SHORT THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

Examination of the formation of estrogen receptor alpha-driven super-enhancers

by Dóra Bojcsuk

Supervisor: Dr. Bálint László Bálint

UNIVERSITY OF DEBRECEN
DOCTORAL SCHOOL OF MOLECULAR CELL AND IMMUNE BIOLOGY

DEBRECEN, 2019

Examination of the formation of estrogen receptor alpha-driven super-enhancers

by **Dóra Bojcsuk**, MSc

Supervisor: Dr. Bálint László Bálint, PhD

Doctoral School of Molecular Cell and Immune Biology, University of Debrecen

Examination Committee:

Head:	Prof. Dr. László Fésüs, PhD, DSc, MHAS
Members:	Prof. Dr. Péter Viktor Nagy, PhD, DSc
	Dr. Csaba Barta, PhD

The Examination takes place at the 3.009-010 Library of the Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen at 11 a.m., 13th of December, 2019.

Reviewers of the thesis: Prof. Dr. Imre Miklós Boros, PhD, DSc
Prof. Dr. Gábor Szabó, PhD, DSc

Defense Committee:

Head:	Prof. Dr. László Fésüs, PhD, DSc, MHAS
Members:	Prof. Dr. Imre Miklós Boros, PhD, DSc
	Prof. Dr. Gábor Szabó, PhD, DSc
	Prof. Dr. Péter Viktor Nagy, PhD, DSc
	Dr. Csaba Barta, PhD

The PhD Defense takes place at the Lecture Hall of Building A of the Department of Internal Medicine, Faculty of Medicine, University of Debrecen, at 1 p.m., 13th of December, 2019.

1. Introduction

1.1. Organization of the human genome

The human genome can be described with a well-defined sequence of approximately 3.2 billion ($3.2x10^9$) base pairs and the first version of it was released in 2001 by the Human Genome Project. The cell nucleus contains two copies of 3.2 megabase pair long DNA molecules, which are altogether about 2 meters long. In order to store this massive information in the nucleus of 6 μm diameter, a well-organized and higher-order structure of chromatin is required. For this purpose, different DNA segments form a bead-like structure, which are called nucleosome; later, these nucleosomes stack together to form 30-nm fibers which have a coil-like solenoid-type or zig-zag structure. However, the highest level of organization is achieved by the chromosome structure. The 30-nm fiber first forms large loops held together by Condensin II rings at the stem. These stems form essentially a core rich in Condensin II, from which the loops of different sizes originate radially. These are eventually further subdivided by Condensin I rings to form a solid chromosome. It is imperative to mention that chromatin packaging into chromosomes is only observed during cell division; most DNA is found in an easily accessible form that allows genes to be regulated by DNA-binding proteins.

1.2. Transcriptional regulation levels in eukaryotes

The Human Genome Project provided an insight into our protein-coding gene pool and the location of approximately 20-25,000 genes within the genome. Except for the so-called housekeeping genes, which are responsible for the basic cellular processes, up to half of our ~25,000 genes show some activity in every cell. Moreover, different cells are regulated by different gene sets. This cell and/or tissue-specific gene expression pattern and timely genes'

on/off switches determine the molecular mechanisms of tissue development, cell fate, metabolic status, and the eventual the transcriptional regulation of the diverse biological processes.

Gene regulation is a fundamental biological process for an appropriate cell function, and tightly controlled by perfectly sophisticated regulatory mechanisms. The first gene regulatory model was described in 1960 by *François Jacob* and *Jacques Monod*, who proposed that the resulting gene products (RNA, protein) are able to regulate genes expression through feedback control. This idea formed the basis of the contemporary gene expression models.

According to our present knowledge, DNA-binding gene regulatory proteins, namely the transcription factors, act as *trans*-regulators and exert their stimulatory or inhibitory effects via interacting with *cis*-regulatory elements (CREs), such as promoters, enhancers, silencers, and insulators, or even the entire of *cis*-regulatory modules (CRMs), which may include even more than ten CREs. During eukaryotic development, the well-defined set of transcription factors is responsible for the proper gene function. The composition of the transcription factors could be combinatorial at different stages of development thus allowing the cell type- and context-dependent regulation of the genes.

