INFORMATION INTEGRATION APPROACHES FOR INVESTIGATING ESTROGEN RECEPTOR MEDIATED TRANSCRIPTION

By

Hatice Ulku Osmanbeyoglu

BS in Computer Engineering, Northeastern University, 2004

MS in Electrical and Computer Engineering, Carnegie Mellon University, 2006

MS in Bioengineering, University of Pittsburgh, 2009

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2012

UNIVERSITY OF PITTSBURGH

School of Medicine

This dissertation was presented

by

Hatice Ulku Osmanbeyoglu

It was defended on

November 16, 2012

and approved by

Steffi Oesterreich, Ph.D.

Professor Department of Pharmacology and Chemical Biology

Panayiotis V. Benos, Ph.D.

Associate Professor, Department of Computational and Systems Biology

Roger Day, Sc.D.

Associate Professor, Department of Biomedical Informatics

Dissertation Advisor: Xinghua Lu, M.D., Ph.D., M.S.,

Associate Professor, Department of Biomedical Informatics

INFORMATION INTEGRATION APPROACHES FOR INVESTIGATING ESTROGEN RECEPTOR MEDIATED TRANSCRIPTION

Hatice Ulku Osmanbeyoglu, PhD

University of Pittsburgh, 2012

Estrogen plays essential roles in the function of normal physiology and diseases. Its effects are mainly mediated through two intracellular estrogen receptors, ERα and ERβ, which belong to a family of nuclear receptors (NRs) functioning as transcription regulators. In the first part of this thesis, we aim to derive a holistic view of the transcription machineries at estrogen-responsive genes and further, to reveal different mechanisms of estrogen-mediated transcription regulation. In order to achieve this, we integrated and systematically dissected a variety of genome-wide high-throughput datasets, including gene expression arrays, ChIP-seq, GRO-seq, and ChIA-PET. Our analyses have led to the following novel findings: In the absence of the ligand, most of the estrogen-responsive genes assumed a high-order chromatin configuration that involved Pol II, ERα and ERα-pioneer factors. Without the ligand, estrogen-induced genes showed active transcription at promoters but failed to elongate into gene bodies, and such a pause was lifted after estrogen treatment. However, the estrogen-repressed genes showed coordinated transcription at promoters and gene bodies in the absence and presence of estrogen. Through information integration, we inferred that, for estrogen-repressed genes, the majority of the high-order chromatin complexes containing actively transcribed genes were disrupted after estrogen treatment. The analyses led to the hypothesis that one mechanism for estrogen-mediated repression is through disrupting the original transcription-favoring chromatin structures.

Further, nuclear receptors such as ERs interact with co-regulators to regulate gene transcription. Understanding the mechanism of action of co-regulator proteins—which do not bind DNA directly, but exert their effects by binding to transcription factors—is important for the study of normal physiology as well as diseased conditions. However, due to the nature of detecting indirect protein-DNA interaction, ChIP-seq signals from co-regulators can be relatively weak and thus biologically meaningful interactions remain difficult to identify. In the second part of this thesis, we investigated and compared different machine learning approaches to integrate multiple types of genomic and transcriptomic information derived from our experiments and from public databases. This helped us to overcome the difficulty of identifying functional DNA binding sites of the co-regulator SRC-1 in the context of estrogen response. Our results indicate that supervised learning with the naïve Bayes algorithm significantly enhanced the peak calling of weak ChIP-seq signals and outperformed other machine learning algorithms. Our integrative approach revealed many potential ERα/SRC-1 DNA binding sites that would otherwise be missed by conventional peak calling algorithms with default settings.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

I am very thankful to many people who were directly and indirectly involved in the completion of my thesis. This work would not have been possible without the support of my mentors, friends and family. I owe a debt of gratitude to my research advisor, Dr. Xinghua Lu, who guided me through the process of dissertation research. Whenever I needed his advice, his door was always open. This dissertation would not have been possible without his support, understanding and encouragement.

I want to thank the rest of the thesis committee for their time, and insightful comments. I would like to extend my special thanks to Dr. Roger Day who has always spared time when I needed advice. I've learned so much from him. Special thanks go to Dr. Steffi Oesterreich for opening doors to this exciting research area on estrogen receptors and Dr. Takis Benos for his valuable suggestions. I would like to also thank everybody at the Department of Biomedical Informatics and specifically my colleagues in the Lu Lab, Toni, Nova, Genine, and Yolanda.

I thank my friends, who shared the pain, joy, laughter, and pride of my research over the years, sometimes from miles away. I also thank all my friends in the Islamic Center of Pittsburgh who have made living in Pittsburgh memorable and invaluable.

Last but not the least; I would like to express my gratitude to my parents and siblings for their unconditional love and moral support throughout my studies. Words fail me to thank them enough. This is the only way to express my sincere gratitude towards them.

INTRODUCTION

Genes indirectly code for proteins. The process of manufacturing proteins from the genetic code in DNA is called *gene expression*. Most of the genes are expressed in a tightly controlled manner in only part of the organism, or under particular conditions. In a diseased state, this organization is disturbed. In order to regulate gene expression, collections of regulatory proteins interact with the specific sequences in the promoter as well as the enhancer regions of targeted genes.

