turkey) as two technical repeats for each biological replicate. The SMB-DNA-protein mixtures were washed with 50 mM NH_4HCO_3, followed by reduction with 100 mM DTT in 50 mM NH_4HCO_3 at 56°C for 45 min, and alkylation with 100 mM iodoacetamide at RT in the dark for 30 min. MS Grade Trypsin Protease (Pierce) was added onto the beads for overnight digestion

at 37°C (enzyme: protein ratio of 1:100). The resulting peptides were purified using C18 StageTips (ThermoFisher). Peptides were analyzed by online C18 nanoflow reversed-phase HPLC (2D nanoLC; Eksigent) linked to a Q-Exactive Orbitrap mass spectrometer (ThermoFisher). The data sets were searched against the human SWISS-PROT database version 2014_08. Proteome Discoverer (version 1.4; ThermoFisher) was used to identify proteins. The final protein lists were analyzed using the STRING v11 [46] and DAVID [47], [48] databases.

2.11 *In silico* Analysis of Transcription Factor Binding Motifs for the *CXXC5* Promoter

To find TF binding motifs, we developed a motif search tool using all the available ChIP-Seq datasets at the Cistrome [49] database (http://cistrome.org/). This tool obtains 1) a set of binding locations on a sample of Chip-Seq reads using MACS2 peak locations, 2) the reference sequence of the genomic locus to analyze, and 3) the binding motifs for a specific Transcription Factor from the JASPAR [50] database (http://jaspar.genereg.net/), as input and conducts and approximate string search on the reference sequence binding locations using the consensus binding motif as the query sequence. The program outputs both forward and reverse strand hits, and the hits on the binding locations are ranked to a logarithmic sequence similarity score.

2.12 Statistical Analysis

Results were presented as the mean ± standard deviation (SD). Significance was determined using a two-tailed unpaired t-test with a confidence interval, minimum, of 95%.

CHAPTER 3

RESULTS AND DISCUSSION

PART I: *CXXC5* EXPRESSION IS REGULATED BY E2-ERα

3.1 CXXC5 is an E2 Responsive Gene

In my master thesis studies, I treated MCF7 cells without (0,01%) ethanol or with 10^{-9} M E2 or 10^{-7} M ICI, which is the complete antagonist of the ER, or E2 and ICI together for 24h and continue with RNA isolation. Isolated RNAs were converted into cDNA libraries, and *CXXC5* expression changes were monitored using the qPCR approach. I identified that *CXXC5* is an E2-responsive gene, and its RNA levels are increased in the presence of E2 compared to ethanol treatment. On the other hand, when the E2 and ICI treatment was conducted together, ICI completely blocked the E2 mediated upregulation of *CXXC5* levels indicating that the E2-responsiveness of *CXXC5* is ER-dependent. To investigate whether the E2 responsiveness of *CXXC5* RNA levels is reflected in the protein levels, firstly, I started with the antibody characterization studies. Although there are several commercially available CXXC5 antibodies, the protein detection specificity of those antibodies varies in biochemical assays such as western blotting, immunocytochemistry, and chromatin immunoprecipitation, co-immunoprecipitation. I characterized the CXXC5 antibody using CXXC5 targeting or non-targeting siRNAs and overexpression approaches and decipher the CXXC5-specific migration in western blot analysis (Figure 3-1). Results revealed that CXXC5 targeting siRNA #10 is the most effective one compared to others for knocking-down the CXXC5 protein levels.

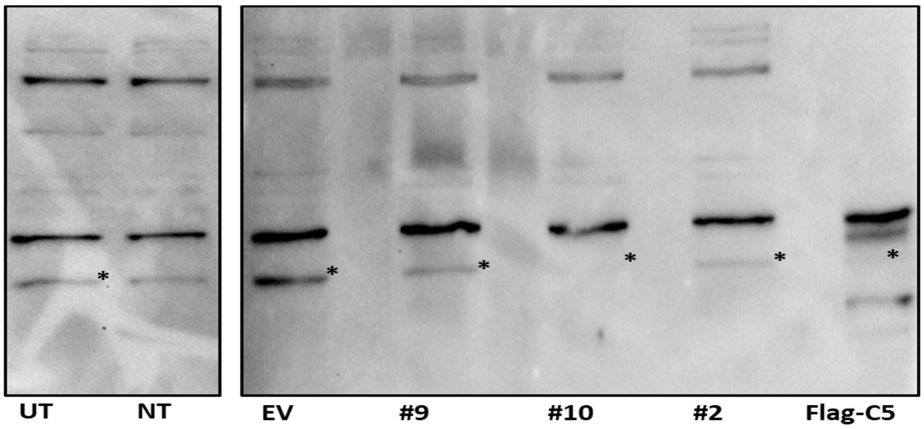

Figure 3-1 Antibody characterization. MCF7 cells were transfected for 24h without (un-transfected, UT) or with AllStars (NT), siRNA#2, #7, #9 or #10. I also transfected cells with pcDNA3.1(-) bearing none (Vector, V) or WT-CXXC5 cDNA as control. 100 µg nuclear protein extracts, with the exception of F-C5 which was 25 µg to prevent the shadowing effect of the overexpressed protein on the endogenous protein, were subjected to WB using CXXC5 antibody (ab106533). Star denotes the endogenous CXXC5, while the arrow indicates the overexpressed Flag-CXXC5. A representative image from two independent experiments is shown.

After I verified that the Abcam106533 CXXC5 antibody which specifically detects the CXXC5 protein on western blots, I subjected the proteins that are obtained from E2-alone, ICI-alone or E2-ICI together treated MCF7 cells to western blot analysis. I observed that in the presence of the E2, CXXC5 protein levels are augmented compared to the vehicle control treatment. While ICI alone did not affect the CXXC5 protein levels, when ICI is introduced together with E2, E2 mediated augmentation of the protein levels is prevented by ICI (Figure 3-2).

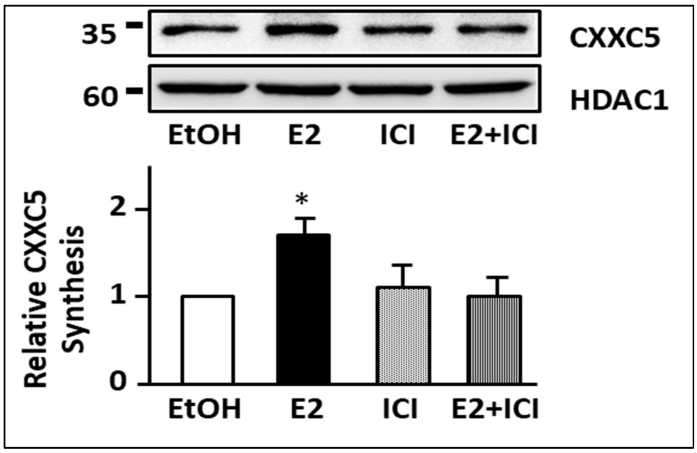

Figure 3-2 CXXC5 protein levels are increased in the presence of E2. MCF7 cells grown in medium containing CD-FBS for 48h were treated without (EtOH, 0.01%) or with 10^{-9} M E2 and/or 10^{-7} M ICI for 24h. Nuclear extracts were subjected to WB using ab106533 or an HDAC1 antibody. A representative image from two independent experiments is shown. Changes in protein levels were quantified with the ImageJ image processing program. Asterisk (*) denotes significant change.

3.2 The Binding of ERα to an Intronic ERE Sequence in the CXXC5 Gene Locus *In Vitro*

I showed that CXXC5 is an E2-responsive gene at the transcript and protein level, and this responsiveness is ER-dependent. To investigate how the E2-ER complex regulates the *CXXC5* expression, I initially subjected the *CXXC5* locus sequence to the Dragon ERE finder version 3 [43]. A stringent analysis that allows only one mismatch in the putative ERE sequence compared to the consensus ERE sequence (GGTCAnnnTGACC) was performed with *in silico* studies. According to our previous studies, one mismatch in the consensus core motif does not affect the ERα binding and the function [51]. Results revealed that in the *CXXC5* locus, there is a putative ERE sequence (GGTCAggaTGAC<u>A</u>), which contains one mismatch represented as an underlined nucleotide and is found at the 242 bp upstream of the translation initiation codon of the CXXC5 protein.

We used electrophoretic mobility shift assay (EMSA) to test whether ERα interacts with this putative CXXC5-ERE *in vitro*. In EMSA, we utilized 5'end biotinylated DNA substrates consisting of a 13 bp core consensus (Con)-ERE or CXXC5-ERE surrounded by the non-specific sequences known to be not affecting the ERα interactions [51]. To test the binding ability of the ERα to ERE sequences, we used extracts of MDAMB231 (triple-negative breast cancer cell line) cells which were transiently transfected with an expression vector bearing none as control (EV) or Flag-tagged ERα ORF.

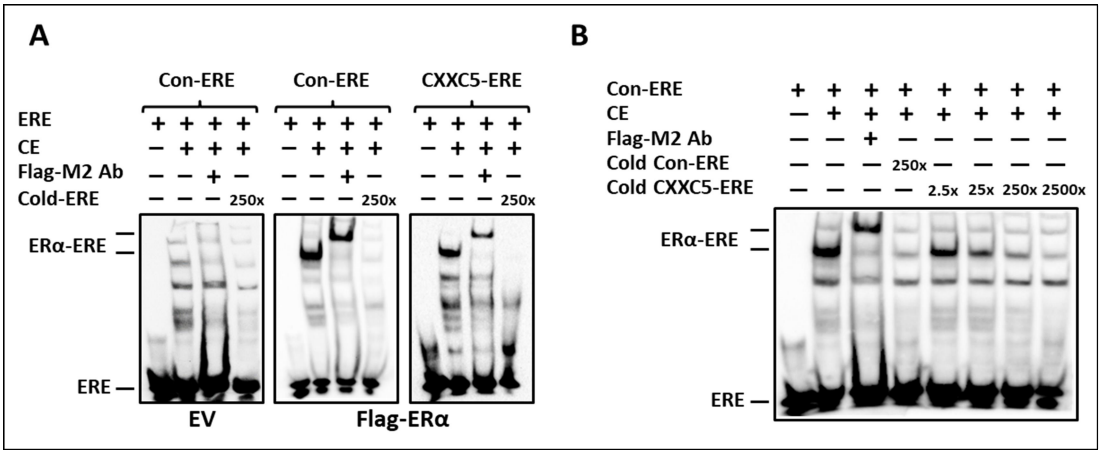

Figure 3-3 Electrophoretic mobility shift assay. A) & B) Cell extracts (CE; 10 μg) of MDAMB231 cells transfected with pcDNA3.1(-) bearing none (Vector, V) or the Flag-ERα cDNA were subjected to EMSA using biotinylated DNA (40 fmol) with (+) or without (-) the Flag-M2 antibody (Flag-M2) in the absence (-) or presence (+) of cold competitor at indicated amounts. ERα-ERE denotes the protein-bound biotinylated ERE. ERE indicates the unbound (free) biotinylated ERE. A representative result from three independent determinations is shown.

In our negative control, which is the extracts of the parental vector-transfected (EV) cells, there is no specific retardation caused by the DNA interaction as expected (Figure 3-3A-left panel). On the other hand, with extracts of the Flag-tagged ERα ORF vector-transfected cells, the migration retardation has been detected due to the interaction of the Con-ERE or CXXC-ERE (Figure 3-3A-middle & right panels, respectively). The addition of the Flag-M2 antibody to the EMSA reaction resulted in further retardation due to the increased molecular weight of the DNA-ERα complex, which specifies the DNA-protein interaction. I also used a cold-competitor approach, in which the unlabeled Con-ERE or

CXXC5-ERE as in 250-fold molar excess were used to prevent the formation of the biotinylated DNA-ERE complex. Furthermore, when I used the unlabeled CXXC5-ERE with increasing molar ratios (2.5x, 25x, 250x, and 2500x) together with biotinylated Con-ERE, I observed that CXXC5-ERE effectively compete with Con-ERE and abated the formation of the ERα-Con-ERE complex (Figure 3.3B). I also tested the diminishing effect of the cold-competitor addition to the EMSA reaction on the ERα-CXXC5-ERE complex using an unlabeled non-ERE as 2500-fold molar excess did not affect the interaction of the ERα and CXXC5-ERE (Figure 3-4). Our results indicated that ERα specifically interacts with the CXXC5-ERE *in vitro*.

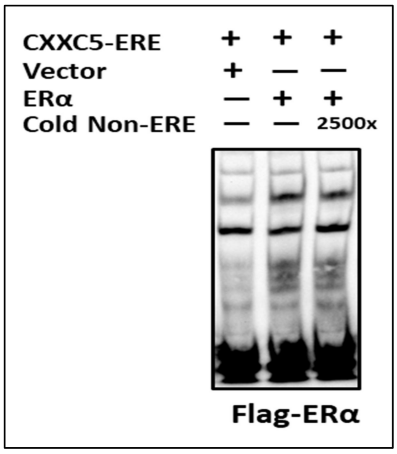

Figure 3-4. Competition assay using unlabeled non-ERE sequence. MDAMB231 cells were transfected with pcDNA3.1 (-) bearing none (V) or Flag-ERα cDNA. Cell extracts were subjected to EMSA using biotinylated CXXC5-ERE (40fmol). Specific binding of ERα to CXXC5-ERE was examined in the presence (+) of excess (2500 fold) cold (unlabeled) DNA fragment with no-ERE features (5'-CTGTCAGTATCAGTTACAGGTCTA -3', upper strand)

3.3 The Binding of ERα to an Intronic ERE Sequence in the CXXC5 Gene Locus *In cellula*

After I showed that ERα binds to the potential ERE sequence in the CXXC5 gene locus *in vitro,* I conducted a chromatin immunoprecipitation assay (ChIP) to assess whether ERα also binds this CXXC5-ERE *in cellula*. I grow MCF7 cells in CD-FBS supplemented medium for 72h to provide the internalization of the steroid hormones followed by the treatment of vehicle (0,01% ethanol) or at the saturating concentration 10^{-8} M E2 for one hour. After that, cells were collected and subjected to the ChIP protocol using ERα (HC-20x) or species-specific rabbit IgG antibody overnight. PCR and qPCR were performed after DNA isolation.

I found that the DNA isolation was successfully performed in the IgG or ERα inputs' PCR amplification using primers that amplify the CXXC5-ERE containing region in the conventional PCR in both vehicle or E2 treated samples. On the other hand, CXXC5-ERE containing region could be only amplified in the E2 treated and ERα precipitated sample, indicating that ERα interacts with CXXC5-ERE in the presence of E2 (Figure 3-5A). qPCR studies further demonstrated that E2 increases the ERα binding to the CXXC5-ERE (Figure 3-5B) as it was observed in our positive control non-consensus ERE sequence of TFF1 gene [3], [52], [53] that intervenes the E2-ER responsiveness of the gene (Figure 3-5C). Our results indicate that in the presence of E2, ERα interacts with the CXXC5-ERE sequence *in cellula*.

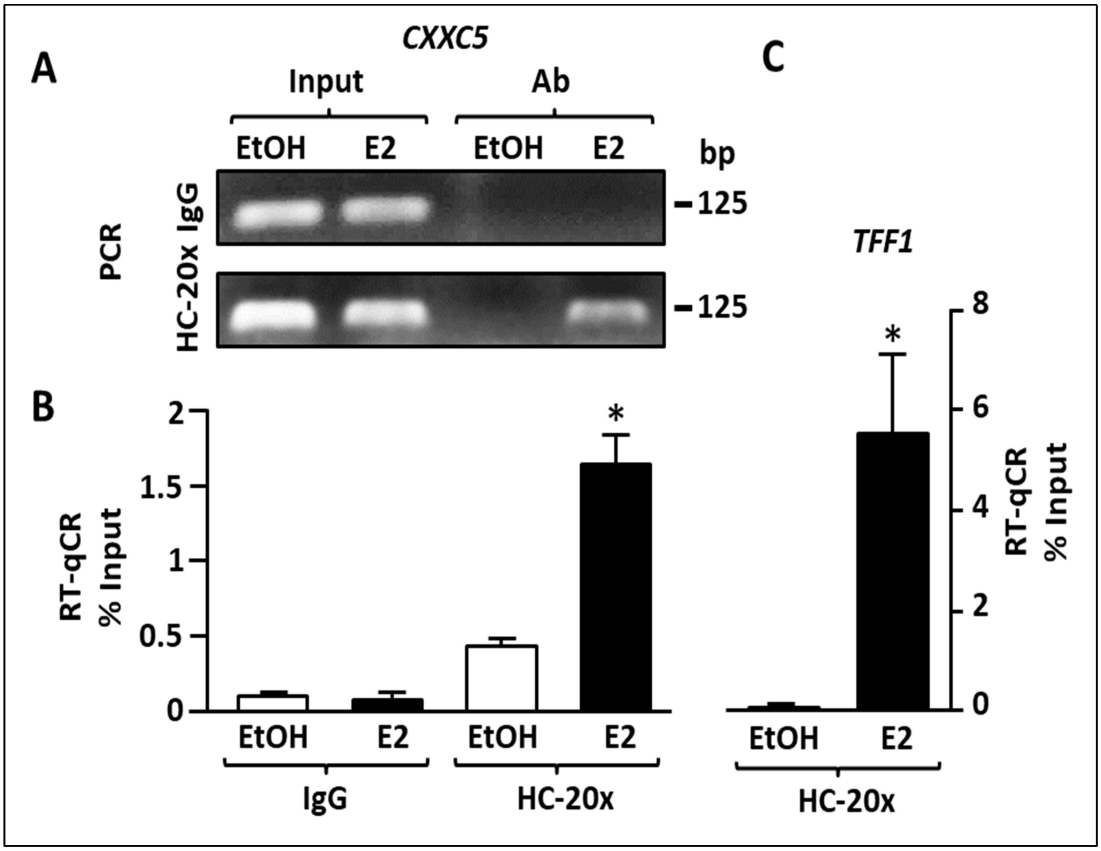

Figure 3-5 Chromatin Immunoprecipitation Assay. MCF7 cells grown in medium containing CD-FBS for 72h treated without (EtOH, 0.01%) with 10^{-8} M E2 for one h prior to ChIP. Cells were fixed with 0.75% paraformaldehyde, lysed, sonicated and subjected to ChIP using IgG or an ERα specific HC20x antibody followed by the incubation with Protein A/G conjugated magnetic beads. **A)** PCR reactions subjected to 2% agarose gel electrophoresis from a representative experiment performed three independent times. **B)** Samples were also subjected to qPCR for quantitative analysis with primers specific to the estrogen responsive region of *CXXC5*. **C)** qPCR results the estrogen responsive region of *TFF1* with the same experimental inputs described in (B) with primers specific to the estrogen responsive region of *TFF1*. Sizes of the DNA fragments in base pairs are indicated. Asterisk (*) denotes significant change depicted as percent (%) of input.