Transcription factors are able to bind with a thousand fold or even higher affinity to their recognition motifs compared to a random sequence motif and this primary selection allows them to exert their effects on the target genes. If chromatin is accessible for a transcription factor and this transcription factor occupies its own motif, co-factors (e.g. subunits of the Mediator complex or P300 protein with acetyltransferase activity) or post-translational modifications are able to further improve or catalyze its regulatory effects on the genes. This protein complex is already capable of anchoring RNA polymerase II to the *core* promoter and activating it.

In addition to the regulatory elements of the DNA, epigenetic regulation is also an important level of eukaryotic transcriptional processes. It includes modulating target chromatin accessibility and regulating the degree of DNA methylation.

1.3. ChIP-seq: genome-wide identification of transcriptional factor binding sites

Chromatin immunoprecipitation (ChIP) coupled with next-generation sequencing is a widely used technique for studying chromatin-associated proteins e.g. transcription factors or histone-modifying enzymes at a genome-wide level. This method typically requires 1 to 10 million cells and the proteins can be detectable by using antibody against the protein of interest.

1.4. Super-enhancers

Super-enhancers are large regulatory regions of the genome where the enhancers form clusters through the interaction of several enhancers and a large number of proteins, including transcription factors and co-regulators. Within the super-enhancers, presence of critical components of active enhancers, such as P300, BRD4, and MED1 proteins, histone marks H3K27ac and H3K4me1/2 can be observed. Compared to the typical enhancers, their importance lies in the fact that they are involved in the regulation of genes which largely determine the cell fates.

1.5. Identification of super-enhancers

The two most widely-used bioinformatic programs for identifying super-enhancers are Rank Ordering of Super-Enhancers (ROSE) and Hypergeometric Optimization of Motif EnRichment (HOMER). Super-enhancers can be identified based on the presence of master transcription factor binding sites or co-factor (MED1, BRD4) bindings predicted from ChIP-seq experiment.

However, the most commonly used prediction method is based on the H3K27ac-labeled nucleosome regions.

Basically, both ROSE and HOMER work with the same logic: in the first step of the prediction, the individual enhancers are clustered together using one of the aforementioned programs and the only condition is that the distance between two enhancers should not exceed 12.5 kb. In the second step, the enhancer clusters are ranked according to their read coverage (density) and below the point where the slope does not exceed 1, enhancers are considered as typical enhancers and enhancer clusters above that point are considered super-enhancers. In total, only a few hundred super-enhancers can be identified from each experiment and the resulting super-enhancer numbers might differ with the use of the two methods due to the different mathematical approaches.

1.6. Nuclear receptors

The currently known 48 members of the nuclear receptor superfamily are involved not only in the regulation of many physiological processes in humans, such as metabolism, cell proliferation, inflammation or even circadian rhythm, but also play an important role in pathological processes, like most cancers. Their central role is also supported by the fact that approximately 13% of pharmaceutical drugs are targeted by nuclear receptors.

Most nuclear receptors represent a group of transcription factors that are able to directly bind to various small, fat-soluble molecules, endogenous ligands (such as steroid hormones) via their ligand-binding domain, which can easily penetrate the lipid bilayer due to its lipophilic properties. The formed hormone-receptor complex – together with a number of co-regulator (activator or repressor) proteins can bind directly to a specific DNA sequence and thereby directly regulate the expression of target genes. However, some nuclear receptors do not have

ligand-binding pockets; they are probably regulated by other mechanisms, such as post-transcriptional modifications.

1.7. Estrogen receptor alpha

In the late 1920's, Edward A. Doisy, Clement D. Veler and Sidney A. Thayer were the firsts who isolated the estrone (E1) and the estriol (E3); later Doisy successfully isolated the estradiol (E2), as well. These estrogen derivatives differ in the number of their hydroxyl groups and in their affinity for the receptor. Although E1 and E3 are both high-affinity ligands, they are much weaker agonists of the estrogen receptors (ERs) than E2. The two estrogen receptors, alpha (ERα) and beta (ERβ) are responsible for the regulation of the physiological functions of E2.

ERα is one of the most extensively studied member of the nuclear receptor superfamily. It can be found in the mammary gland, uterus, ovarian theca cells, bone, liver, adipose tissue, and also in the male reproductive system and in the prostate stromal cells. However, its prominent role is due to the fact that ERα is a key hormone-regulated transcription factor in the three-quarters of breast cancer cases.

The most widely used human cell line to study the behavior of ERα is MCF-7. MCF-7 is a well-established *in vitro* model for the investigation of estrogen-dependent biological processes of breast cancer development, as this model is an ER^+ breast cancer-derived cell line isolated from the pleural effusion of a patient with metastatic breast cancer.