Transcription factors (TFs) often interact with other proteins, which further modulate the function and the efficacy of TFs to achieve fine-tuned regulation of gene expression. Studying such interactions and regulations is an increasingly important component of studying gene expression systems. In cancer formation, gene expression might be deregulated by misactivated TFs and/or by mutations and translocations of DNA regions. A large amount of oncogenic signaling pathways converge on such sets of TFs that ultimately control gene expression patterns resulting in cancer development [9].

Nuclear receptors (NRs), such as estrogen receptor alpha (ERα), are transcription factors that migrate to the nucleus (often as a result of binding ligand) to regulate downstream target genes. NRs play important biological roles in normal physiology and certain diseases. Upon ligand binding, ERα and other NRs are bound by proteins called co-regulators that recruit transcriptional machinery and chromatin modifying enzymes. Co-regulators are therefore critical

in NR activity. Understanding the composition of functional NR/co-regulator complexes in specific signaling contexts could provide a basis for the development of novel NR- and co-regulator-targeted therapeutics.

The last decade has seen an explosion in the amount of genomic, proteomic and phenotypic data and new high-throughput technologies related to NRs: the expression patterns of NRs, co-regulators and their target genes (transcriptomics); the binding of ligand- and tissue-specific functional NR and co-regulator sites to DNA (cistromics); the organization of NRs into higher order complexes; and the downstream effects of NRs on homeostasis and metabolism (metabolomics). Integrating and synthesizing this rich and heterogeneous information has great potential to provide novel insights into the biological mechanisms of NRs and the diseases related to them. Significant bioinformatics challenges lie ahead. We need new bioinformatics frameworks a) to solve issues related to noise in high-throughput datasets and b) to integrate these large-scale datasets into meaningful models of NR and co-regulator biology.

1.1 THE ROLE OF BIOINFORMATICS IN NUCLEAR RECEPTORS

In this section, I will describe two different bioinformatics challenges related to NRs specifically ERα and its co-regulators.

1.1.1 Integrative analysis of discovery-driven datasets

The emergence and application of genome-wide high-throughput technologies have enabled us a new way of understanding how ERα functions in the cell. For instance, microarray technology has enabled us to explore global features of hormone-regulated gene expression through nuclear receptors (NRs). Moreover, by measuring instantaneous transcriptional activity, global run-ons sequencing (GRO-seq), can help us to address direct transcriptional responses for a particular hormone signaling pathway. In addition, ChIP-seq (chromatin immunoprecipitation followed by sequencing) enables the accurate genome-wide profiling of transcription factors, co-regulators, RNA Pol II, histone modification binding sites as well as DNA methylation. Lastly, ChIA-PET **(**chromatin interaction analysis with paired-end tag sequencing**)** captures long-range chromatin looping on a genome-wide scale. However, each data source provides a very narrow view of the transcription regulation. An important bioinformatics goal is the successful integration of these high-content data-sets utilizing computational and mathematical approaches to detect statistically significant patterns and trends across these diverse datasets. This approach allows testing new hypothesis, eventually leads biological discovery.

In the first part of this thesis, I integrated a variety of recent genome-wide high-throughput datasets, including gene expression arrays, ChIP-seq, GRO-seq and ChIA-PET in order to derive a holistic view of the transcription machineries at estrogen-responsive genes, and reveal different mechanisms of estrogen-mediated transcription regulation in MCF-7 breast cancer cell line. In doing so, our analyses have led to the many novel findings.

1.1.2 Dealing with experimental noise in co-regulator ChIP-seq dataset

Recently, chromatin immunoprecipitation coupled with high-throughput next-generation sequencing (ChIP-seq) has become the main technology for global characterization of the transcriptional impact of NRs and their co-regulators [1-3]. ChIP-seq involves the short-read (~30bp) sequencing of the ChIP-enriched DNA fragments. These short sequence reads/tags are then aligned to a reference genome. Then the actual binding loci from the positional tag distributions (i.e. sequenced DNA fragments mapped onto a reference genome sequence) are determined using 'peak calling' algorithms. Studying the indirect interactions between TFs and their co-regulators through ChIP-seq technology poses an additional challenge since co-regulators do not directly bind DNA. Co-regulator ChIP-seq measures the indirect protein-DNA binding through primary TFs and leads to relatively weak sequencing signals—i.e. relatively small number of sequence tags above noise. As such, it remains a challenge for contemporary peak calling methods to detect weak indirect protein-DNA-binding signals and simultaneously maintain a high specificity.

The ability to improve ChIP-seq peak calling by utilizing available sources of biological information for indirect co-regulator binding in the presence of weak ChIP-seq signal is an important research area. Due to the intrinsic variability in the affinity of interactions between a TF and its co-regulators, it is inevitable that the ChIP-seq signal of these types of studies would span a broad spectrum and that the weak signal scenario would be likely to occur often. The need for the methods to address this problem is acute considering the increasing number of studies using ChIP-seq to study NR and their co-regulators due to their importance in normal development and in many diseases such as breast cancer.