3.4 The Regulatory Effect of E2-ERα on the CXXC5-ERE Containing Region

We amplified using MCF7 genomic DNA as the template and cloned CXXC5-ERE containing region of the *CXXC5* locus (-305 to +1, +1 represents the A residue of

the encoding ATG) into a reporter vector that bears *Firefly Luciferase* ORF as the reporter enzyme. We used MCF7 cells that were grown in CD-FBS supplemented medium for 48h followed by treatment of the vehicle control (ethanol 0.01%) or 10^{-9} M E2 and/or 10^{-7} M ICI for 24h to test whether the CXXC5-ERE containing region is indeed regulated by E2-ERα complex or not. Cells were transfected with the reporter vector bearing none (Basic-Luc), CXXC5-ERE, or E2-ERα responsive region of the *Oxytocin* promoter as a control [39], [51].

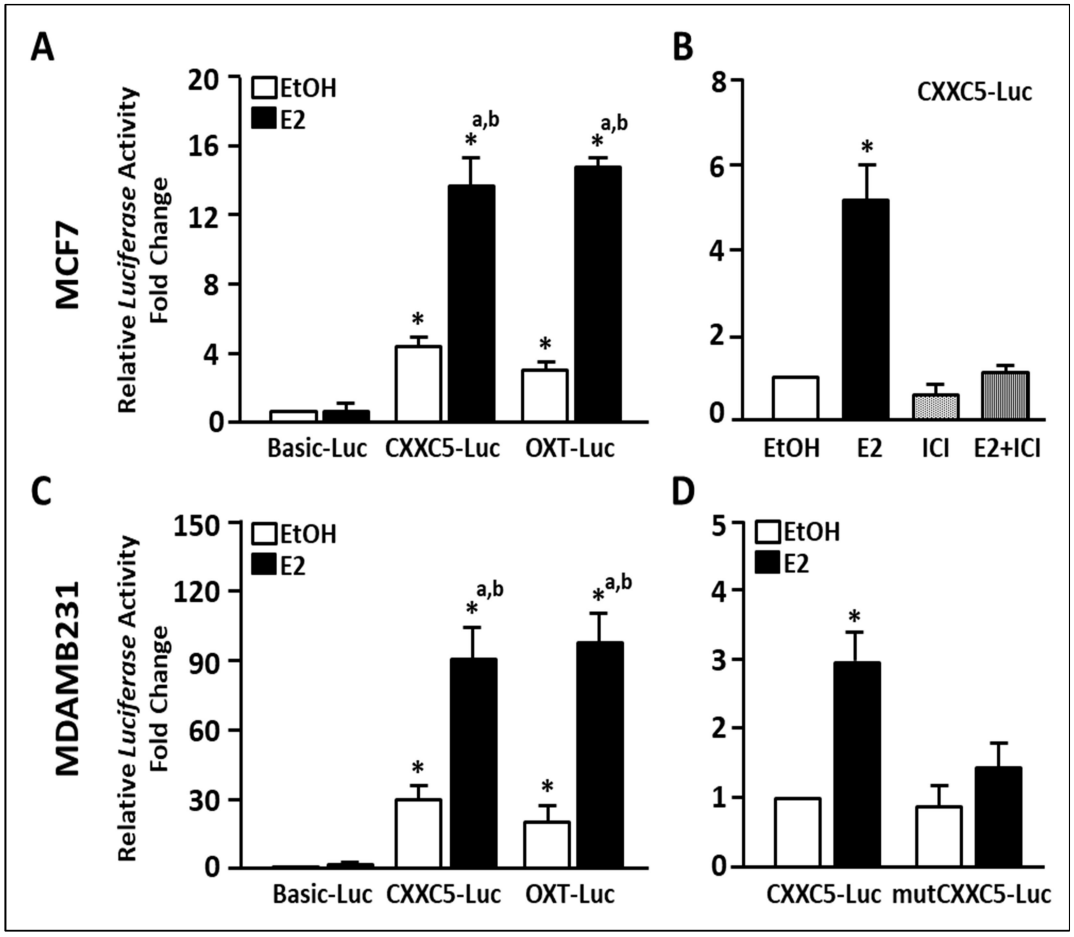

Figure 3-6 Transcriptional responses from the estrogen responsive CXXC5 region-driven gene reporter. A) MCF7 cells grown in medium containing CD-FBS for 48h were transiently transfected with pGL3 bearing none (Basic-Luc), the estrogen responsive region of *CXXC5* (CXXC5-Luc) or *OXT* (OXT-Luc) driving *Firefly Luciferase* cDNA expression as the reporter in the absence (EtOH, 0.01%) or presence of 10^{-9} M E2 for 24h.

Figure 3-6 (cont'd)
The transfection efficiency was monitored by the co-expression of pCMV–RL that drives the expression of *Renilla Luciferase* cDNA. 24h later, cellular extracts were subjected to luciferase assays. Shown is the mean ± SD of three independent experiments performed in triplicate. Firefly/Renilla luciferase activities are presented as fold change compared to EtOH control of pGL3-Basic, which is set to 1. *a and *b indicate significant difference from E2 of Basic-Luc and the corresponding EtOH control, respectively. **B)** MCF7 cells transfected with CXXC5-Luc treated without (EtOH, 0.01%) or with 10^{-9} M E2 and/or 10^{-7} M ICI for 24h were subjected to luciferase assays. Shown is the mean ± SD of three independent experiments performed in triplicate. Firefly/Renilla luciferase activities are presented as fold changes compared to EtOH, which was set to 1. **C)** MDAMB231 cells were transfected as described in (A) with Basic-Luc, CXXC5-Luc, or OXT-Luc reporter together with pCDNA-Flag-ERα vector. Cells were also co-transfected with pCMV-RL for monitoring transfection efficiency. Results are the mean ± SD of three independent experiments performed in triplicate. Firefly/Renilla luciferase activities are presented as fold changes compared to EtOH of pGL3-Basic, which is set to 1. *a indicate significant change from EtOH of Basic-Luc; while *b denotes significant change of E2 compared to EtOH of CXXC5-Luc or OXT-Luc. **D)** MDAMB231 cells were transfected with CXXC5-Luc or mutCXXC5-Luc vector, the latter which bears a mutant sequence that changes the ERE sequence in *CXXC5* to a non-ERE, together with pcDNA-Flag-ERα vector. Cells were treated without (EtOH, 0.01%) or with 10^{-9} M E2 for 24h. Shown is the mean ± SD of three independent experiments performed in triplicate. The normalized Firefly/Renilla luciferase activities are presented as fold change compared to EtOH of CXXC5-Luc, which was set to 1.

E2 treatment did not change the luciferase response in the Basic-Luc transfected cells. Basal luciferase activity of the CXXC5-Luc or OXT-Luc was significantly increased compared to responses observed with Basic-Luc transfected cells. On the other hand, in the presence of the E2, the relative luciferase activities in both CXXC5-Luc and OXT-Luc were increased compared to Basic-Luc responses (Figure 3-6A). When the cells were treated with ICI together with E2, the E2-mediated augmentation of the reporter response was completely blocked, suggesting that the E2 response of the CXXC5-Luc (or OXT-Luc, data not shown) is ER-dependent (Figure 3-6B). To investigate the ERα dependency of the CXXC5-Luc in response to E2, we transiently co-transfected MDAMB231 cells with an expression vector bearing none (data not shown) or the Flag-ERα ORF and reporter enzyme constructs of CXXC-Luc or OXT-Luc. I observed an increase in

the luciferase activities of CXXC5-Luc and OXT-Luc in the presence of ERα when the cells were treated with E2 compared to those observed with the Basic-Luc vector-transfected cells, indicating that the augmentation of the reporter enzyme activity is indeed E2-ERα dependent (Figure 3-6C). Finally, when we introduce a mutation to the CXXC5-ERE sequence (5′-GGTCAggaTGACA-3′ changed into 5′-TTTGATCCCTCAA-3′ using overlapping PCR approach) with the remaining of the surrounding sequences are the same, the relative luciferase activity was lost, indicating that the CXXC5-ERE is the E2-ERα responsiveness region (Figure 3-6D). As a summary, our results altogether demonstrate that E2-ERα takes part in the regulation of the *CXXC5* expression through interactions with the CXXC5-ERE.

PART II: IDENTIFICATION OF THE *CXXC5* PROMOTER

3.5 The Identification of the Main *CXXC5* Transcript in MCF7 and HL60 Cells

The E2-ERα complex regulates the CXXC5 gene expression through an intronic ERE found at the 240 bp upstream of the first nucleotide of the translation initiation codon of the CXXC5 protein (Figure 3-7).

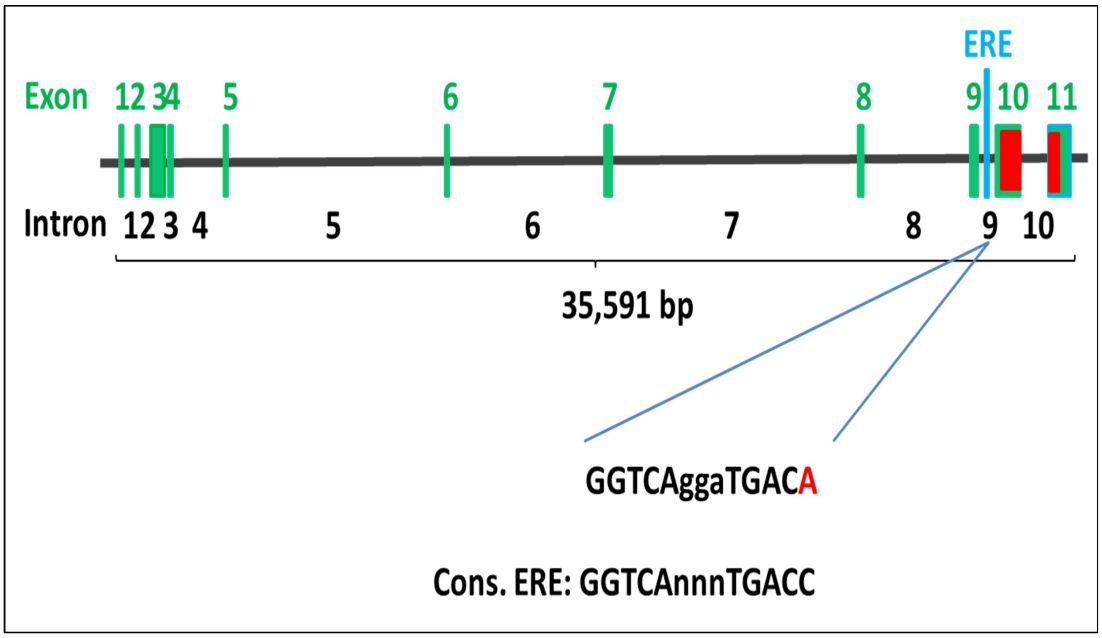

Figure 3-7 Representation of the location of the CXXC5-ERE in the *CXXC5* locus. CXXC5-ERE is found in the Intron 9 and has a one mismatch compared to consensus ERE sequence depicted in red.

Although various signaling pathways were shown to be involved in the regulation of *CXXC5* expression, including E2-ER signaling, as I showed here, how the CXXC5 gene expression is accomplished is mostly unknown. The second part of my doctoral studies aims, therefore, to elucidate mechanisms responsible for the CXXC5 gene expression by defining and dissecting features of the primarily *CXXC5* promoter elements.

CXXC5 is located on the 5q31.2, oriented on the forward strand, and encompasses 35-kb DNA. According to the ENSEMBL and NCBI genome databases, 14 protein-coding transcript variants in different tissues are identified by the Expressed Sequence Tags (ESTs) approach, which is widely utilized to identify alternative splicing products of genes [54], [55]. There are 11 different exons and 10 introns, depicted in Figure 3-8, describing the exon usage for the *CXXC5* transcript variants. The last two exons (Exon 10 and 11) on the gene locus are found in all 14 TVs and contain the coding region of 969 nucleotides, 924 bp of Exon 10, and 45bp, including a stop codon, is in Exon 11. All 14 TVs of *CXXC5*

encode the same protein with a calculated molecular mass (MM) of 33 kDa. Although some of the transcript variants display small nucleotide changes at their 3'UTR (untranslated region), the transcript diversity is mostly caused by the 5' ends of the variants. The 5' end heterogeneity could be the result of alternative promoter usage and/or alternative splicing [56]–[58].

To reveal the promoter(s) elements of the *CXXC5*, we initially wanted to identify which of the variants found in different databases (ENSEMBL, NCBI, UCSC) are present in our cell models through cloning, quantification, and northern blotting studies. Therefore, we wanted to assess the expression of the *CXXC5* transcript variants in MCF7 and the HL60 cell line derived from acute promyelocytic leukemia, and in both of which *CXXC5* contributes to phenotypic features of the cell models [9], [14]. We isolated total RNA from the cells, and in collaboration with Kerim Yavuz, M.Sc. we cloned, sequenced the *CXXC5* variants, and quantified their expressions with RT-qPCR. Results revealed that the CXXC5-TV2 consisting of Exon 3, 10, and 11 is the primary transcript in both MCF7 and HL60 cells (Kerim Yavuz, Utilization of Transcript Variants of CXXC5 Gene To Predict Its Promoter, Master Thesis, METU, 2020).

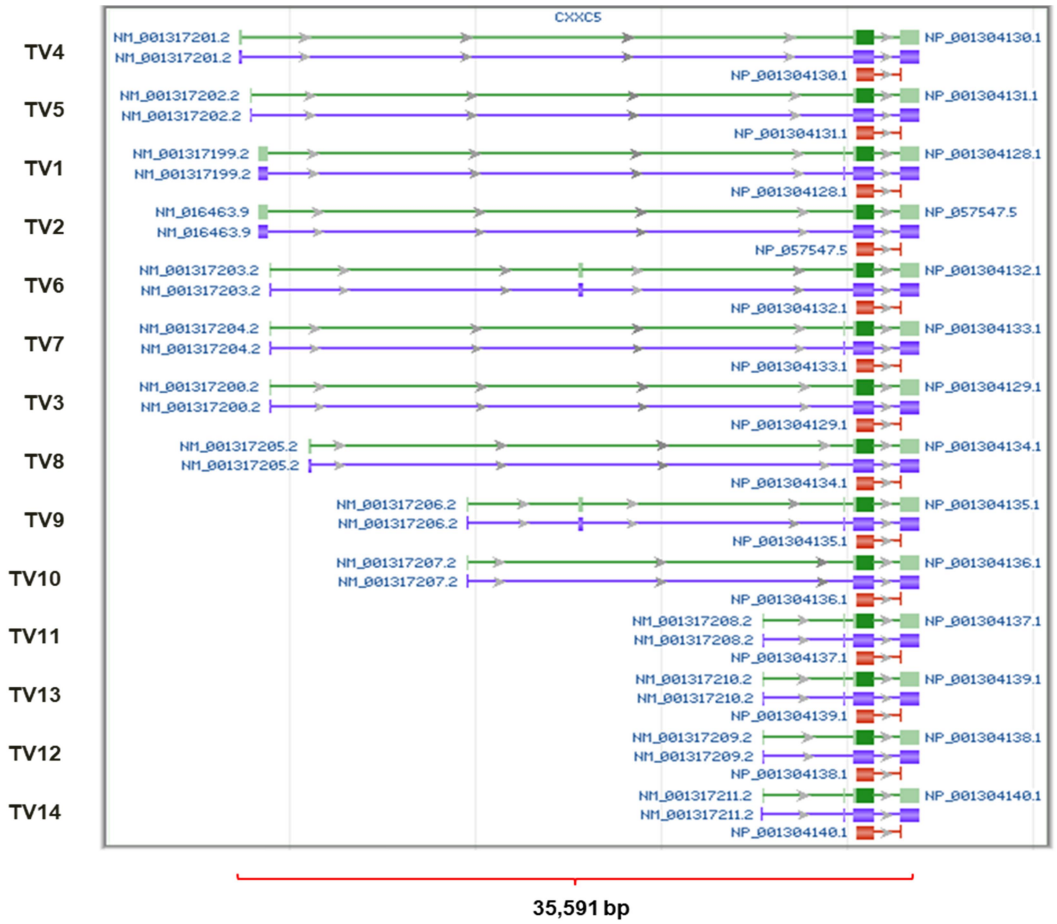

Figure 3-8 Representation of the Exon Usage of the *CXXC5* Transcript Variants. Annotation of the 14 transcript variants of the CXXC5 gene is represented in Genome Data Viewer of NCBI format.