Although ERα is located predominantly in the nucleus even in the absence of the ligand, it continuously moves between the nucleus and the cytoplasm. Upon estradiol activation (which has a $K_d=10^{-10}$ M affinity for the ligand-binding pocket of the receptor), first, the receptor dissociates from the heat shock protein 90 (hsp90), which inhibits its effect. Then, due to the conformation change of the ligand-binding domain, the receptor is translocated from the cytoplasm to the

nucleus where it can directly binds to a specific DNA sequence, called estrogen response element (ERE). Using general bioinformatic motif scans, more than 1 million putative ERα transcription factor binding sites were identified in the reference human genome but only a portion of these sites is functionally relevant. The discrepancy between the putative ERα binding sites and those identified in experimental models may reflect the accessibility of the sites. Chromatin accessibility has been suggested to drive, in general, the site selection of transcription factors but can also be explained by the sequence differences within the EREs or by the different cell types that regulate the required genes for their function through different sets of enhancers. In addition to that, the presence of various co-factors are also required to facilitate the availability of chromatin.

1.8. Co-factors in the estrogen receptor alpha-driven transcriptional regulation

ERα is controlled by various co-factors or co-factor complexes that can act as co-activator or co-repressor proteins that may interact with a number of DNA-binding transcription factors to control the proper functioning of the genes. The most common transcription factors that can interact with ERα are, fore example, Forkhead box 1 (FoxA1), Activator Protein 2 gamma (AP2γ), GATA-binding protein 3 (GATA3), Retinoic acid receptor alpha/gamma (RARα/γ), Activator protein 1 (AP-1), or Signal transducer and activator of transcription 1 (STAT1). FoxA1 plays a role as a pioneer factor, and it has been proposed that FoxA1 binds to ~50% of the regions occupied by ERα, even in the absence of estradiol. Moreover, FoxA1 is indispensable for any ERα recruitment.

AP2γ and GATA3 are also recruited to ERα binding sites and may be involved in stabilizing ERα chromatin interaction. The motif of these proteins also shows a significant

enrichment in ERα-bound chromatin regions. Based on ChIA-PET data, 88% of ERα binding sites that are interacting with distant DNA regions, show co-occurrence with FoxA1 and GATA3 proteins. This suggests that these transcription factors are involved in the fine-tuning of the transcriptional responses to E2 treatment.

2. Aims of the study

2.1. Examination of the functioning of ligand-inducible super-enhancers

During (re)analyzing several publicly available ERα transcription factor ChIP-seq data we observed that already in the absence of added ligand there is ERα detectable typically at one or few high-affinity response element(s) that will serve as the nucleating point of the ERα-driven super-enhancers activated by the specific estradiol tretment.

Hypothesis: the primary enhancers form the basis of super-enhancers activating upon ligand treatment.

To prove these observations, we globally determined the differences between:

1. the effect of ligand treatment on the binding of the primary (we referred to them as 'mother enhancers') and subsequently appearing secondary ('daughter enhancers') enhancers;
2. dynamics of their binding;
3. the sequence motifs occupied by them;
4. between their co-factors and co-regulators, and
5. whether these phenomena can be observed in the case of other ligand-inducible transcription factors.

Because MCF-7 cell line is a very important model, many research groups used it for the genome-wide examination of ERα transcription factor and its co-factors. Thus, the raw sequenced data generated by others can be available in publicly available databases. Therefore, we have chosen ERα as our primary model to present our results.

2.2. Characterization of the ERα *cistrome* depending on the presence of ERE and the inducibility of the bindings

We categorized and characterized more than 88,000 ERα binding events depending on whether they were present prior or after E2-treatment and further divided them into additional clusters based on the presence or absence of ERE. During these steps, we identified further 6,535 ERα binding sites that are not part of super-enhancers but can be characterized by similar features to those of mother enhancers.

Our goal was to determine whether:

1. there is difference between the effects of mother enhancers on gene expression and the effecs of those 6,535 ERα binding sites that are not part of super-enhancers but can be characterized by similar features;
2. there is difference betwen the preferred DNA sequence motifs, co-factors, and the gene expressional effects of the four different clusters of ERα binding sites.

3. Materials and methods

3.1. Data collection

For the characterization of ligand-inducible super-enhancers, we investigated seven transcription factors in five different human and mouse cell types: ERα, FoxA1 and AP2γ in the MCF-7 cell line, androgen receptor (AR) in the prostate cancer-derived LNCaP cell line, JUNB in primary bone marrow-derived macrophages (BMDM) of the C57BL/6 mouse strain, VDR in mouse intestinal epithelial cells and RAR in the F9 mouse embryonic testis carcinoma cell line. Control and ligand-treated ChIP-seq samples for the above-mentioned transcription factors were selected from the publicly available Sequence Read Archive (SRA) or Gene Expression Omnibus (GEO) databases.

Additionaly, samples of ERα ChIP-seq experiments were included in our analysis ranging from vehicle-treated samples, untreated samples, vehicle- and E2-treated *FoxA1 knock-down* samples, tamoxifen- and fulvestrant-treated samples as well as a time-course experiment of E2-treatment.

In order to characterize different types of ERα binding sites, additional ChIP-, DNase-, and E2-treated time-course RNA-seq data carried out in the MCF-7 cell line were downloaded from the SRA and GEO databases. The analyzed factors are the following: vehicle- and E2-treated FoxA1, AP2γ, GATA3, ERα, siCTL ERα, siFoxA1, H3K27ac, P300, DNase I, HDAC2, and SIN3.

3.2. ChIP-seq data analysis

Raw sequence data were downloaded from the GEO database and processed according to the following steps: ChIP-seq reads were aligned to the human hg19 or mouse mm10 reference genome assembly with the BWA tool (v07.10)followed by generation of BAM files with SAMtools (v0.1.19). Peaks were predicted with the MACS2 tool (v2.0.10) then the artifacts were removed according to the ENCODE blacklisted genomic regions. Tag directories were generated

with the *makeTagDirectory* program of HOMER (Hypergeometric Optimization of Motif EnRichment) (v4.2). Fragment lengths were set to 150 nucleotides. Bedgraph files were generated from the generated HOMER tag directories by use of *makeUCSCfile* command. DNase-seq data were analyzed with the same way; for the prediction of the H2K27ac-labeled regions, broad regions were targeted.

3.3. Super-enhancer prediction

Super-enhancers were predicted from the ligand-treated samples using HOMER *findPeaks* and *style super*. Reads per kilobase per million mapped reads (RPKM) values for both the control and ligand-treated samples were calculated on the summit ±50 bp region of the peaks determined from the corresponding ligand-treated samples. Peaks with the highest read density (referred to as 'mother'peaks) of each future SE were selected from the control samples. The emerging SE peaks (referred to as 'daughter' peaks) were identified in the ligand-treated samples in subsequent analysis.

3.4. *De novo* motif analysis

Motif enrichment analysis was performed on the summit 100-bp regions of the peaks by the *findMotifsGenome.pl* of HOMER. The targeted motif lengths were 10, 12, 14, and 16 bp and the background sequences were randomly generated by HOMER.

3.5. RNA-seq analysis

Raw sequence data were downloaded from the GEO database and aligned to the hg19 reference genome assembly by the use of TopHat program (v2.0.7). The Fragments Per Kilobase of transcripts per Million mapped reads (FPKM) values were calculated by Cufflinks (v2.0.2) with default parameters. Genes were annotated using PeakAnnotator program to the nearest protein-coding genes.

4. Results

4.1. Examination of the functioning of ligand-inducible super-enhancers

4.1.1. Identification of mother- and daughter enhancers

In the study of gene regulation by ERα, 392 highly covered ERα-driven super-enhancer regions with 4,042 unique binding sites were predicted from the GSM614610 sample of the breast cancer-derived MCF-7 cell line. Focusing on some super-enhancers, we observed that in the absence of stimulation, future super-enhancers are represented by one or a few transcription factor binding event(s). According to the assumption that primary enhancers form the basis of super-enhancers, we referred to these elements as 'mother enhancers' (n=392), and the subsequently appearing secondary enhancers were referred to as 'daughter enhancers' (n= 3,650).

4.1.2. Characterization of the transcriptional activity of super-enhancers

As mentioned in the introduction, super-enhancers are associated with an unusual high amount of co-factors, especially MED1 binding. As MED1 was reported to be the key component of the Mediator complex bridging super-enhancers with transcription start sites, we visualized the presence of this component within the ERα-driven super-enhancers. We found that MED1 is preferentially recruited to ERα-bound sites with a high binding affinity to the buds or initiator(s) of SEs, namely, mother enhancers, upon estradiol treatment.