I also verified that the expression and the presence of the CXXC5-TV2 in our cell models. I performed a Northern Blot (NB) analysis by hybridizing the biotinylated probes with the ribosomal-RNA (rRNA) depleted RNA samples. I used two different probes to detect the *CXXC5* transcript; one targets the Exon10-11, a common region for all the variants (Figure 3-9A), or Exon3, the first exon of the main transcript (Figure 3-9B). The *CXXC5* transcript variants' sizes vary between 2309-2776; I observed one prominent RNA band around 2700 nucleotides which is about the size of the CXXC5-TV2: 2682 nucleotides. My NB findings further support the conclusion that the CXXC5-TV2 is the primary transcript in both

MCF7 and HL60 cell lines. *GAPDH* transcript variant 4 (NCBI, Accession: NM_001289746.2, 1525 nucleotides linear mRNA) targeting biotinylated probe was used as a positive control which was observed around 1500 nucleotide according to RNA ladder (Figure 3-9C).

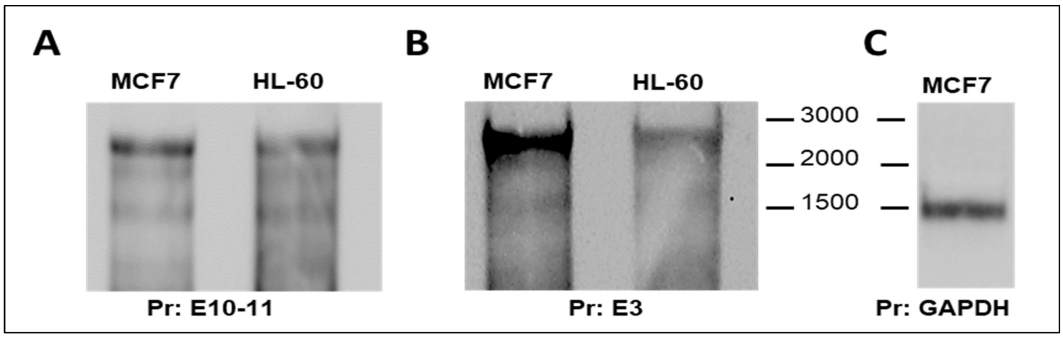

Figure 3-9 Northern blot analysis of the main *CXXC5* transcript variant. RNA samples were subjected to NB using biotinylated probes targeting **A)** a shared region by all variants which complementary to Exon 10-11 or **B)** a Exon 3-specific region. **C)** A *GAPDH* targeting probe was also used for the detection of *GAPDH* transcripts as a control. The molecular marker (nt) is indicated.

3.6 Characterization of the 5'and 3' ends of the main *CXXC5* transcript variant

3.6.1 5' Rapid Amplification of cDNA Ends (5'RACE)

Since annotation of promoters in the human genome relies on the experimental evidence of 5′-ends of mRNA transcripts, which primarily correspond to the transcription start sites [59], [60], I predicted that the identification of TSS(s) of the main transcript variant could be a starting point for the deciphering of the promoter region(s) of *CXXC5*.

For the identification of TSS(s), I used the 5' Rapid Amplification of cDNA Ends (5'RACE) approach, designed for the amplification of nucleic acid sequences from a messenger RNA (mRNA) template between a defined internal site and unknown

sequences at the 5'-end of the mRNA through the use of an adaptor RNA probe [61]. Although prone to biases introduced by various factors, including RNA secondary structures, G-C nucleotide content, adaptor ligation efficiency [62], 5'RACE has been successfully used to identify 5'-ends of numerous RNA transcripts [63]. We also used TFF1, a well-studied estrogen-responsive and TATA-box promoter-driven gene [21], [52], [64], as a control for 5'RACE studies.

In previous *in silico* studies in collaboration with Gizem Kars M.Sc., we revealed that the first exon of the CXXC5-TV2, Exon 3 is found within a CpG island, and it has a high GC content (>70%) (Gizem Kars, Epigenetic Characterization of CXXC5 Gene Locus, Master Thesis, METU, 2018). Likely due to the high GC content of the Exon 3 and surrounding sequences 5'RACE of CXXC5-TV2, in contrast to *TFF1,* which generates a transcript with a single TSS (Figure 3-10A) proved to be complicated. Nevertheless, our results based on the sequencing of PCR amplicons generated from 5'RACE-cDNA libraries of MCF7 cells indicated that several 5'-ends of the *CXXC5* transcript could be detected, suggestive of multiple TSSs (Figure 3-10B). These results suggest the presence of a transcription start region for CXXC5-TV2 rather than a distinct TSS.

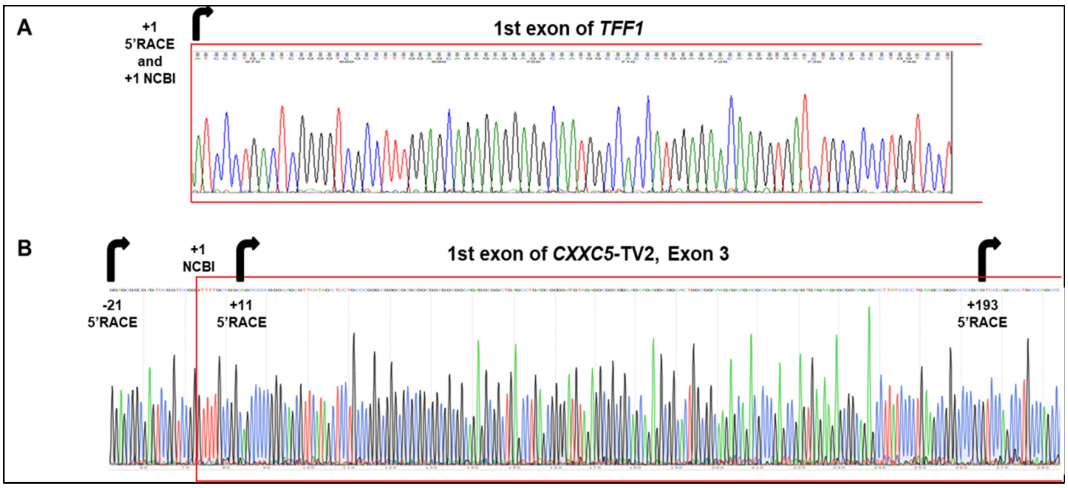

Figure 3-10 Transcription Start Sites of the *TFF1* and CXXC5-TV2. The chromatography images of the sequencing results are represented. **A)** There is only one transcription start site for *TFF1* as our 5'RACE results correlates with the NCBI annotated initiation site. **B)** There are multiple transcription start sites for CXXC5-TV2. The NCBI annotated start site is represented as +1 and the 5'RACE results were shown accordingly as -21, +11, and +193. The arrows indicate the transcription start sites for the genes.

Collectively, results suggest that CXXC5-TV2, composed of Exons 3, 10, and 11 is the main transcript with multiple TSSs, a transcription start region.

3.6.2 3' Rapid Amplification of cDNA Ends (3'RACE)

I also identify the 3'end of the main transcript of the *CXXC5* using 3'RACE. Primers that were targeting the common region of *CXXC5* variants were used in the 3'RACE studies. Since the 3'UTRs of the transcript variants of *CXXC5* are highly conserved and representing negligible variations, I observed consistently the same 3'end in the biological replicates (Figure 3-11A). I also showed one 3'end for the *TFF1* gene in biological replicates (Figure 3-11B).

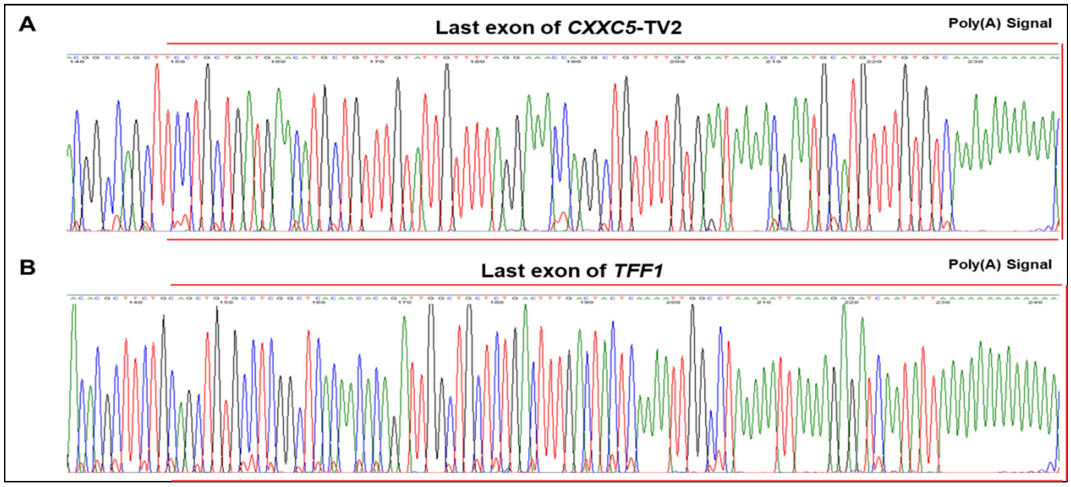

Figure 3-11 Termination Sites of the *TFF1* and *CXXC5*. For the investigation of the termination of the *TFF1* and *CXXC5*, 3'RACE studies were conducted. The chromatography images display the ends of the transcripts together with the Poly(A) signal.

3.7 Promoter Identification of the *CXXC5*

After the identification of the transcription start region of the CXXC5-TV2, I cloned 1548 bp long region including all the TSSs which were found in the

5’RACE studies together with the 5’ upstream Exon 3, Exon 3, and Exon 4 into pGL3-Basic reporter vector to assess the promoter activity of the putative promoter region of *CXXC5* (Figure 3-12A). I introduced peripheral deletions from 5’ or 3’ end to decipher the core promoter region of *CXXC5*. The TFF1 gene promoter containing the estrogen-responsive region is also used as a control in the luciferase reporter assays. I transiently co-transfected the MCF7 cells with *Firefly Luciferase* vector bearing none or the putative *CXXC5* promoter regions as full-length or truncated forms together with the *CMV* promoter-driven *Renilla Luciferase* containing reporter vector. The relative luciferase activity was significantly higher in cells transfected with the reporter construct containing the putative *CXXC5* promoter region as we similarly observed from the positive control, the *TFF1* promoter (Figure 3-12B).

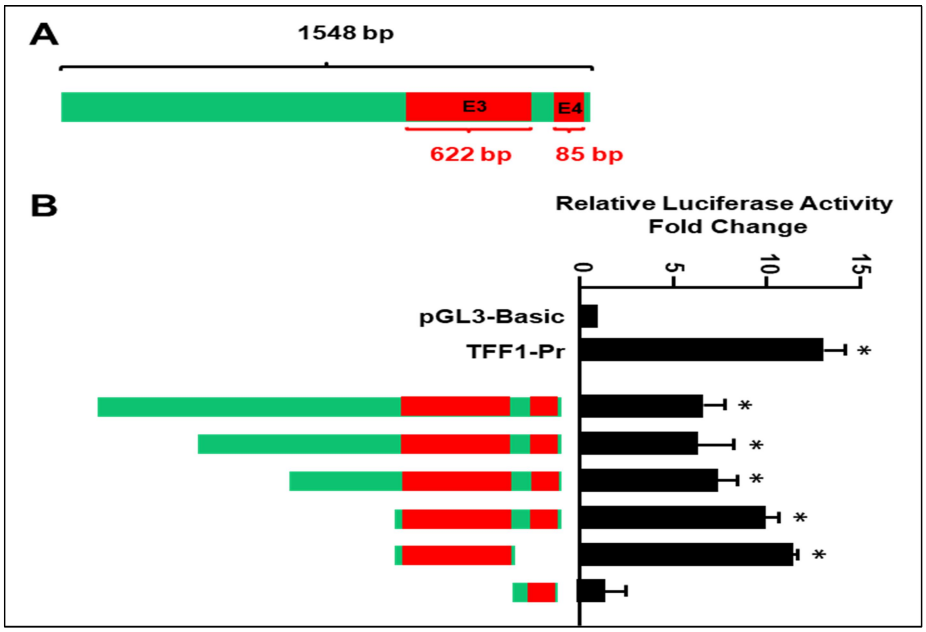

Figure 3-12 Relative Luciferase Activities of the Putative *CXXC5* Promoter Containing Regions of *CXXC5* locus. A) Schematic representation of the *CXXC5* region encompassing 3' end of Intron 2, Exon 3, and Exon 4 as the full-length template for reporter assays. **B)** MCF7 cells grown in growth medium for 48h with 10% FBS were transiently transfected with pGL3 bearing none (Basic-Luc), the promoter of estrogen-responsive TFF1 gene, or the template *CXXC5* region bearing full-length, or 5'- and/or 3'end truncated *CXXC5* regions that drive the *Firefly Luciferase* cDNA expression as the reporter enzyme for 24h. The transfection efficiency was monitored by the co-expression of pCMV-RL that drives the expression of the *Renilla Luciferase* cDNA. Cellular extracts were then subjected to dual-luciferase assays. Results are shown in the mean ± SD of three independent experiments performed in triplicate. Firefly/Renilla luciferase activities are presented as fold change compared to pGL3-Basic, which is set to one. Asterisk (*) indicates a significant difference from the Basic-Luc control.

Results revealed that the Exon 3-alone containing construct resulted in the highest relative luciferase activity (Figure 3-12B). Therefore, I continued further internal and/or peripheral deletions of the Exon 3 sequence as Segments of A to D to increase the resolution of the core elements of the *CXXC5* promoter. Additional deletions indicated that the 5' end of the Exon 3, corresponding to Segment A, retains the promoter activity (Figure 3-13).

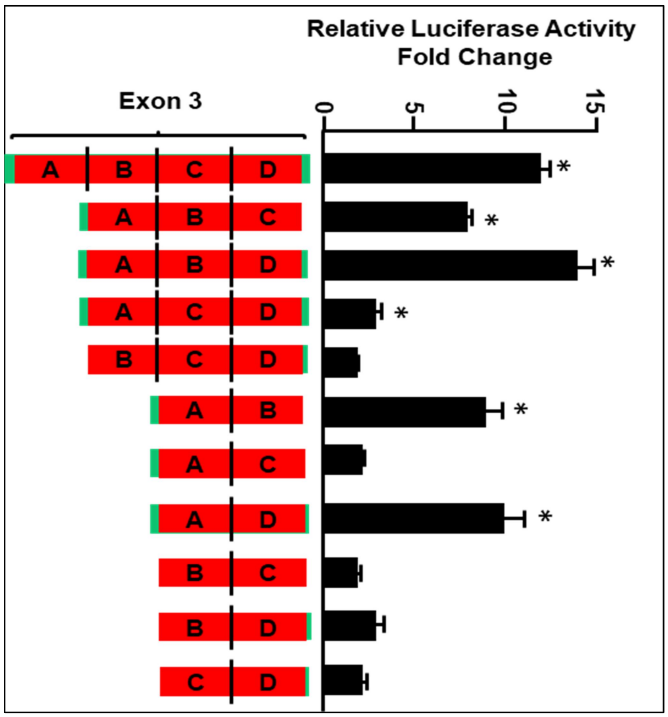

Figure 3-13 Relative Luciferase Activities of the Exon 3 and the truncated forms of Exon 3. MCF7 cells were grown and transiently transfected with reporter enzyme vectors. Exon 3 with four segments (A-D) were cloned as the full-length, or various combinations of segments into the Basic-Luc reporter vector. Results are shown in the mean ± SD of three independent experiments performed in triplicate. Firefly/Renilla luciferase activities are presented as fold change compared to pGL3-Basic, which is set to one. Asterisk (*) indicates a significant difference from the Basic-Luc control.

While Segment A increased the transcriptional activity, in the absence of Segment A, transcriptional activity is decreased (Figure 3-13). On the other hand, when Segment A and Segment C are cloned together, Segment A mediated increased relative luciferase activity is lost due to Segment C. Although Segment C seemed to have a repressive effect on the transcriptional activation, when Segment B is present, Segment C's suppression activity is attenuated, suggesting that Segment B has a neutralizing impact on the repressive ability of Segment C (Figure 3-13). Also, Segment A alone has a transcription activation activity; Segment C display transcription repression features (Figure 3-14A). These findings suggest that Segment C has also DNA elements negatively affecting the transcription.

To test this prediction, I genetically fused Segment C to the 3' end of the *TFF1* or the strong human *cytomegalovirus* (*CMV*) promoter. I also cloned Segment D as a negative control, which shows no significant effects on transcription activation. In contrast to Segment D, which has minimal impacts on the luciferase activity from the *CMV* promoter, Segment C dramatically repressed the transcriptional activation (Figure 3-14B).

Thus, these results collectively suggest that the core promoter elements of the *CXXC5* reside in Segment A, which consists of all three TSSs.

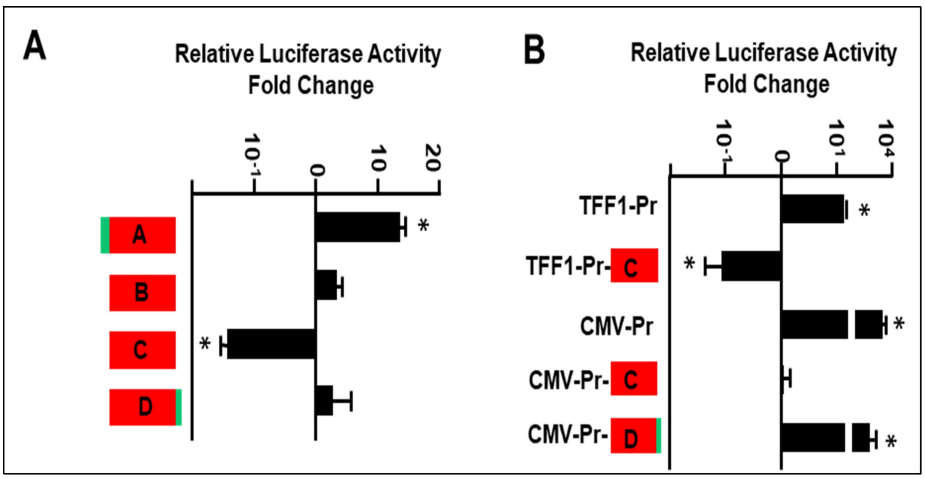

Figure 3-14 Relative Luciferase Activities of the Segments of Exon 3 Alone and Genetically Fused Segment C. MCF7 cells were grown and transiently transfected with reporter enzyme vectors. **A)** Segments of the Exon 3 were cloned alone into Basic-Luc reporter vector. **B)** Segment C or Segment D of Exon 3 was genetically fused to the 3'-end of *TFF1* or *CMV* promoter driving the expression of the *Firefly Luciferase* cDNA as the reporter enzyme. Results are shown in the mean ± SD of three independent experiments performed in triplicate. Firefly/Renilla luciferase activities are presented as fold change compared to pGL3-Basic, which is set to one. Asterisk (*) indicates a significant difference from the Basic-Luc control.

3.8 The Promoter of the *CXXC5* is Transcriptionally Active

As I located the core promoter elements of the *CXXC5*, we carried out *in silico* analyses of the genomic region of the *CXXC5* locus wherein Segment A containing Exon 3 is situated as the putative promoter region. The *CXXC5* promoter containing region of the locus shows a remarkably high (>70%) GC content and is greatly enriched CpG dinucleotide repeats (https://www.biologicscorp.com/tools/GCContent/) (Figure 3-15A). In addition, it has an asymmetric GC distribution called GC skew, which is used as a measure of the DNA strand asymmetry in the GC nucleotide distribution (http://genskew.csb.univie.ac.at/GenSkewServlet) (Figure 3-15B). These features are the characteristics of CpG islands [65]. Based on these results, the transcription start region, including Segment A, of the CXXC5-TV2 located in a CpG island, a conclusion is also consistent with the CGI annotation track. We therefore suggest that the CXXC5 gene is driven by a CGI promoter (https://genome.ucsc.edu/) (Figure 3-15C).

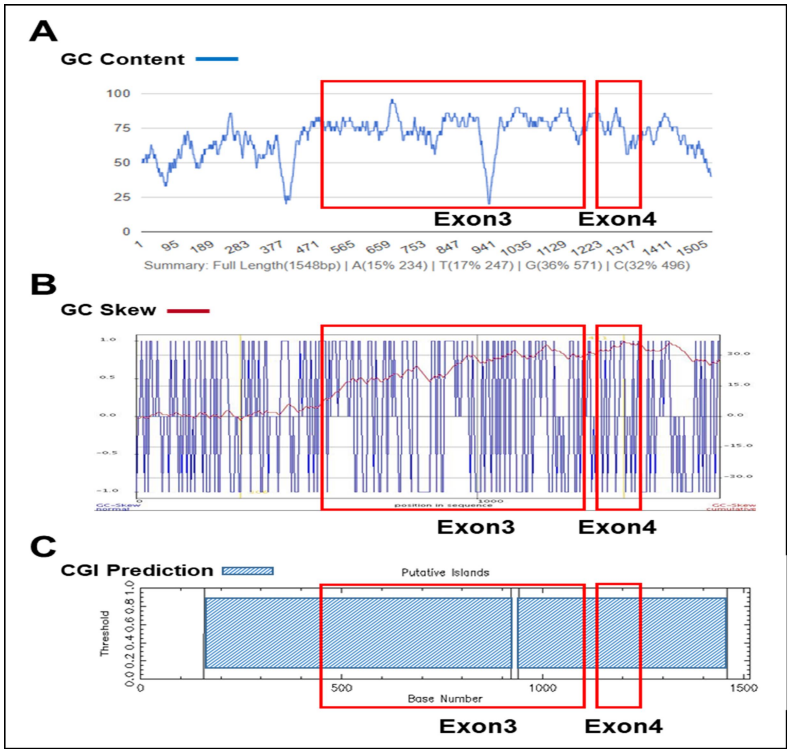

Figure 3-15 *In silico* Studies for the Exon 3 and Exon 4 that encompassing the promoter of *CXXC5*. *In silico* analyses of the *CXXC5* promoter region containing sequences for **A)** G-C content, **B)** asymmetric GC distribution, GC skew (GC skew), and **C)** the presence of a CpG island (CGI) (Gizem Kars, Epigenetic Characterization of CXXC5 Gene Locus, Master Thesis, METU, 2018).

DNA methylation plays a critical role in gene expression regulation and, consequently, in a wide variety of biological processes in both physiological and pathophysiological states [66]. DNA methylation is a dynamic epigenetic mark, cell type or tissue-specific, and heritable. Methylated CpGs present in 80% of the CpGs in the human genome, which are found in both genic and intergenic regions [67]. Although most CpGs are methylated, about 70% of human gene promoters are associated with unmethylated CGIs [68], [69]. To examine whether the Exon 3 and the surrounding sequences, including Segment A are methylated or not, bisulfite conversion and subsequent sequencing were conducted in collaboration with Gizem Kars M.Sc. Results revealed that while the upstream and the

downstream sequences of the Exon 3 are highly methylated, the Exon 3, which contains the promoter region, is non-methylated (Figure 3-16A), suggesting that the *CXXC5* promoter is found within a transcriptionally permissive region in an CpG island (Gizem Kars, Epigenetic Characterization of CXXC5 Gene Locus, Master Thesis, METU, 2018). In addition, the nucleosome occupancy studies also

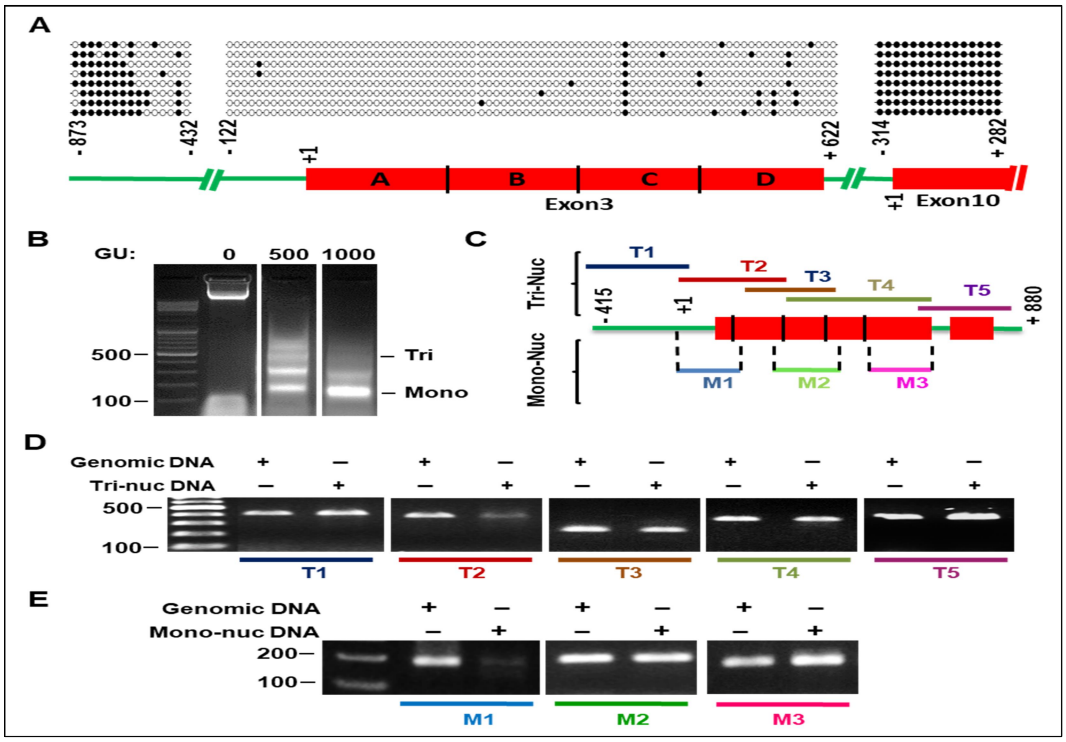

Figure 3-16 Features of the region encompassing the promoter of *CXXC5*. A) The methylation state of the *CXXC5* promoter region (-873 to -473 and -122 through the Exon 3; +1 indicates the beginning of Exon 3) together with the 3'-end of Intron 9 and Exon 10 (-195 through +330; +1 marks the beginning of Exon 10) as controls was examined with bisulfite sequencing. Isolated genomic DNA of MCF7 cells was subjected to bisulfite reaction for the conversion of unmethylated cytosine residues to uracil followed by bisulfite PCR. PCR amplicons produced with bisulfite primers were cloned and sequenced. Aligned sequences to the corresponding *CXXC5* regions were depicted as a lollipop distribution. Filled circles indicate methylated and empty circles denote non-methylated CpG dinucleotides. **(B-E)** Nucleosome occupancy at the *CXXC5* promoter elements was assessed with Micrococcal Nuclease (MNase) assay **B)**. MCF7 cells were fixed, permeabilized, and treated without (0) or with 500 or 1000 gel units (GU) of MNase for 30 min at 37°C for chromatin digestion. Isolated DNA was analyzed with agarose gel electrophoresis. **C)** Representation of the *CXXC5* regions that are monitored. Isolated DNA fragments corresponding to **D)** tri-nucleosomal (T1-5) and **E)** mono-nucleosomal (M1-3) DNAs were subjected to PCR using region-specific primer pairs (Gizem Kars, Epigenetic Characterization of CXXC5 Gene Locus, Master Thesis, METU, 2018).

revealed that the 5'-surrounding sequences of Exon3 and Segment A are nucleosome-deficient, and the remaining segments of Exon3 contain nucleosomes (Figure 3-16B-E).

To confirm that Segment A is indeed nucleosome-deficient; MCF7 cells processed for chromatin digestion by the use of MNase described for nucleosome occupancy were subjected to ChIP using antibodies specific to unmodified H3 (Figure 3-17A) or tri-methylated histone H3 lysine 4, H3K4me3, (Figure 3-17B), a histone modification used as a marker for actively transcribed genes [70]. Purified DNA was then subjected to qPCR using primers specific to Segments of Exon 3. We found that Segment A is devoid of H3 but the remaining segments of Exon 3 bear H3 decorated with K4me3 modification.

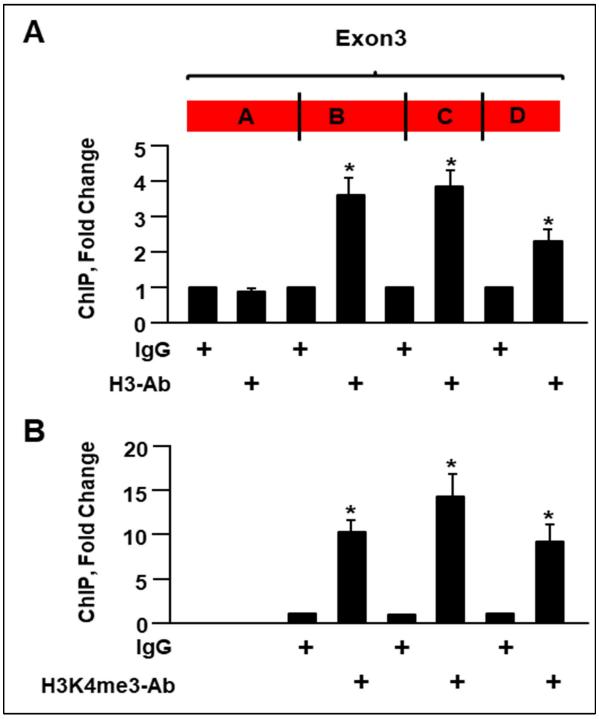

Figure 3-17 ChIP-qPCR Results of the H3 and H3K4me3. Following chromatin digestion of MCF7 cells by the use of MNase, chromatin was subjected to ChIP using species-specific IgG, an antibody specific to **A)** unmodified H3 or **B)** H3K4me3 (Gizem Kars, Epigenetic Characterization of CXXC5 Gene Locus, Master Thesis, METU, 2018).

I also performed ChIP assays using antibodies against RNA Polymerase II (RNAP II) and Serine 5 phosphorylated RNAP II, which is a representative of the active

RNAP II. Modifications of the C-terminal repeat domain (CTD) of the RNAP II, which is the largest subunit, particularly phosphorylation, determines the fate of the RNA generation as well as processing [71]. During the promoter clearance and RNA transcription, Ser5 residues are phosphorylated [71]; therefore, it is used as a marker of the regions where the transcription initiation occurs. I observed the active RNAP II as well as the RNAP II presence at Segment A, as we observed in the positive control *GAPDH* promoter (Figure 3-18).

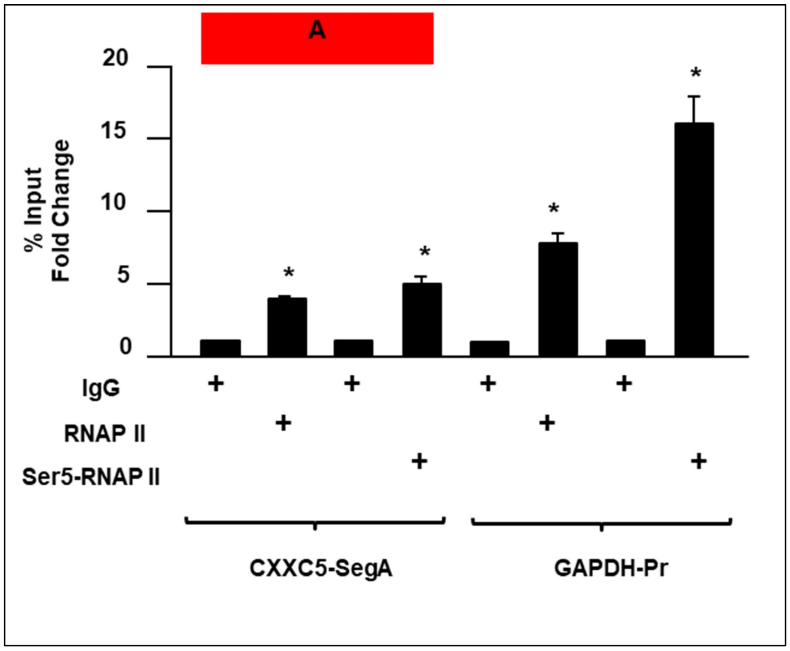

Figure 3-18 ChIP-qPCR Results of the RNAPII and Ser5P-RNAPII. MCF7 cells were sonicated and chromatin is subjected to ChIP using species-specific IgG, an antibody specific to RNAPII or Ser5P-RNAPII.

Our results collectively indicate that Segment A of Exon 3 constitutes the core promoter element of the CXXC5 gene located in a CGI and this promoter is transcriptionally active.

3.9 Identification of the *CXXC5* Core Promoter Binding Proteins

To initially evaluate proteins that potentially engage with the core promoter of *CXXC5*, I used a pull-down approach. Nuclear extracts of MCF7 cells were incubated overnight with a 5'-end biotinylated PCR amplicon (232 bp in length) containing Segment A or a fragment of Exon 10 as the control, followed by incubation with streptavidin-conjugated magnetic beads. Proteins bound to beads/DNA were then subjected to MS. Subtractive analysis of MS results obtained with proteins bound to beads, the control DNA, and Segment A revealed 94 proteins that specifically associate with the core *CXXC5* promoter (Appendix B). Analyses using STRING v11 [46] and DAVID [47] databases suggest that Segment A associated-proteins are mainly grouped in the regulation of gene expression, which can further be sub-grouped into proteins as transcription factors (TFs) and transcription co-regulatory proteins as well as proteins involved in histone/chromatin, DNA, and RNA processing (Figure 3-19).

Proteins identified as transcription factors include AFF1, ATF7, CCGBP1, CREB1, ELF1, MAZ, MGA, MYNN, NF1A, NF1B, PRDM10, TFAP2C, TPAP4, ZBTB2, ZBTB7A, ZBTB7B, ZNF596, and ZNF625. Transcription co-regulatory proteins comprise ANKRD12, ATXN7, BCOR, BRD2, BRD3, CBX8, MTA1, RBBP6, RB1, TADA2B, TRRAP, and WIZ. The group of proteins involved in RNA processing includes BUD31, CNOT1, DDX41, DDX49, DDX50, DDX54, RANBP2, and YBX1. The protein group associated with chromatin/histone binding, modifications, and organization as well as DNA conformational changes encompasses GATAD2A, INO80, JMJD1C, KAT2A, KDM1A, KDM2A, MCRS1, ORC5, RPA2, TAF6, TAF6L, and TOP3A.