By examining the presence of further co-factors and chromatin marks, we found that mother enhancers are located in the most accessible chromatin regions with high levels of DNase I signal and show the highest P300, H3K27ac and BRD4 coverage upon induction, suggesting that the top ERα enhancers represent the most active regulatory regions. However, this pattern was not depicted by the daughter enhancers.

4.1.3. Identification of specific sequence motifs

We assumed that different co-factors are responsible for the differences in the activity of mother and daughter enhancers. Therefore, we applied a motif enrichment analysis for these regions. Considering the strong *P*-value (1e-200) and the high enrichment of ERE (62.2%) (compared to the background of 3.37%), despite the small number of target sequences (n = 328), we concluded that the high level of ERα recruitment at mother enhancers, even in the absence of estradiol, reflects strong canonical DNA elements. However, for the emergence of daughter peaks, other transcription factors, such as FoxA1 and AP-1 act in concert with the increased level of ERα after estradiol treatment.

4.1.4. Examination of further ligand-inducible super-enhancers

Primary binding sites, that do not require ligand stimulation to interact with the DNA, are driven by significantly stronger response elements compared to the secondary binding sites. We assumed that they do not required the presence of other transcription factor(s). These observations were supported by data from six additional ligand-inducible transcription factors, namely: FoxA1 and AP2γ in the MCF-7 cell line, androgen receptor (AR) in the prostate cancer-derived LNCaP cell line, JUNB in primary bone marrow-derived macrophages (BMDM) of the C57BL/6 mouse strain, VDR in mouse intestinal epithelial cells and RAR in the F9 mouse embryonic testis carcinoma cell line.

All examined transcription factors, playing roles in super-enhancer formation, showed a phenomenon similar to that of ERα. Their primary regulatory regions possessed canonical

elements specific to dominant transcription factor(s) while the further occupied regions had fewer specific elements together with their collaborative factors.

4.2. Characterization of the ERα *cistrome* depending on the presence of ERE and the inducibility of the bindings

4.2.1. Classification of the binding sites

We investigated whether further ERα binding sites that are not part of super-enhancers but can be characterized by similar features to those of mother enhancers could be identified in the genome and we examined if there is a difference in their effects on gene expression or their co-factors. To answer this question, we categorized and characterized more than 88,000 ERα binding sites depending on whether they were present prior to E2-treatment or only after and further divided into additional clusters based on the presence or absence of ERE.

In addition to super-enhancer (Type I, n=4,042), we have identified a second type of ERα transcription factor binding sites (n=6,535) referred to as Type II. These binding sites are able to bind to the DNA through the ERE without E2-treatment and upon hormonal stimulation, the EREs will be occupied by a large amount of ERα protein. We have also identified a third type of ERα transcription factor binding sites (Type III) with much larger numbers (n=16,533), where despite the presence of the ERE, the receptor binding was established only upon adding E2. ERα binding sites that lack ERE were part of Type IV (n=54,408).

4.2.2. Quantifying the ERα binding density within the four types

ERα binding conditions in Type II and Type I. are similar to each other since the EREs can be found in each case and bind to the protein even in the absence of E2-treatment. However, protein density is a ten times lower compared to the density of mother enhancers. This is probably due to the fact that mother enhancers have more stronger response element and as part of super-enhancers, other super-enhancer constituents can further strengthen their binding affinity**.** Even though E2-stimulation had a robust effect on protein recruitment in all types, *de novo* enhancers (Type III) and those binding sites that were established in the absence of ERE (Type IV) have less enrichment in protein binding compared to the other ones.

4.2.3. Presence of the FoxA1, AP2γ, and GATA3 proteins within the different types of ERα binding sites

The motif preferences of the ~88,000 ERα binding sites partially suggested which protein(s) allow(s) ERα molecules to bind DNA. Therefore, we depicted the presence of the main co-factors such as FoxA1, AP2γ, and GATA3. Moreover, E2-induced ERα bindings upon silencing the FoxA1 (siFoxA1). Despite the fact that E2 treatment has generally no genome-wide effect on FoxA1 binding, ERα dominated mother enhancers recruit FoxA1 upon estradiol treatment.