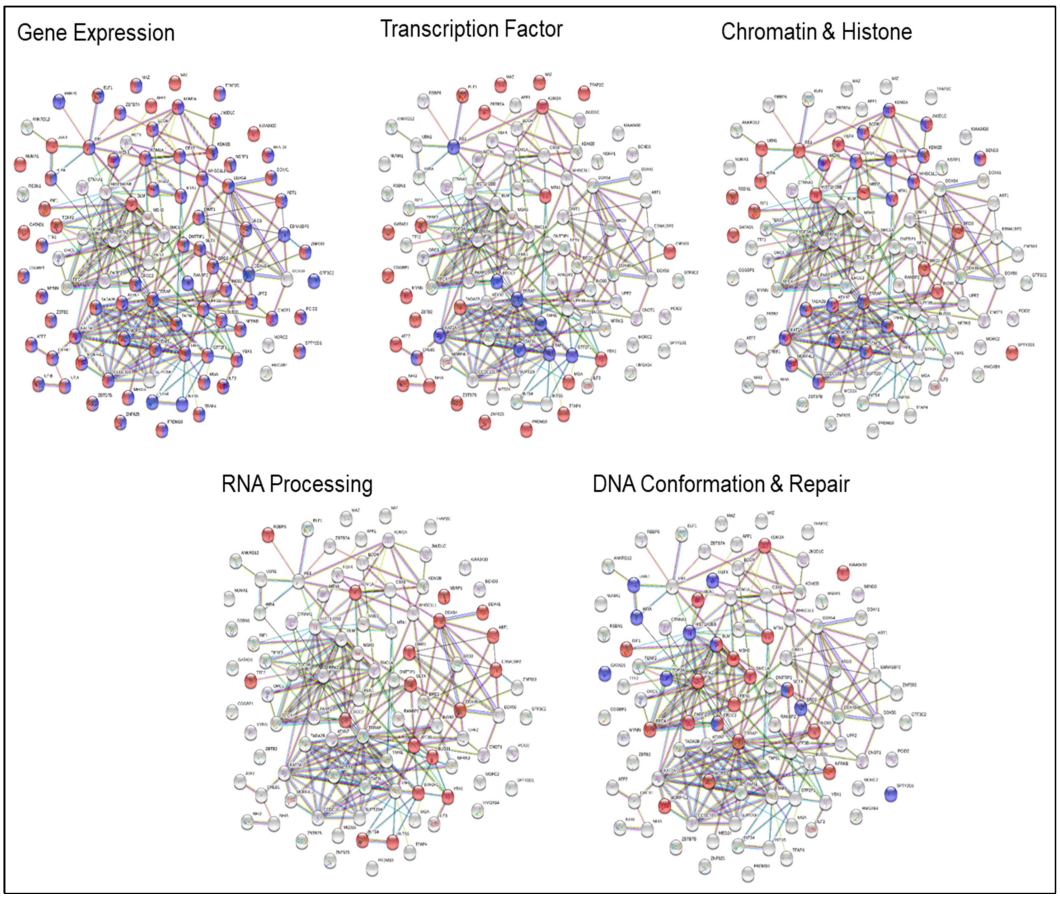

Figure 3-19 GO term enrichment analyzes of proteins associated with Segment A as the core *CXXC5* promoter element using the STRING interaction database. Nuclear extracts of MCF7 cells were incubated overnight with a 5'-end biotinylated PCR amplicon containing Segment A or a fragment of Exon 10 as the control followed by precipitation with streptavidin-conjugated magnetic beads. Bound proteins were then subjected to MS for protein IDs. Proteins are grouped under the headings of Gene Expression [GO: 0010467 Regulation of Gene Expression (red) and GO: 0010468, gene expression (Blue)]; Transcription Factor [GO: 0000981, PolII-specific DNA-binding transcription factor activity (red) and GO: 0090575, PolII transcription factor complex (blue)]; Chromatin and Histone [GO: 0006325, chromatin organization (red) and GO: 0016570, histone modification (blue)]; RNA [GO: 0006396, RNA processing (red)]; and DNA [GO: 0006281, DNA repair (red) and GO: 0071103, conformational changes (blue)].

To initially assess the binding of 15 TFs obtained with the promoter pull-down approach as the putative binders to sequences of Segment A, we carried out bioinformatics analyses using the Cistrome database, a resource of human and

mouse cis-regulatory information derived from ChIP-seq, DNase-seq, and ATAC-seq chromatin profiling assays to map the genome-wide locations of transcription factor binding sites [49]. Due to the availability of information on TFs in the Cistrome database, the possible association of 15 TFs (AFF1, ATF7, CREB1, ELF1, MAZ, MGA, MYNN, NFIA, NFIB, PRDM10, TFAP2C, TFAP4, ZBTB2, ZBTB7A, and ZBTB7B) with Exon 3 and surrounding sequences were analyzed with datasets generated by the use of MCF7 cells and/or of other cell lines. While ATF7, CREB1, MGA, MYNN, NFIA, NFIB, ZBTB2, and ZBTB7B do not appear to interact with the Exon 3 region, the association of ELF1, TFAP4, or TFAP2C with the Exon 3 in cells seems to be dependent on tissue-of-origin. On the other hand, AFF1, MAZ, PRDM10, RB1, TFAP2C, TFAP4, and ZBTB7A could be involved in the regulation of *CXXC5* expression in MCF7 and also other cells by interacting with the Exon 3 region. In collaboration with Kerim Yavuz, M.Sc. and Tolga Can Ph.D. layered images of the Cistrome analyses of CREB1, ELF1, and MAZ, as well as RNAP II and Ser5P-RNAP II which we previously showed the binding to *CXXC5* promoter, were obtained, and the results are represented in UCSC browser in Figure 3-20.

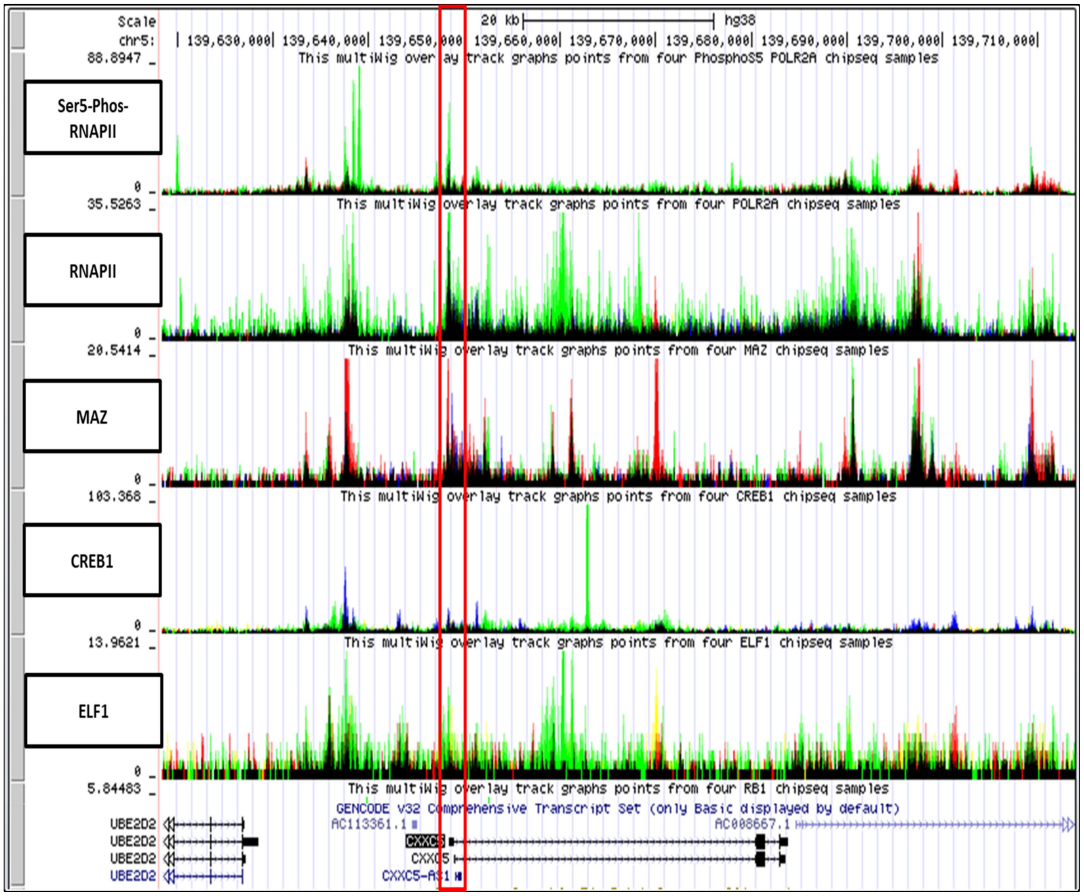

Figure 3-20 ChIP-Seq Results of the *CXXC5* Promoter Binder Proteins. Layered representation of the distinct ChIP-Seq results of Ser5P-RNAPII, RNAPII, MAZ, CREB1, and ELF1 were shown on *CXXC5* promoter. Cistrome results were visualized in UCSC Browser. Colors of Red, Green, Blue and Yellow represents different cell lines (Appendix C) for corresponding proteins and black indicates the overlapping binding regions.

3.10 Interactions of Transcription Factors with Segment A *In Cellula*

Based on our *in silico* analyses, I performed ChIP assays in cellular extracts of MCF7 cells using an antibody for ELF1 or MAZ. I also carried out ChIP for RB1, which associates with DNA indirectly through interactions with, for example, members of the E2F family proteins and with hematopoietic transcription factors, including ELF1 (PMID: 10477726), which we identified in our promoter pull-down assays as one of the Segment A interacting proteins as well. Initially, I

performed western blot analysis followed by ChIP with the corresponding TF antibodies to show the precipitation is accomplished (Figure 3-21A). To assess the possible presence of TFs on Segment A, isolated DNA samples following ChIP were subject to qPCR using primer sets specific for Segment A, for the promoter of *OAS1* as the positive control for ELF1 and promoter of *C-MYC* as for the positive control for MAZ and RB1 or Myoglobin (MB) which I used as a negative control. I also conducted ChIP using an antibody specific to CREB1 to ensure that CREB1 does not interact with the Exon 3 region as the Cistrome database suggested. Results revealed that ELF1, MAZ, or RB1 indeed associates with Segment A, as each interacts with the respective control's promoter elements but not with the MB (Figure 3-21B). CREB1 did not show an association with Segment A or the MB, but it interacts with the promoter elements of the *CCNA2* promoter as it was shown before [72].

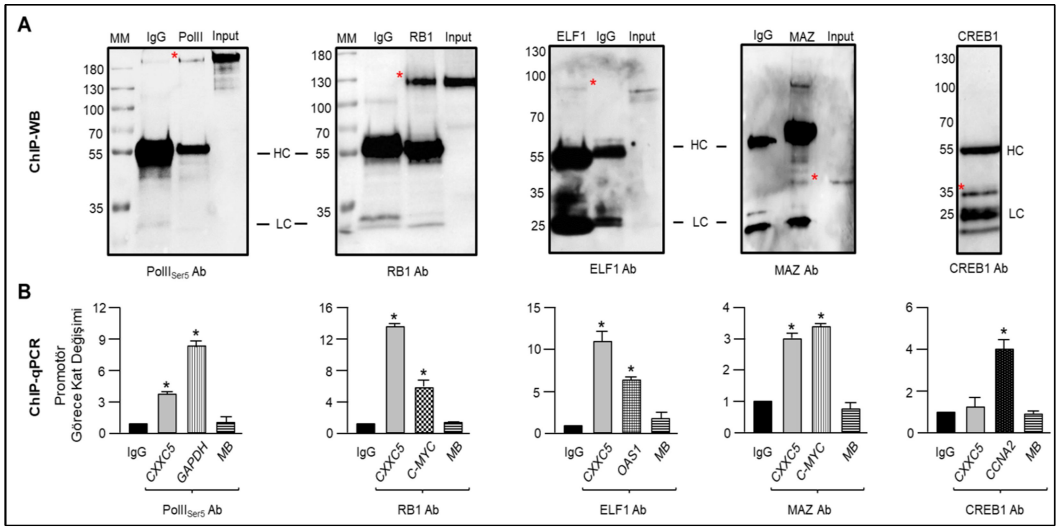

Figure 3-21 Assessing the Binding of CREB1, ELF1, MAZ, and RB1 to the Core *CXXC5* Promoter Elements *In Cellula*. A) ChIP-WB. Fragmented chromatin from MCF7 cells processed for ChIP and immunoprecipitated with a species-specific IgG or an antibody specific to CREB1, ELF1, MAZ, or RB1. One aliquot of precipitates was subjected to WB using antibodies for ChIP. Asterisk (*) indicates the protein of interest. Input, IgG together with heavy (HC) and light (LC) chains of IgG are indicated. Molecular masses (MM) in kDa are denoted. **B)** ChIP-qPCR. The remaining aliquot of ChIP was subjected to qPCR. While identical primer sets for each antibody were used in assessing the interaction of a transcription factor (TF) to Segment A as the core *CXXC5* promoter elements (CXXC5) or to the *Myoglobin* (MB) as a control, distinct primer sets were used for the promoter of a target gene of each TF. The mean ± SD of three independent experiments performed in triplicate is shown. Asterisk (*) indicates significant differences depicted as fold change compared to IgG.

3.11 ELF1 and MAZ Regulate the Expression of *CXXC5*

After I showed that ELF1 and MAZ proteins interact with Segment A, I wanted to assess these proteins' effects on the expression of the *CXXC5*. I performed overexpression using expression vectors bearing none, or the cDNA of ELF1, or MAZ; knocking-down studies using non-targeting (Allstar siRNA), ELF1, or MAZ targeting siRNAs by transient transfection into MCF7 cells. The levels of ELF1 and MAZ protein synthesis are effectively decreased when I transfected the

cells with corresponding siRNAs compared to the non-targeting siRNA (Figure 3-22A&B). Overexpressed proteins of ELF1 and MAZ were used to verify the proteins' migration on the SDS-PAGE, together with the siRNA samples.

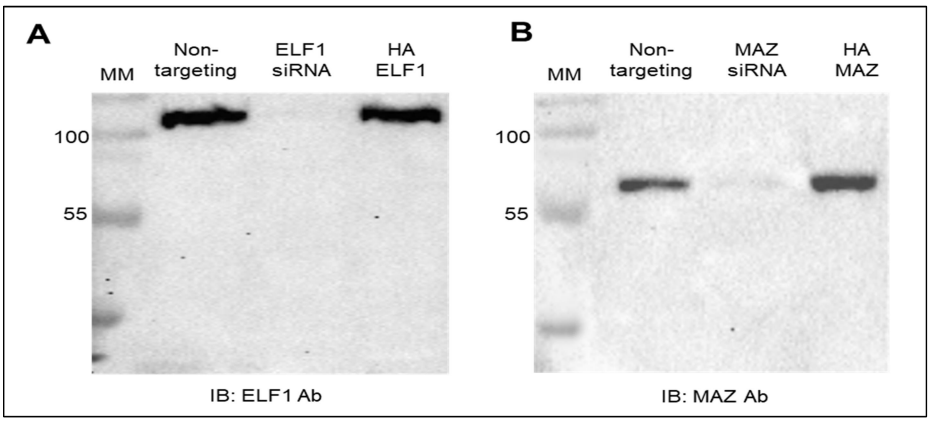

Figure 3-22 Western Blot Analyses of the Knocking-down and Overexpression of ELF1 or MAZ Proteins. MCF7 cells were transiently transfected with siRNAs of non-targeting, ELF1, or MAZ targeting, and expression vector bearing Flag-tagged ELF1, or MAZ cDNA. 50 µg nuclear extracts were loaded on 8% SDS-PAGE for ELF1 and 70 µg nuclear extracts were loaded on 10% SDS-GEL MAZ WB. **A)** sc-133096 anti-ELF1 and **B)** sc-130915 anti-MAZ antibodies used for the blotting.

Overexpression of both ELF1 or MAZ cDNAs increased the expression of *CXXC5*, as well as *OAS1* and *C-MYC*, positive controls for the ELF1 and MAZ, respectively (Figure 3-23A&B). While ELF1 targeting siRNA transfection causes a significant decrease in the expressions of both *CXXC5* and the control *OAS1*, knocking-down of the MAZ protein does not affect the expressions of either *CXXC5* or positive control *C-MYC* (Figure 3-23C&D), suggesting that there might be other mechanisms involved in the MAZ directed regulation in the cell such as potential compensation of the lack of MAZ protein by the proteins that could bind same/similar motifs as the MAZ, for example SP1 protein.

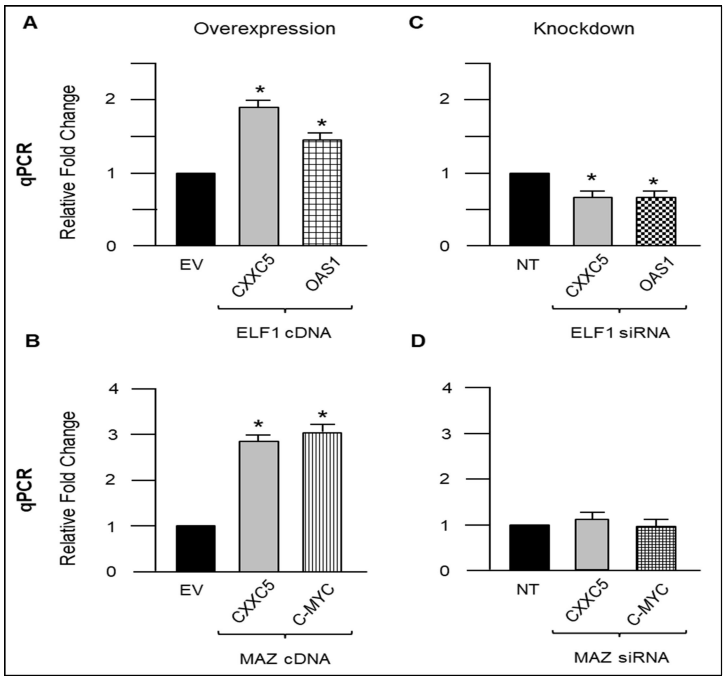

Figure 3-23. Effects of the Overexpression and Knocking-down of ELF1 or MAZ on the *CXXC5* Expression. MCF7 cells were transiently transfected with expression vectors bearing none, the ELF1 or MAZ cDNA; or siRNAs targeting non-specific, ELF1, or MAZ RNA. After total RNA isolation and cDNA conversion, samples were subjected to qPCR using transcript specific primers. Asterisk (*) indicates significant differences depicted as fold change compared to EV or NT.

Findings collectively suggest that ELF1 and MAZ participate in the regulation of *CXXC5* expression in MCF7 cells.

3.12 Sequence Motif Analyses for ELF1 or MAZ on Segment A

After I showed that both ELF1 and MAZ contribute to the expression of *CXXC5*, I examined the putative binding sites for ELF1 or MAZ proteins on Segment A. We performed sequence motif analyses using a motif analysis tool that we developed in collaboration with Çerağ Oğuztüzün and Tolga Can Ph.D., using the JASPAR database [50], which is a resource for curated, non-redundant TF-binding profiles stored as position frequency matrices (PFMs) for TFs. We identified potential binding sites for MAZ and ELF1 (Figure 3-24A) proteins in Segment A (Figure 3-24B). Moreover, one of CGI promoters' characteristics is the lack of sequence motifs for TATA-box, downstream promoter element (DPE) positioned at distinct locations relative to TSS that define non-CpG promoters [68], [69], [73]. Consistent with this, we found no such elements.

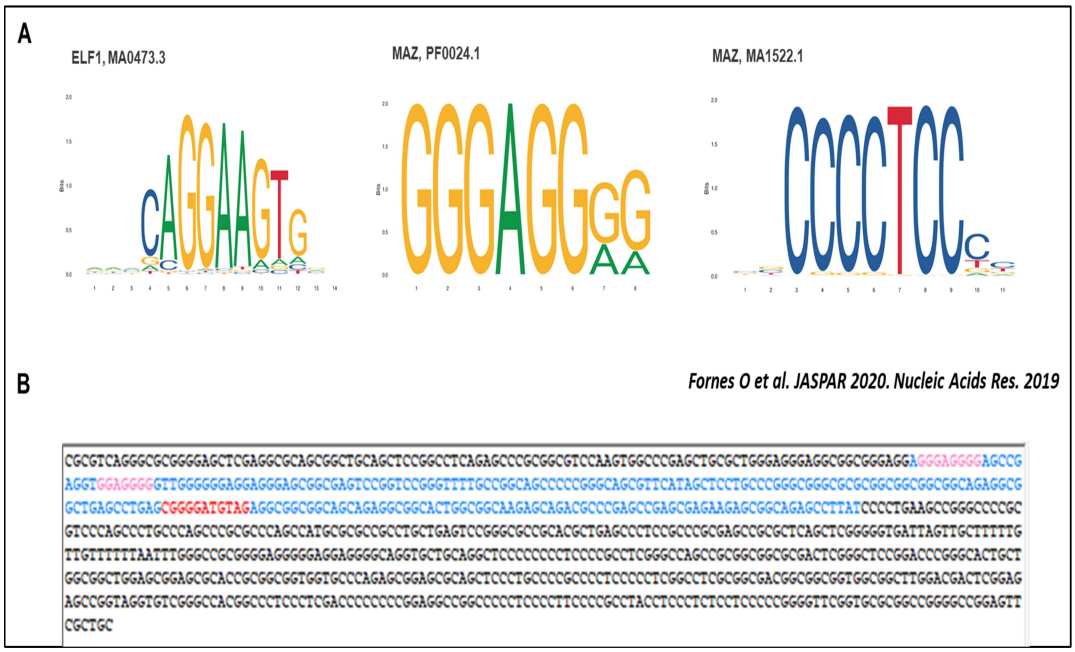

Figure 3-24. Putative Binding Motifs of ELF1 or MAZ on Segment A. A) JASPAR database used for the analyses of the ELF1 or MAZ binding motifs. **B)** Binding motif analysis tool that we developed used for the searching of the putative ELF1 and MAZ binding sites in Segment A. Putative binding sites for ELF1 or MAZ proteins in Segment A were shown in red and pink colors, respectively.

To corroborate the binding of ELF or MAZ to the *in silico* identified motif, we performed EMSA using 5'end biotin-labeled oligonucleotides together with the nuclear extracts of MCF7 cells. For both of the proteins, in the presence of the biotin-labeled oligo, retardations in the migrations are observed. The addition of the protein-specific antibody further retards the migration, which indicates the observed shifts are indeed due to the DNA-ELF1 or DNA-MAZ complexes (Figure 3-25A&B). Gel retardation is lost when we used biotin-labeled mutated DNA and/or unlabeled cold competitor motifs, which confirms that the motif sequence-protein interactions are specific (Figure 3-25A&B).

Based on these results, I also performed reporter assays. I found in transiently transfected MCF7 cells using promoter constructs bearing deleted binding motif sequences of ELF1 or MAZ that the absence of the sequence motif for the binding of ELF1 or MAZ resulted in a significant decrease in the reporter enzyme activity compared to that observed with the native sequence while they still preserve the transcription activation features compared to Basic-Luc.

These results collectively suggest that ELF1 and MAZ contribute to the expression of *CXXC5* by binding to the motifs reside in Segment A.

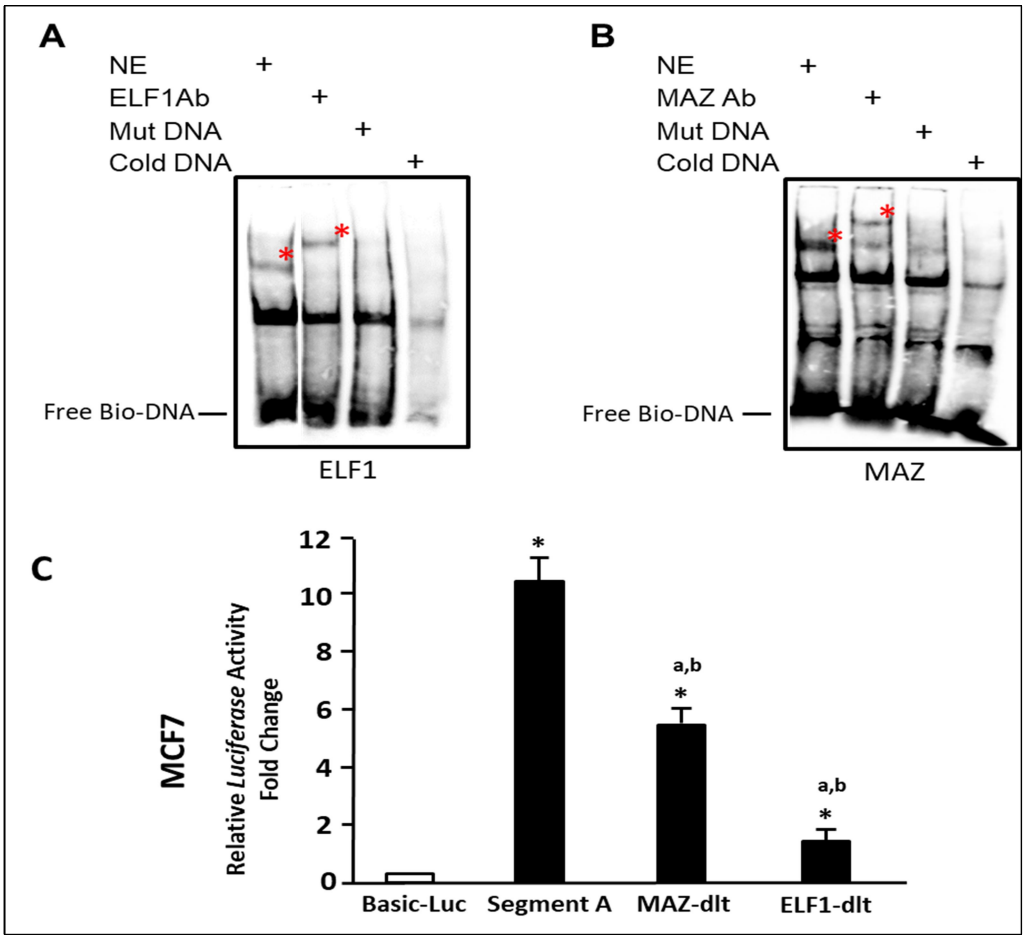

Figure 3-25 Testing the Interaction of the ELF1 or MAZ Proteins with the Binding Motifs on Segment A. 45 µg of the nuclear extracts of the MCF7 cells were used together with 40 fmol 5' biotin-labeled **A)** ELF1 or **B)** MAZ binding motif sequences found in Segment A and run on 5% PAGE. When protein specific antibodies were used due to the increased molecular weight super-shifts were observed. Mutated binding sites for ELF1 or MAZ and 250x cold competitor sequences were also used as controls and the shifts were lost as expected. A representative result from two independent determinations is shown. **C)** MCF7 cells, grown in 10% FBS containing medium 48h, were transiently transfected with pGL3 bearing none (Basic-Luc), the MAZ (MAZ-dlt) or ELF1 (ELF-dlt) motif deleted driving *Firefly Luciferase* cDNA expression as the reporter. The transfection efficiency was monitored by the co-expression of pCMV–RL that drives the expression of *Renilla Luciferase* cDNA. 24h later, cellular extracts were subjected to luciferase assays. Shown is the mean ± SD of three independent experiments performed in triplicate. Firefly/Renilla luciferase activities are presented as fold change compared to Basic-Luc, which is set to 1. *a and *b indicate significant difference from Segment A and the Basic-Luc control, respectively.

CHAPTER 4

CONCLUSION AND FUTURE DIRECTIONS

Breast cancer accounts for 25% of all cancer types that are diagnosed in women in the world. The survival rate of breast cancer is improved in the last decades with the latest developments in prognostic and therapeutic approaches. About 80% of all breast cancers are "ER-positive." It is therefore crucial to understand the underlying mechanisms of E2-ER signaling since it is one of the primary factors that participate in the initiation and progression of breast cancer. Consequently, identifying genes regulated by and/or involved in the E2-ER signaling has critical importance for establishing the new diagnostic and treatment targets and/or markers. Our previous microarray studies suggested that *CXXC5* is an E2-responsive gene and it is regulated through the ERE-dependent signaling pathway.

CXXC5 is a member of the ZF-CXXC domain protein family consisting of 12 proteins. CXXC5 is suggested to be involved in the modulation of cellular proliferation, differentiation, and death as a transcription factor, transcription co-regulator and/or chromatin modifier in response to various signaling pathways.

In this dissertation, my goal was to investigate how the CXXC5 gene expression is accomplished, in breast cancer cell line models, to provide a better understanding of the effects of the *CXXC5* expression and synthesis in physiological and pathophysiological states. To accomplish this, 1) I confirmed that *CXXC5* is indeed E2-responsive, and its regulation is ERα-dependent. An intronic ERE in the *CXXC5* locus is bound by the E2-ERα complex. 2) I identified the location of the core promoter elements of *CXXC5* and proteins engaged with these core elements participated in the regulation of *CXXC5* expression.

PART I: CXXC5 gene is regulated by E2-ERα signaling through an intronic ERE

1) In my master thesis studies, I showed that *CXXC5* mRNA levels are elevated in the presence of E2. Therefore, in my doctoral thesis studies, I first investigated whether the increased levels of *CXXC5* expression are reflected in protein levels or not using western blot analysis. To assess the protein levels of CXXC5, I initially characterized the CXXC5 ab106533 antibody using siRNA and overexpression approaches. After that, I treated MCF7 cells with or without E2 at physiological levels and conducted WB. I showed that CXXC5 synthesis is also increasing in response to E2.

2) Then, I performed *in silico* analysis to predict putative ERE sequence(s) in the *CXXC5* locus using dragon ERE finder version 3. I found a putative ERE sequence containing one mismatch compared to the consensus ERE sequence located 240 bp upstream of the translation initiation site of the CXXC5, which I called CXXC5-ERE.

3) I performed electrophoretic mobility shift assays to test whether the ERα binds this putative ERE sequence within the *CXXC5* locus or not. I transfected MDAMB231cells, ER-negative breast cancer cell line, with expression vector bearing none or Flag-tagged ERα cDNA and cellular extracts were used together with 5' biotinylated CXXC-ERE or consensus ERE. I showed that ERα binds to the CXXC5-ERE *in vitro*.

4) Next, I continued chromatin immunoprecipitation assay to investigate whether ERα binds to CXXC5-ERE in the cellular context. I treated MCF7 cells with or without E2 for one hour, followed by crosslinking and sonication of the chromatin. Then the chromatin samples were subjected to immunoprecipitation using ChIP-grade ERα antibody. After the reversal of the crosslinking and protein degradation, I recover the DNA using ethanol precipitation. DNA samples were subjected to conventional PCR and qPCR reactions using CXXC5-ERE- or TFF1-ERE-, which is found in the *TFF1*

promoter region, specific primers. As a result, I showed that ERα binds CXXC5-ERE *in cellula*.

5) After that, I tested whether the binding of ERα to CXXC5-ERE is functional or not using the reporter enzyme system. MCF7 cells were transiently transfected with *Firefly Luciferase* cDNA containing vector bearing CXXC5-ERE or ERE containing *Oxytocin* promoter region. I also transiently transfected MDAMB231 cells with reporter enzyme vectors and expression vectors bearing none or Flag-tagged ERα. Then dual-luciferase assays were performed. I showed that in the presence of E2, the relative luciferase activities resulted from CXXC5-ERE containing reporter enzyme is significantly enhanced compared to responses observed from the Basic-Luc construct. In addition, ICI, the complete antagonist of the ER, treatment blocks the E2 mediated augmentation of the luciferase activity, indicating that the transcriptional activity caused by the CXXC5-ERE is E2-responsive and ERα dependent. Therefore, the binding of ERα to CXXC5-ERE is a functional interaction.

Conclusions; A) *CXXC5* gene is an E2-responsive gene, and the E2-responsiveness of *CXXC5* is ERα-dependent. B) A distally located ERE is involved in the *CXXC5* expression.

PART II: CXXC5-TV2 is the primary transcript variant of the *CXXC5* gene in MCF7 and HL60 cells. The core promoter elements are located at the immediate 5' region and within the beginning of the first exon of the main transcript. ELF1 and MAZ proteins contribute to the *CXXC5* expression.

1) There are 14 annotated *CXXC5* transcript variants in both MCF7 and HL60 cells. In collaboration with Kerim Yavuz M.Sc., I conducted cloning and sequencing studies and showed that 8 and 9 of the 14 *CXXC5* transcript variants are expressed in MCF7 and HL60 cells, respectively. We performed qPCR assays to establish which of these expressed variants are

the main transcript(s). Our results showed that CXXC5-TV2 is the primary transcript in both MCF7 and HL60 cells.

2) Since the main transcript, CXXC5-TV2, is expressed substantially higher than the other variants, I continued investigating the transcription start site(s) of the CXXC5-TV2 as the main transcript to localize the core promoter elements of the *CXXC5* using the 5'RACE approach. I showed that there are multiple transcription start sites (TSSs) for *CXXC5* which are located upstream and downstream of the annotated transcription start site. I also performed 3'RACE studies to characterize the termination of the *CXXC5* transcript, which was relatively conserved compared to 5' UTR of the *CXXC5* transcript variants. I showed a single transcription termination site of the *CXXC5* transcript with a Poly(A) signal.

3) After identifying the transcription start region of the CXXC5-TV2, I conducted luciferase studies to locate the core promoter elements. I cloned a large region (1548 bp) of *CXXC5* locus consisting of the first exon, upstream and downstream of the CXXC5-TV2 (Exon 3). I subsequently introduced deletions 5' and 3' ends. Initial reporter enzyme studies revealed that the highest transcriptional activity is observed in the full-length first exon of the CXXC5-TV2. Therefore, I continued to introduce peripheral and internal deletions to Exon 3. Following dual-luciferase assays indicated that the initiation of the Exon 3 (Segment A) resides the core promoter elements. A region which is found in the middle of the Exon 3 (Segment C) also has regulatory element sites that show repressive activity when alone for transcription: Segment C significantly decreased relative luciferase activities caused by *TFF1* and *CMV* promoters when cloned into these promoters' 3' ends.

4) In collaboration with Gizem Kars M.Sc., I performed *in silico* analyses, bisulfite conversion, and nucleosome occupancy assays to investigate the features of the *CXXC5* promoter. Our results reveal that the *CXXC5*

promoter is GC-rich and has high GC content, indicates an asymmetry of the GC sequences called GC-skew, and is predicted to be found within a CpG island. Our bisulfite sequencing studies indicated that upstream and downstream Exon 3 are methylated, but Exon 3 and the promoter region are non-methylated, as one of the features of the CGI promoters. Then we investigated the nucleosome occupancy of the *CXXC5* promoter. Micrococcal Nuclease (MNase) assay was conducted using mono- or tri-nucleosomal regions in conventional PCR reactions. Results revealed that the *CXXC5* promoter (Segment A) seemed to be nucleosome-deficient, unlike the upstream and downstream regions in the Exon 3.