Effect of the siRNA-mediated FoxA1 silencing also showed a similar effect; namely that ERα binding depends largely on the presence of FoxA1. By increasing the FoxA1 binding, AP2γ had similar behavior, with a prominent binding upon E2-treatment in case of mother enhancers while GATA3 did not show any binding. In the case of Type II binding sites, both FoxA1 and AP2γ appear upon E2-treatment and slight recruitment of GATA3 has also been observed.

Silencing of the FoxA1 did not produce a dramatic impact on Type II ERα bindings compared to the super-enhancer constituents in Type I, which may be due to the compensatory role of AP2γ.

Although Type III binding sites are independent on motif level from any major direct co-factor presence, in the lack of FoxA1, ERα binding completely abolished, suggesting an apparently distant regulation of FoxA1 likely in physical proximity due to the chromosomal loops.

4.2.4. Expression of ERα target genes

Super-enhancers are not only exceptional because of the amount of attracted co-activators but also their effect on genes is determinative. It is known that super-enhancers regulate cell type-specific genes that largely determine the cell identity. Therefore, we calculated the average expression of the genes possibly regulated by the different types of ERα transcription factor binding sites. We determined the closest protein-coding genes to each super-enhancers, then plotted their average gene expression values in 10 different time points (from 0 to 1280 minutes) upon E2-treatment. The time-course RNA-seq data clearly showed that ERα-driven super-enhancers indeed regulate genes in each time point with 2-times higher expression compared to the typical enhancers in all groups. However, there is no difference between the expression patterns of the remaining three types.

4.2.5. Genes regulated by ERα-driven super-enhancers

ERα-driven super-enhancers regulate genes such as BCAS*3*, KRT8, KRT19, AZIN1, SLC9A3R1, CXXC5, SLC7A5, HES1 or CCND1 with pivotal roles in breast cancer. While BCAS3 (Breast Carcinoma Amplified Sequence-3) is known as a co-activator of ERα in breast cancer cell, altered expression of the KRT8 and KRT19 (Keratin 8 and 19), AZIN1 (Antizyme

inhibitor 1), SLC9A3R1 (SLC9A3 Regulator 1), SLC7A5 (Solute Carrier Family 7 Member 5), CXXC5 (CXXC Finger Protein 5) and HES1 (Hes Family BHLH Transcription Factor 1) genes is correlates with poor prognosis or metastatic conditions in breast cancer patients.

As super-enhancers are responsible for cell identity, understanding the mechanism of action of super-enhancers is indispensable for improving knowledge of the regulation of gene expression in general and cellular identity in particular. Our results proved that super-enhancers regulate genes with high expression levels compared to the typical enhancers – through the collaboration of the individual enhancers that made up them – and most of these are indeed linked to breast cancerous processes.

5. Discussion

Continous development of sequencing techniques has made the understanding of regulatory regions at the genome level possible. There is a boosting need for processing the huge dataset resulting from sequencing associated with the exponentional development of bioinformatics. Therefore, we succeded in analyzing our data in an increasingly efficient way.

ERα is the primary drug target in estrogen sensitive breast cancer patients and it is known for almost three decades. However, its exact function has only been outlined in the last decades by the ChIP-seq technique. Contrary to our knowledge, genome-wide data indicates that only 7% of ERα transcription factor binding sites are located in the proximity of the promoter regions, while 93% of them act via distal, *cis* regulatory elements.

In 2013, *Richard A. Young* et al. identified a group of regulatory elements that can be found near cell type-specific genes that are suitable for the specification of different cell types. It has been proved that their effects are additive and together they could induce the extremely high expression of the target genes. These regulatory regions have been termed super-enhancers.

This work was based on an ERα ChIP-seq sample which enabled us to identify 81,518 ERα binding sites in the MCF-7 cell line upon estradiol treatment. During the identification of ERα-driven super-enhancers, we observed that in the absence of stimulation, future super-enhancers are represented by one or a few transcription factor binding event(s) even in the absence of stimulation.

According to the assumption that primary enhancers form the basis of super-enhancers, we referred to these elements as ‘mother enhancers’ and the subsequently appearing secondary enhancers we referred to as ‘daughter enhancers’. Our goals were to characterize mother and daughter enhancers and understand their role in ERα-driven transcriptional regulation.

In summary, the primary ChIP-seq peaks which were present prior to estradiol stimulation had significantly stronger binding elements than the activated peaks. This suggests that certain elements have high DNA–protein interaction affinities and that there is no need for further binding with other factors e.g. FoxA1 or AP2γ. Mother enhancers can be characterized by the presence of MED1 and upon ligand activation, further co-regulators such as P300 and BRD4 and the active histone mark H3K27ac also appear. However, FoxA1 and AP2γ also appeared where they have only secondary stabilizing role.

Unlike mother enhancers, transcription factor recruitment at daughter enhancers is limited. These regions are similar to any further enhancers and can be characterized by the presence of the more general, non-canonical estrogen response elements. They might be discriminated from typical enhancers by the proximity to mother enhancers or their higher ERα occupancy.

Our conclusion is that the existence of a canonical element provides competition between ERs, and the attracted ERs likely bind to neighboring non-canonical EREs rather than to similar non-canonical EREs in the distal regions of the genome. The super-enhancer region becomes more acetylated and MED1 and P300 bind to ERα and subsequently bind to a canonical ERE. To validate the role of canonical elements in super-enhancer formation, we investigated 5 additional ligand-inducible transcription factors, confirming the pivotal role of strong binding sites in super-enhancer formation.

The 81,518 ERα transcription factor binding sites that can be predictable from the MCF-7 cell line were categorized depending on whether they were present prior to E2-treatment or only after it and further divided into additional clusters based on the presence or absence of ERE. We characterized not just their dynamics but also investigated the presence of the possible co-factors

and the gene expression differences, resulted from the different types of ERα binding sites. We found that neither Type II ERα binding sites, which are able to bind to the DNA through ERE even in the absence of ligand, nor the Type III / Type IV binding sites, acting by binding to ERE only upon stimulation or facilitated by co-factors were unable to produce as strong gene expression as super-enhancers.

Our results further strengthen the much-debated view that super-enhancers are working as regulatory units and they can produce much robust transcriptional output than typical enhancers, even if they have the same characteristics and/or dynamics like mother and daughter enhancers. In this study, we first described and characterized ligand-inducible super-enhancers and our results have encouraged other research groups to further investigate the function and importance of the ERα-driven super-enhancers and the possible interaction between them – which was our assumption in this work.

6. Summary

Super-enhancers are regulatory units of the genome where enhancers are clustered and act together and have an extraordinary effect on the target genes. The concept was first mentioned in 2013. Since then, more than 500 studies have been published focusing on super-enhancers. Many of these studies identified the role of super-enhancers in regulating genes predisposing to various diseases. In this present work, we investigated the function of the ERα-driven super-enhancers in breast cancer-derived MCF-7 cell line and found that super-enhancers are typically marked prior to ligand activation by the presence of a pre-occupied ERα binding and these primary bindings can occur through a canonical response element in all cases. These response elements are so specific and capable of attracting a large amount of ERα molecules to a particular region of the genome. They also facilitate the occupation of the neighboring weaker response elements by ERα molecules rather than a distant genomic region thus forming super-enhancers. We referred to ERα-dense primary, "nucleating" binding sites as 'mother enhancers', while the additional, secondary regions appearing upon ligand activation were designated as 'daughter enhancers'. The mechanism of super-enhancer formation was validated on 5 additional human and mouse cell lines for 7 further ligand-inducible transcription factors as well.

By grouping and examining approximately 80,000 ERα transcription factor binding sites identifiable in the genome, it has been found that typical enhancers with the same properties and dynamics as the mother-enhancers are not capable of inducing the same transcription output; presumably because mother enhancers function as units of super-enhancers. Our results confirm not only the exceptional role of mother enhancers but also the importance of daughter enhancers. Altogether, these justify the identification of super-enhancers in different cells and understanding of the genes that they are regulated.

Acknowledgement

First and foremost, I would like to express my deepest gratitude to my supervisor, *Dr. Bálint László Bálint,* who encouraged me to learn bioinformatics during my M.Sc. studies. I am so greatful for his help, continuous support and motivation.

I am grateful to *Dr. Attila Horváth* and *Erik Czipa* who supported my work with patience in the first decade. Without them, I could not have to acquire the basics of bioinformatics. I am thankful to *Edina Erdős, Lilla Ozgyin*, and *Noura Faraj*, who not just provided helpful discussions and insightful comments but they were beside me in everyday life as my friends.

I am grateful to *Prof. Dr. József Tőzsér* and *Prof. Dr. László Fésüs,* the former and recent heads of the Department of Biochemistry and Molecular Biology respectively, who made it possible to do my Ph.D. studies in the Doctoral School of Molecular Cell and Immune Biology in a very good atmosphere and among researchers with great professional knowledge.