5) ChIP assays were conducted using H3 and H3K4me3 specific antibodies to confirm nucleosome-deficiency in Segment A. We showed that Segment A could not be immunoprecipitated with histone specific antibodies, verifying that the promoter region of *CXXC5* is indeed nucleosome-depleted. In addition, the surrounding areas of the promoter contain H3K4me3 modification, which is an indication of the transcriptionally active promoters. The occupancy of RNA Polymerase II (RNAP II) or Serine 5 phosphorylated RNAP II (Ser5P-RNAP II) occupancy on Segment A further supports the conclusion that the *CXXC5* promoter is transcriptionally active.

6) Then, I continued with promoter pull-down assay to identify the proteins that are engaged with the *CXXC5* promoter. 5' biotinylated *CXXC5* promoter and a control sequence which corresponds to the Exon 10 were incubated with nuclear extract overnight, and the bound proteins were analyzed with LC/MS-Mass Spectrometry. The proteins obtained from MS were subjected to the STRING and subgrouped according to their biological functions. As expected, the proteins interacting with the *CXXC5* promoter are mainly involved in gene expression regulation and composed of transcription factors, chromatin & histone modifiers, RNA processing proteins, and DNA conformation & repair proteins.

7) Transcription factors that are obtained from MS were further analyzed using publicly available ChIP-Seq datasets at the Cistrome database. ELF1 and MAZ proteins are bound to the *CXXC5* promoter in multiple ChIP-Seq datasets that stem from different tissue origins.

8) I tested the transcription factors ELF1, MAZ, or RB1, whether they indeed bind to *CXXC5* promoter in MCF7 cells using ChIP assays. Cells were grown in 10% FBS supplemented medium and collected, followed by ChIP protocol using protein-specific antibodies. Results revealed that these proteins bind to the *CXXC5* promoter *in cellula*.

9) I continued to investigate the effects of the ELF1 and MAZ on the *CXXC5* gene expression using overexpression and knocking-down approaches. I verified that the siRNAs targeting ELF1 or MAZ mRNAs are effectively working, and I can overexpress the ELF1 and MAZ proteins using expression vectors bearing the corresponding cDNAs by western blot analysis. After the verification, I performed transient transfection using either siRNAs or expression vectors and isolated the total RNA followed by cDNA conversion. cDNAs were used in qPCR assays together with *CXXC5*-specific primers. Results indicated that ELF1 and MAZ contribute to the *CXXC5* expression regulation.

10) I performed *in silico* analyses for the binding motifs of ELF1 and MAZ using the JASPAR database. The putative binding motifs of ELF1 or MAZ were searched in Segment A utilizing a tool that we developed in collaboration with Çerağ Oğuztüzün and Tolga Can Ph.D. After that, the putative binding sites were tested whether they are indeed the targets of ELF1 and MAZ proteins by electrophoretic mobility shift assays. I showed that ELF1 or MAZ binds the corresponding motif found in Segment A *in vitro*. Finally, I deleted the ELF1 or MAZ binding motifs from Segment A and cloned them into the reporter enzyme vector, and performed dual-luciferase assays. Results revealed the deletion of the binding sequence for

ELF1 or MAZ resulted in a decrease in the transcriptional activation, suggesting that both ELF1 and MAZ proteins participate in the regulation of CXXC5 gene expression by binding their binding motifs on Segment A.

Collectively, I identified the *CXXC5* promoter and confirmed that CXXC5 is an E2-responsive, ERα dependent regulated gene through an intronic ERE, which is located approximately 31 kb downstream of the transcription start region. We are currently trying to understand how the CXXC5-ERE is involved in the CXXC5 gene expression. We forsee that there a chromatin loop in the *CXXC5* locus brings the promoter and the putative enhancer region in close proximity. We will aim to elucidate the underlying mechanisms of how the distally located ERE sequence participates in regulating *CXXC5* expression using chromosome conformation capture approaches.

REFERENCES

[1] Y. Huang, X. Li, and M. Muyan, "Estrogen receptors similarly mediate the effects of 17β-estradiol on cellular responses but differ in their potencies.," *Endocrine*, vol. 39, no. 1, pp. 48–61, Feb. 2011.

[2] K. Polyak, "Science in medicine Breast cancer : origins and evolution," *J. Clin Invest*, vol. 117, no. 11, pp. 3155–3163, 2007.

[3] J. Huang, X. Li, C. a Maguire, R. Hilf, R. a Bambara, and M. Muyan, "Binding of Estrogen Receptor β to Estrogen Response Element in Situ Is Independent of Estradiol and Impaired by Its Amino Terminus," *Mol. Endocrinol.*, vol. 19, no. 11, pp. 2696–2712, Nov. 2005.

[4] J. S. Lewis-Wambi and V. C. Jordan, "Estrogen regulation of apoptosis: how can one hormone stimulate and inhibit?," *Breast Cancer Res.*, vol. 11, no. 3, p. 206, Jan. 2009.

[5] C. M. Klinge, "Estrogen receptor interaction with estrogen response elements," *Nucleic Acids Res.*, vol. 29, no. 14, pp. 2905–2919, Jul. 2001.

[6] M. Muyan, L. M. Callahan, Y. Huang, and A. J. Lee, "The ligand-mediated nuclear mobility and interaction with estrogen-responsive elements of estrogen receptors are subtype specific," *J. Mol. Endocrinol.*, vol. 49, no. 3, pp. 249–266, Dec. 2012.

[7] K. J. Hamilton, Y. Arao, and K. S. Korach, "Estrogen hormone physiology: Reproductive findings from estrogen receptor mutant mice," *Reprod. Biol.*, vol. 14, pp. 3–8, 2013.

[8] 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, May 2009.

[9] G. Ayaz *et al.*, “CXXC5 as an unmethylated CpG dinucleotide binding protein contributes to estrogen-mediated cellular proliferation,” *Sci. Rep.*, vol. 10, no. 1, p. 5971, Dec. 2020.

[10] Q. Zhang *et al.*, “Cloning and Functional Analysis of cDNAs with Open Reading Frames for 300 Previously Undefined Genes Expressed in CD34 + Cloning and Functional Analysis of cDNAs with Open Reading Frames for 300 Previously Undefined Genes Expressed in CD34 + Hematopoietic,” *Genome Res.*, vol. 10, pp. 1546–1560, 2000.

[11] H. K. Long, N. P. Blackledge, and R. J. Klose, “ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection,” *Biochem. Soc. Trans.*, vol. 41, no. 3, pp. 727–40, Jun. 2013.

[12] X. Xiong, S. Tu, J. Wang, S. Luo, and X. Yan, “CXXC5: A novel regulator and coordinator of TGF-β, BMP and Wnt signaling,” *J. Cell. Mol. Med.*, vol. 23, no. 2, pp. 740–749, Feb. 2019.

[13] C. Xu *et al.*, “DNA Sequence Recognition of Human CXXC Domains and Their Structural Determinants,” *Structure*, vol. 26, no. 1, pp. 85-95.e3, Jan. 2018.

[14] 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, Apr. 2009.

[15] X. Yan, J. Wu, Q. Jiang, H. Cheng, J. J. Han, and Y. 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, Feb. 2018.

[16] H. 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, Feb. 2014.

[17] 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, Feb. 2009.

[18] H. Kim *et al.*, "CXXC5 is a negative-feedback regulator of the Wnt/β-catenin pathway involved in osteoblast differentiation," *Cell Death Differ.*, vol. 22, no. 6, pp. 912–920, Jun. 2015.

[19] 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, Jun. 2015.

[20] 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, Mar. 2016.

[21] 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, no. 1, p. 37808, Dec. 2016.

[22] S. Choi *et al.*, "CXXC5 mediates growth plate senescence and is a target for enhancement of longitudinal bone growth," *Life Sci. Alliance*, vol. 2, no. 2, p. e201800254, Apr. 2019.

[23] M. Ravichandran *et al.*, "Rinf Regulates Pluripotency Network Genes and Tet Enzymes in Embryonic Stem Cells," *Cell Rep.*, vol. 28, no. 8, pp. 1993-2003.e5, Aug. 2019.

[24] 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, May 2017.

[25] 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,

Feb. 2013.

[26] G. Li *et al.*, "CXXC5 regulates differentiation of C2C12 myoblasts into myocytes," *J. Muscle Res. Cell Motil.*, vol. 35, no. 5–6, pp. 259–265, 2014.

[27] 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, Aug. 2016.

[28] P. A. Marshall *et al.*, "Discovery of novel vitamin D receptor interacting proteins that modulate 1,25-dihydroxyvitamin D3 signaling," *J. Steroid Biochem. Mol. Biol.*, vol. 132, no. 1–2, pp. 147–159, Oct. 2012.

[29] 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, Jun. 2009.

[30] X. Wang *et al.*, "CXXC5 Associates with Smads to Mediate TNF-α Induced Apoptosis," *Curr. Mol. Med.*, vol. 13, no. 8, pp. 1385–1396, Aug. 2013.

[31] S. Knappskog *et al.*, "RINF (CXXC5) is overexpressed in solid tumors and is an unfavorable prognostic factor in breast cancer," *Ann. Oncol.*, vol. 22, no. 10, pp. 2208–2215, Oct. 2011.

[32] P. May-Panloup *et al.*, "Molecular characterization of corona radiata cells from patients with diminished ovarian reserve using microarray and microfluidic-based gene expression profiling," *Hum. Reprod.*, vol. 27, no. 3, pp. 829–843, Mar. 2012.

[33] D. L'Hôte *et al.*, "Discovery of novel protein partners of the transcription factor FOXL2 provides insights into its physiopathological roles," *Hum. Mol. Genet.*, vol. 21, no. 14, pp. 3264–3274, Jul. 2012.

[34] M. B. Treppendahl, L. Möllgård, E. Hellström-Lindberg, P. Cloos, and K.

Grønbaek, “Downregulation but lack of promoter hypermethylation or somatic mutations of the potential tumor suppressor CXXC5 in MDS and AML with deletion 5q,” *Eur. J. Haematol.*, vol. 90, no. 3, pp. 259–260, Mar. 2013.

[35] F. Centritto *et al.*, “Cellular and molecular determinants of all- trans retinoic acid sensitivity in breast cancer: Luminal phenotype and RAR α expression,” *EMBO Mol. Med.*, vol. 7, no. 7, pp. 950–972, Jul. 2015.

[36] I. Benedetti, A. M. De Marzo, J. Geliebter, and N. Reyes, “CXXC5 expression in prostate cancer: implications for cancer progression,” *Int. J. Exp. Pathol.*, vol. 98, no. 4, pp. 234–243, Aug. 2017.

[37] X. Chen and X. Wang, “The KN Motif and Ankyrin Repeat Domains 1 / CXXC Finger Protein 5 Axis Regulates Epithelial-Mesenchymal Transformation , Metastasis and Apoptosis of Gastric Cancer via Wnt Signaling,” 2020.

[38] M. Muyan *et al.*, “Modulation of Estrogen Response Element-Driven Gene Expressions and Cellular Proliferation with Polar Directions by Designer Transcription Regulators,” *PLoS One*, vol. 10, no. 8, p. e0136423, Aug. 2015.

[39] J. Huang, X. Li, P. Yi, R. Hilf, R. A. Bambara, and M. Muyan, “Targeting estrogen responsive elements (EREs): design of potent transactivators for ERE-containing genes,” *Mol. Cell. Endocrinol.*, vol. 218, no. 1–2, pp. 65–78, Apr. 2004.

[40] M. Muyan, P. Yi, and G. Sathya, “Fusion estrogen receptor proteins: toward the development of receptor-based agonists and antagonists,” *Mol. Cell. Endocrinol.*, vol. 182, pp. 249–263, 2001.

[41] P. Yi, S. Bhagat, R. Hilf, R. A. Bambara, and M. Muyan, “Differences in the abilities of estrogen receptors to integrate activation functions are critical for subtype-specific transcriptional responses,” *Mol. Endocrinol.*, vol. 16, no. 8,

pp. 1810–27, Aug. 2002.

[42] 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, Apr. 2009.

[43] V. B. Bajic *et al.*, “Dragon ERE Finder version 2: a tool for accurate detection and analysis of estrogen response elements in vertebrate genomes,” *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3605–3607, Jul. 2003.

[44] M. Haring, S. Offermann, T. Danker, I. Horst, C. Peterhansel, and M. Stam, “Chromatin immunoprecipitation: optimization, quantitative analysis and data normalization,” *Plant Methods*, vol. 3, no. 11, 2007.

[45] Y. Zhao, L. Du, and N. Zhang, “Sensitivity of Prestaining RNA with Ethidium Bromide Before Electrophoresis and Performance of Subsequent Northern Blots Using Heterologous DNA Probes,” pp. 204–210, 2013.

[46] D. Szklarczyk *et al.*, “The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D362–D368, Jan. 2017.

[47] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources,” *Nat. Protoc.*, vol. 4, no. 1, pp. 44–57, Jan. 2009.

[48] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists,” *Nucleic Acids Res.*, vol. 37, no. 1, pp. 1–13, Jan. 2009.

[49] R. Zheng *et al.*, “Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D729–D735, Jan. 2019.

[50] O. Fornes *et al.*, “JASPAR 2020: update of the open-access database of transcription factor binding profiles,” *Nucleic Acids Res.*, Nov. 2019.

[51] P. Yi *et al.*, “The effects of estrogen-responsive element- and ligand-induced structural changes on the recruitment of cofactors and transcriptional responses by ERα and ERβ,” *Mol. Endocrinol.*, vol. 16, no. 4, pp. 674–93, Apr. 2002.

[52] M. Berry, A. M. Nunez, and P. Chambon, “Estrogen-responsive element of the human pS2 gene is an imperfectly palindromic sequence.,” *Proc. Natl. Acad. Sci.*, vol. 86, no. 4, pp. 1218–1222, Feb. 1989.

[53] R. Métivier *et al.*, “Estrogen Receptor-α Directs Ordered, Cyclical, and Combinatorial Recruitment of Cofactors on a Natural Target Promoter,” *Cell*, vol. 115, no. December, pp. 751–763, 2003.

[54] R. Sorek, “A Non-EST-Based Method for Exon-Skipping Prediction,” *Genome Res.*, vol. 14, no. 8, pp. 1617–1623, Aug. 2004.

[55] K. D. Pruitt, T. Tatusova, and D. R. Maglott, “NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins,” *Nucleic Acids Res.*, vol. 35, no. Database, pp. D61–D65, Jan. 2007.

[56] A. R. Kornblihtt, I. E. Schor, M. Alló, G. Dujardin, E. Petrillo, and M. J. Muñoz, “Alternative splicing: a pivotal step between eukaryotic transcription and translation,” *Nat. Rev. Mol. Cell Biol.*, vol. 14, no. 3, pp. 153–165, Mar. 2013.

[57] 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.

[58] J.-R. Landry, D. L. Mager, and B. T. Wilhelm, “Complex controls: the role of alternative promoters in mammalian genomes,” *Trends Genet.*, vol. 19, no. 11, pp. 640–648, Nov. 2003.

[59] J. L. Ashurst and J. E. Collins, “G <scp>ENE</scp> A

<scp>NNOTATION</scp> : P <scp>REDICTION AND</scp> T <scp>ESTING</scp>," *Annu. Rev. Genomics Hum. Genet.*, vol. 4, no. 1, pp. 69–88, Sep. 2003.

[60] N. D. Trinklein, "Identification and Functional Analysis of Human Transcriptional Promoters," *Genome Res.*, vol. 13, no. 2, pp. 308–312, Feb. 2003.

[61] M. A. Frohman, M. K. Dush, and G. R. Martin, "Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer.," *Proc. Natl. Acad. Sci.*, vol. 85, no. 23, pp. 8998–9002, Dec. 1988.

[62] C. A. Raabe, T.-H. Tang, J. Brosius, and T. S. Rozhdestvensky, "Biases in small RNA deep sequencing data," *Nucleic Acids Res.*, vol. 42, no. 3, pp. 1414–1426, Feb. 2014.

[63] O. Yeku and M. A. Frohman, "Rapid Amplification of cDNA Ends (RACE)," in *Methods in molecular biology (Clifton, N.J.)*, 2011, pp. 107–122.

[64] G. F. Sewack, T. W. Ellis, and U. Hansen, "Binding of TATA Binding Protein to a Naturally Positioned Nucleosome Is Facilitated by Histone Acetylation," vol. 21, no. 4, pp. 1404–1415, 2001.

[65] S. R. Hartono, I. F. Korf, and F. Chédin, "GC skew is a conserved property of unmethylated CpG island promoters across vertebrates," *Nucleic Acids Res.*, vol. 43, no. 20, p. gkv811, Aug. 2015.

[66] W. S. Yong, F. M. Hsu, and P. Y. Chen, "Profiling genome - wide DNA methylation," *Epigenetics Chromatin*, pp. 1–16, 2016.

[67] Z. D. Smith and A. Meissner, "DNA methylation: roles in mammalian development," *Nat. Rev. Genet.*, vol. 14, no. 3, pp. 204–220, Mar. 2013.

[68] 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.*, vol. 103, no. 5, pp. 1412–1417, Jan. 2006.

[69] A. M. Deaton and A. Bird, "CpG islands and the regulation of transcription," *Genes Dev.*, vol. 25, no. 10, pp. 1010–1022, May 2011.

[70] A. Barski, S. Cuddapah, K. Cui, T. Roh, D. E. Schones, and Z. Wang, "Resource High-Resolution Profiling of Histone Methylations in the Human Genome," pp. 823–837, 2007.

[71] B. Bartkowiak and A. L. Greenleaf, "Phosphorylation of RNAPII," no. June, pp. 115–119, 2011.

[72] C. Zhang *et al.*, "Definition of a FoxA1 Cistrome That Is Crucial for G 1 to S-Phase Cell-Cycle Transit in Castration-Resistant Prostate Cancer," vol. 71, no. 10, 2011.