I am also thankful to the collaborators: Dr. Katalin Dánielné Sándor, Dr. Adeline Divoux, Dr. Timothy F. Osborne, and Steven R. Smith who not just provided data for me but also whom I could learn a lot from. Special thanks to *Lívia Szántó-Kiss* and *Nóra Elek* for their immense patience and help; without them I could have given up the opportunity offered by many professional events.

I would like to express my deepest gratitude to *my beloved family* and *Gergő.* I am thankful for their love and for Gergő's critical attitude; he always encouraged me to be better and he is not only as a friend but also as a partner that was with me all the time.

This work was supported by the ÚNKP-17-3-DE-140 and ÚNKP-18-3-III-DE-253 New National Excellence Program of the Ministry of Human Capacities and by the EFOP-3.6.3-VEKOP-16-2017-00009 scholarships.

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/296/2019.PL
Subject: PhD Publikációs Lista

Candidate: Dóra Bojcsuk
Neptun ID: BD1BZT
Doctoral School: Doctoral School of Molecular Cellular and Immune Biology

List of publications related to the dissertation

1. **Bojcsuk, D.**, Bálint, B. L.: Classification of different types of estrogen receptor alpha binding sites in MCF-7 cells.
   J. Biotechnol. 299, 13-20, 2019.
   DOI: http://dx.doi.org/10.1016/j.jbiotec.2019.04.016
   IF: 3.163 (2018)

2. **Bojcsuk, D.**, Nagy, G., Bálint, B. L.: Inducible super-enhancers are organized based on canonical signal-specific transcription factor binding elements.
   Nucleic Acids Res. 45 (7), 3693-3706, 2017.
   DOI: http://dx.doi.org/10.1093/nar/gkw1283
   IF: 11.561

List of other publications

3. Divoux, A., Sándor, K., **Bojcsuk, D.**, Talukder, A., Li, X., Bálint, B. L., Osborne, T. F., Smith, S. R.: Differential open chromatin profile and transcriptomic signature define depot-specific human subcutaneous preadipocytes: primary outcomes.
   Clin Epigenet. 10 (1), 1-15, 2018.
   DOI: http://dx.doi.org/10.1186/s13148-018-0582-0
   IF: 5.496

4. Márton, J., Péter, M., Balogh, G., Bódi, B., Vida, A., Szántó, M., **Bojcsuk, D.**, Jankó, L., Bhattoa, H. P., Gombos, I., Uray, K., Horváth, I., Török, Z., Bálint, B. L., Papp, Z., Vígh, L., Bai, P.: Poly(ADP-ribose) polymerase-2 is a lipid-modulated modulator of muscular lipid homeostasis.
   Biochim. Biophys. Acta. Mol. Cell Biol. Lipids. 1863 (11), 1399-1412, 2018.
   DOI: http://dx.doi.org/10.1016/j.bbalip.2018.07.013
   IF: 4.402

UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

5. Vető, B., **Bojcsuk, D.**, Bacquet, C., Kiss, J., Sipeki, S., Martin, L., Buday, L., Bálint, B. L., Arányi, T.: The transcriptional activity of hepatocyte nuclear factor 4 alpha is inhibited via phosphorylation by ERK1/2.
   PLoS One. 12 (2), 1-19, 2017.
   DOI: http://dx.doi.org/10.1371/journal.pone.0172020
   IF: 2.766

6. Ozgyin, L., Erdős, E., **Bojcsuk, D.**, Bálint, B. L.: Nuclear receptors in transgenerational epigenetic inheritance.
   Prog. Biophys. Mol. Biol. 118 (1-2), 34-43, 2015.
   DOI: http://dx.doi.org/10.1016/j.pbiomolbio.2015.02.012
   IF: 2.581

7. **Bojcsuk, D.**, Erdős, E., Bálint, B. L.: Az ösztrogénreceptor mûködése a legújabb genomikai kutatások tükrében.
   LAM KID. 4 (2), 79-84, 2014.

8. **Bojcsuk, D.**, Sipos, L., Bálint, B. L.: A mikro-RNS-ek mint a hormonok egy új családja.
   LAM KID. 3 (4), 29-33, 2013.

Total IF of journals (all publications): 29,647
Total IF of journals (publications related to the dissertation): 14,724

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

22 August, 2019