[73] T. Vavouri and B. Lehner, "Human genes with CpG island promoters have a distinct transcription-associated chromatin organization," *Genome Biol.*, vol. 13, no. 11, p. R110, 2012.

APPENDICES

A. Primer Sequences

Primer Name	Primer Sequence (5' to 3')
CXXC5 Expression FP	CGGTGGACAAAAGCAACCCTAC
CXXC5 Expression REP	CGCTTCAGCATCTCTGTGGACT
TFF1 Expression FP	TTGTGGTTTTCCTGGTGTCA
TFF1 Expression REP	CCGAGCTCTGGGACTAATCA
RPLP0 Expression FP	GGAGAAACTGCTGCCTCATA
RPLP0 Expression REP	GGAAAAAGGAGGTCTTCTCG
CXXC5-ERE FP	AATGCCTGGTCAAGCACATG
CXXC5-ERE REP	TCTTCACTCTGTCACAAGAGGA
TFF1-ERE FP	CCTGTGGCCCAGCCACTGCGTCTTTCAG
TFF1-ERE REP	CCTATCTCCTTGGGAGAGCTGTGAG
CXXC5_Promoter FP	AGGGAGGGGAGCCGAGGTGGA
CXXC5_Promoter REP	CGGGCAGGAGCTATGAAC
OAS1_Promoter FP	AATTCAGCACTGGGATCAGG
OAS1_Promoter REP	TTGGCTGGGGTATTTCTGAG
C-MYC_Promoter FP	CTTTATAATGCGAGGGTCTGG
C-MYC_Promoter REP	TGCCTCTCGCTGGAATTACT
CCNA2_Promoter FP	ATCCCGCGACTATTGAAATG
CCNA2_Promoter REP	CGCTCACTAGGTGGCTCAG
MB_Exon2 FP	AAGTTTGACAAGTTCAAGCACCTG
MB_Exon2 REP	TGGCACCATGCTTCTTTAAGTC
GAPDH Promoter FP	TACTAGCGGTTTTACGGGCG
GAPDH Promoter REP	CGCATAAGCTTCCAGCGTCAAAGGTGG
pGL3_Probe FP	5'Biosg/CTAGCAAAATAGGCTGTCCCC
pGL3_Probe REP	5'Biosg/TGGCTTTACCAACAGTACCGGAAT

B. Oligomer List for EMSA

Consensus ERE_Up	5'Biosg/GCGAGATATGGTCACGCTGACCGAGATTCCTTA
Consensus ERE_Down	5'Biosg/ATTCCTTAGAGGGTCAGCGTGACCTATAGAGCG
Cold Consensus ERE_Up	GCGAGATATGGTCACGCTGACCGAGATTCCTTA
Cold Consensus ERE Down	ATTCCTTAGAGGGTCAGCGTGACCTATAGAGCG
CXXC5 ERE_Up	5'Biosg/GCGAGATATGGTCAGGATGACAGAGATTCCTTA
CXXC5 ERE_Down	5'Biosg/ATTCCTTAGAGTGTCATCCTGACCTATAGAGCG
Cold CXXC5 ERE_Up	GCGAGATATGGTCAGGATGACAGAGATTCCTTA
Cold CXXC5 ERE_Down	ATTCCTTAGAGTGTCATCCTGACCTATAGAGCG
ELF1_Up	5'Biosg/CTGAGCGGGGATGTAGAGGCGG
ELF1_Down	CCGCCTCTACATCCCCGCTCAG
Cold ELF1_Up	CTGAGCGGGGATGTAGAGGCGG
ΔELF1_Up	5'Biosg/CTGAGCTACCTTCTAGAGGCGG
ΔELF1_Down	CCGCCTCTAGAAGGTAGCTCAG
MAZ_Up	5'Biosg/GGGAGGAGGGAGGGGAGCCGAG
MAZ_Down	CTCGGCTCCCCTCCCTCCTCCC
Cold MAZ_Up	GGGAGGAGGGAGGGGAGCCGAG
ΔMAZ_Up	5'Biosg/GGGATTTTGGAAGTTTTCCGAG
ΔMAZ_Down	CTCGGAAAACTTCCAAAATCCC

C. List of the Proteins Identified by LC/MS-MS Analysis

Accession	Gene Name	ΣCoverage	Σ# Peptides	Σ# PSMs	calc. pI
Q9ULW3	ABT1	9,56	2	3	9,876465
		7,72	2	3	9,876465
P51825	AFF1	14,13	13	35	9,202637
		16,86	20	26	9,202637
Q6UB98	ANKRD12	19,64	32	54	7,005371
		18,82	30	57	7,005371
P17544	ATF7	10,12	4	14	8,645996
		17,21	5	11	8,645996
O15265	ATXN7	14,13	10	21	9,847168
		15,92	10	17	9,847168
Q6W2J9	BCOR	16,01	22	42	6,480957
		18,01	27	52	6,480957
Q5T5X7	BEND3	37,56	27	58	5,427246
		27,9	22	51	5,427246
P54132	BLM	22,72	28	70	7,48877
		16,94	20	43	7,48877
P25440	BRD2	20,85	15	21	9,085449
		20,72	11	14	9,085449
Q15059	BRD3	19,15	14	27	9,36377
		24,93	16	33	9,36377
P41223	BUD31	31,25	6	11	8,821777
		21,53	3	4	8,821777
Q9HC52	CBX8	13,11	5	7	9,905762
		7,71	2	3	9,905762
Q9UFW8	CGGBP1	11,98	2	3	8,953613
		11,98	2	5	8,953613
A5YKK6	CNOT1	10,52	19	33	7,10791
		10,65	18	27	7,10791
P16220	CREB1	9,09	3	4	5,566895
		5,87	2	3	5,566895
P35221	CTNNA1	27,81	17	27	6,290527
		24,72	16	29	6,290527
Q9UJV9	DDX41	10,77	5	8	6,844238
		27,49	12	20	6,844238
Q9Y6V7	DDX49	13,66	8	17	9,056152
		22,36	12	20	9,056152
Q9BQ39	DDX50	33,51	19	36	9,17334

		21,3	15	24	9,17334
Q8TDD1	DDX54	21,68	15	26	10,02295
		27,92	20	37	10,02295
Q9UNQ2	DIMT1	26,2	10	17	9,993652
		30,03	9	14	9,993652
Q9H147	DNTTIP1	3,34	1	1	8,968262
		19,15	4	5	8,968262
Q99848	EBNA1BP2	21,9	5	7	10,09619
		22,55	5	9	10,09619
P32519	ELF1	14,05	6	14	5,211426
		16,8	6	10	5,211426
P18074	ERCC2	12,11	8	12	7,151855
		15,13	8	12	7,151855
P39748	FEN1	18,68	8	18	8,616699
		27,89	9	12	8,616699
Q8WUU5	GATAD1	17,84	4	5	9,407715
		13,38	4	6	9,407715
P35269	GTF2F1	15,67	8	18	7,48877
		28,24	15	18	7,48877
Q8WUA4	GTF3C2	11,2	8	13	7,312988
		9	6	8	7,312988
Q92522	H1FX	50,7	12	19	10,75537
		33,33	8	25	10,75537
P54198	HIRA	10,52	8	16	8,074707
		10,91	8	12	8,074707
P33778	HIST1H2BB	61,9	7	19	10,31592
		46,03	6	21	10,31592
Q9UGU5	HMGXB4	24,29	14	20	9,319824
		36,44	22	43	9,319824
Q12906	ILF3	12,53	8	13	8,763184
		9,17	6	11	8,763184
Q9ULG1	INO80	18,06	25	47	9,495605
		13,37	20	36	9,495605
Q96HW7	INTS4	11,21	8	21	6,442871
		7,58	5	14	6,442871
Q6P9B9	INTS5	6,58	5	11	7,049316
		6,48	5	12	7,049316
Q15652	JMJD1C	18,27	37	96	7,869629
		19,8	36	99	7,869629
Q92830	KAT2A	17,92	14	25	9,041504
		8,6	6	11	9,041504
O60341	KDM1A	25,7	20	32	6,521973

		19,13	15	24	6,521973
Q9Y2K7	KDM2A	12,74	12	14	7,57666
		12,65	11	16	7,57666
Q8NHM5	KDM2B	17,51	18	35	8,558105
		20,28	20	44	8,558105
Q9Y4F3	MARF1	13,66	19	37	7,869629
		12,06	17	32	7,869629
P56270	MAZ	4,19	2	3	8,953613
		11,74	4	6	8,953613
Q9UBB5	MBD2	20,92	8	18	10,0376
		26,28	8	22	10,0376
Q96EZ8	MCRS1	19,7	8	19	9,378418
		19,26	9	14	9,378418
O75448	MED24	12,44	10	13	6,946777
		22,45	15	22	6,946777
O00255	MEN1	10,89	5	11	6,55127
		15,12	8	16	6,55127
Q8IWI9	MGA	19,86	47	77	6,785645
		18,34	51	91	6,785645
Q9Y6X9	MORC2	21,9	18	24	8,382324
		20,25	16	29	8,382324
Q15014	MORF4L2	39,58	11	17	9,715332
		45,83	10	24	9,715332
P20585	MSH3	17,06	20	35	8,016113
		17,15	17	28	8,016113
Q13330	MTA1	29,37	16	26	9,26123
		17,2	11	19	9,26123
Q9NPC7	MYNN	12,62	4	7	8,250488
		21,31	7	10	8,250488
Q12857	NFIA	19,06	9	18	8,440918
		14,34	7	14	8,440918
O00712	NFIB	15,24	5	13	8,865723
		23,33	7	24	8,865723
Q6P4R8	NFRKB	21,48	21	46	9,246582
		20,79	23	54	9,246582
Q9H0G5	NSRP1	14,7	6	9	8,836426
		11,65	6	7	8,836426
Q14980	NUMA1	18,91	32	44	5,782715
		14,99	27	40	5,782715
O43913	ORC5	17,93	6	11	7,737793
		13,56	5	12	7,737793
Q9UGN5	PARP2	15,95	6	16	8,880371

		25,04	9	21	8,880371
Q5JVF3	PCID2	26,57	9	15	8,528809
		16,54	6	8	8,528809
Q9NQV6	PRDM10	5,93	5	9	6,741699
		8,2	7	12	6,741699
P49792	RANBP2	16,66	39	73	6,20166
		14,83	36	68	6,20166
P06400	RB1	16,81	13	21	7,942871
		25,65	20	35	7,942871
Q7Z6E9	RBBP6	21,04	36	94	9,64209
		24,27	38	83	9,64209
P35249	RFC4	25,34	9	15	8,016113
		28,1	9	16	8,016113
Q5UIP0	RIF1	16,95	36	51	5,516113
		14,85	31	46	5,516113
P15927	RPA2	13,33	2	5	6,150879
		5,19	1	1	6,150879
Q5VWQ0	RSBN1	10,1	6	18	8,602051
		20,2	12	23	8,602051
Q7Z333	SETX	17,48	41	89	7,166504
		13,07	28	59	7,166504
Q14683	SMC1A	20,28	20	31	7,635254
		14,52	14	21	7,635254
Q68D10	SPTY2D1	31,82	15	29	9,788574
		24,53	10	20	9,788574
Q8NEM7	SUPT20H	10,91	8	17	8,484863
		20,03	15	24	8,484863
Q86TJ2	TADA2B	16,67	8	18	7,825684
		17,14	9	17	7,825684
O75529	TAF5L	6,28	3	5	5,884277
		12,05	5	9	5,884277
P49848	TAF6	21,27	8	20	8,602051
		20,53	12	23	8,602051
Q9Y6J9	TAF6L	17,36	6	9	8,968262
		21,7	8	13	8,968262
Q15554	TERF2	16,42	9	15	9,349121
		12,18	7	9	9,349121
Q92754	TFAP2C	28	8	27	7,825684
		23,11	10	32	7,825684
Q01664	TFAP4	21,3	4	6	5,871582
		13,61	4	5	5,871582
Q13472	TOP3A	8,89	8	10	8,338379

		12,39	11	15	8,338379
Q9Y4A5	TRRAP	16,66	55	88	8,191895
		14,15	40	63	8,191895
Q9UNY4	TTF2	22,12	21	40	8,367676
		22,46	23	36	8,367676
Q9NPG3	UBN1	24,34	22	35	9,334473
		25,4	21	35	9,334473
Q9HAU5	UPF2	13,13	18	22	5,693848
		17,3	18	22	5,693848
Q9BZI7	UPF3B	17,6	8	27	9,480957
		17,18	8	33	9,480957
O95785	WIZ	16,54	16	28	6,858887
		14,78	18	28	6,858887
P67809	YBX1	12,96	2	3	9,876465
		18,52	4	5	9,876465
Q8N680	ZBTB2	15,56	6	9	6,100098
		9,53	3	6	6,100098
O95365	ZBTB7A	13,87	5	19	5,186035
		17,12	7	22	5,186035
O15156	ZBTB7B	8,16	2	4	5,858887
		4,82	1	4	5,858887
O00488	ZNF593	20,15	2	4	9,817871
		6,72	1	2	9,817871
Q96I27	ZNF625	22,88	3	57	8,98291
		20,59	2	43	8,98291

D. Layered Cistorme Analyses for *CXXC5* Promoter

	Ser5P-RNAP II	RNAP II	MAZ	CREB1	ELF1
Red	iPSC	K562	K562	A549	HepG2
Green	Neuron	U87	IMR90	GM12878	HCT-116
Blue	Stomach	MCF7	HEK293	HepG2	GM12878
Yellow	Colon	MCF7	MCF7	LNCaP	K562
iPSC	Induced Pluriotent Stem Cell				
K562	Chronic Myelogenous Leukemia				
U87	Glioblastoma				
MCF7	Breast Adenocarcinoma				
IMR90	Human Lung Fibroblast				
HEK293	Human Embryonic Kidney				
A549	Lung Adenocarcinoma				
GM12878	Chronic Myelogenous Leukemia				
HepG2	Hepatocellular Carcinoma				
LNCaP	Prostate Adenocarcinoma				
HCT-116	Colorectal Carcinoma				

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Yaşar, Pelin
Nationality: Turkish (TC)
Date and Place of Birth: 30 January 1990, İzmir
Marital Status: Single
Phone: +90 312 210 7664
email: pelin.yasar@metu.edu.tr

EDUCATION

Degree	Institution	Year of Graduation
MS	METU Biological Sciences	2015
BS	ITU Molecular Biology and Genetics	2012
High School	Meram Anadolu Lisesi, Konya	2007

WORK EXPERIENCE

Year	Place	Enrollment
2009 July-September	İstanbul Üniversitesi Tıbbi Genetik	Intern Student

FOREIGN LANGUAGES

Advanced English

PUBLICATIONS

1. **Yaşar, P**., Kars, G., Yavuz, K., Ayaz, G., Demiralay, D.Ö., Oğuztüzün, Ç., Bigen, E., Suvacı, Z., Çetinkol, Ö.P., Can, T., and Muyan, M. (2020). A CpG island promoter drives the expression of the CXXC5 gene (to be submitted, the co-corresponding author).

2. Ayaz, G., Turan, G., Olgun, Ç.E., Kars, G., Karakaya, B., Can, T., Muyan, M., and **Yaşar, P**. (2020). A prelude to the proximity interaction mapping implicates CXXC5 as a molecular scaffold for gene expressions (to be submitted, first co-author).

3. Ayaz, G., Razizadeh, N., **Yaşar, P**., Kars, G., Kahraman, D.C., Saatci, Ö., Şahin, Ö., Çetin-Atalay, R., and Muyan, M. (2019). CXXC5 as an unmethylated CpG dinucleotide binding protein contributes to estrogen-mediated cellular proliferation. Scientific Reports 10, 9943 (2020). DOI:10.1038/s41598-020-66682-7

4. Ayaz, G., **Yasar, P**., Olgun, C.E., Karakaya, B., Kars, G., Razizadeh, N., Yavuz, K., Turan, G. & Muyan, M. (2019, January). Dynamic transcriptional events mediated by estrogen receptor alpha. Frontiers in Bioscience. 24, 245-276, DOI: 10.2741/4716

5. **Yasar, P**., Ayaz, G., User, S.D., Güpür, G. & Muyan, M. (2017). Molecular mechanism of estrogenestrogen receptor signaling. Reproductive Medicine and Biology. 5;16(1):4-20. DOI: 10.1002/rmb2.12006

6. **Yasar, P**., Ayaz, G. & Muyan, M. (2016, November). Estradiol-Estrogen Receptor α Mediates the Expression of the CXXC5 Gene through the Estrogen Response Element-Dependent Signaling Pathway. Scientific Reports, 25;6:37808. DOI: 10.1038/srep37808

7. **Yasar, P**., Ayaz, G., Olgun, Ç.E., Karakaya, B., Kars, G. & Muyan, M. (2016). CXXC5 is an estrogen responsive gene. FEBS J, 283: ST-05.02.2-008.

8. Muyan, M., Gupur, G., **Yasar, P**., Ayaz, G., User, SD., Kazan, HH., & Huang, Y. (2015). Modulation of Estrogen Response Element-Driven Gene Expressions and Cellular Proliferation with Polar Directions by Designer Transcription Regulators. PLOS ONE. 21;10(8): e0136423. DOI: 10.1371/journal.pone.0136423

9. **Yasar, P**. & Muyan, M. (2014, April). CXXC5 (CXXC finger protein 5). Atlas of Genetics and Cytogenetics in Oncology and Haematology. 19(1):1-3, DOI: 04-2014-CXXC5ID52549ch5q31.pdf

Dancing, Reading, Camping, Movies