In the second part of this thesis, to overcome this problem, I defined a semi-supervised classification task and integrated multiple types of genomic and transcriptomic information derived from small-scale experiments and public databases. With this approach, I showed a general framework for utilizing limited amounts of prior knowledge (both from small-scale experiments and from multiple types of biological data) to enhance the sensitivity and specificity of results of high-throughput technologies.

1.2 DISSERTATION OVERVIEW

Chapter 2 provides background information on transcription mechanism related to estrogen receptor alpha (ERα). Chapter 3 presents integration of a variety of recent genome-wide high-throughput datasets in order to derive a holistic view of the transcription machineries at estrogen-responsive genes, in order reveal different mechanisms of estrogen-mediated transcription regulation by using hypothesis driven statistical approach. Chapter 4 describes a machine learning framework general framework for utilizing limited amounts of prior knowledge (both from small-scale experiments and from multiple types of biological data) to enhance the sensitivity and specificity of results of ChIP-seq 'peak calling' algorithms. Chapter 5 presents conclusions with discussions on future research work.

2.0 BACKGROUND

2.1 ESTROGEN RELATED DISEASES

Estrogens, the most common estrogen being 17β-estradiol (E2), are a class of sex steroid hormones that are synthesized from cholesterol and are secreted primarily by the ovaries, with contributions from placenta, adipose tissue, and adrenal glands. Estrogen is essential in both sexes and has functions not only in reproductive system but also in the musculoskeletal system, central nervous system, hypothalamic pituitary axes, cardiovascular system and immune system. After arriving at target tissues through the blood, estrogens and their cellular receptors regulate many aspects of healthy physiology. Its effect is mainly mediated through two intracellular estrogen receptors, ERα and ERβ, which belong to a family of nuclear receptors functioning as transcription regulators. However, estrogens and ERs are also involved in human diseases including breast cancer.

2.1.1 Breast cancer

Estrogens play a central role in breast cancer which is the top common cancer type diagnosed in women. Women with higher lifelong exposure to estrogen (resulting from early menses and late menarche) have an elevated risk of breast cancer. In 2012, the American Cancer Society

estimates 226,870 new cases of invasive breast cancer and as many as 39,510 breast cancer deaths in the United States [1]. Since 1990, the death rates from breast cancer have been decreasing as a result of earlier detection through screening and increased awareness, as well as improved treatment.

The first connection between estrogen and breast cancer was recorded in 1896 when Dr. George Beatson of the Glasgow Cancer Hospital observed that bilateral oophorectomy in patients with inoperable neoplasia reduced the aggressiveness of these tumors [2]. Since then it has been established that estrogen metabolites are involved in the initiation process through oxidative DNA damage and estrogens themselves enhance cell proliferation, leading to tumor promotion [3]. Therefore, most breast cancer therapies focus on disruption or modulation of the effects of estrogen signaling using selective estrogen receptor modulators (or SERMs) such as Tamoxifen and Raloxifene. Unlike estrogens, SERMs do not purely act as agonists for ERs, nor are they pure antagonists. They exhibit tissue-specific modulation of ER signaling, activating genes is some tissues which they inhibit in others[4]. In general, ERα has long been determined to be a prognostic marker for breast cancer. Moreover, increased survival is seen with ERα-positive status as these tumors respond to anti-estrogen therapy. Further understanding of the molecular mechanism underlying estrogen-mediated ER action may result in better treatments for breast cancer.

2.2 FACTORS IN ESTROGEN SIGNALING

2.2.1 Estrogen receptor

ERs are classical hormone nuclear receptors and members of the nuclear receptor super family. ERs exist in 2 main forms, ERα and ERβ, which are encoded by separate genes *ESR1* and *ESR2*, respectively. Each has distinct tissue expression patterns, post-translational modifications, and cellular localization in normal and disease states. ERα is the predominant receptor in the bone, uterus, liver, and adipose tissue, whereas ERβ is the predominant receptor in the ovary and intestinal tract. The brain, mammary gland, and cardiovascular system express both ERs.

ERα and ERβ show significant overall sequence similarity. As a member of the nuclear receptor super-family, ERs possesses a zinc finger DNA binding domain (DBD), a ligand binding domain (LBD), and two domains involved in its activation and/or repression, AF-1 and AF-2 respectively. The primary transactivating domain, AF-1 resides near the N-terminus of the protein, occupying the A/B region. The DBD is located in region C, close to the center of the protein, and is flanked on the C-terminal end by an unstructured region D, also known as the hinge region. The LBD resides in region E, following on the C-terminal side of the hinge region. In ERα, the LBD is responsible for the majority of dimer stabilization following E2 binding and consists of 12 alpha helices which form a ligand-binding pocket[5]. This pocket binds estrogens, SERMs, and several estrogen-like polycyclic compounds[4]. A unique feature of this region is helix 12, the second transactivating domain also known as AF-2. The domain exhibits estrogen sensitivity and is only active once the ligand is bound. Once activated by ligand binding, the conformation of AF-2 shifts, enabling ERα to mediate regulation of gene expression.

2.2.2 Co-regulators

Co-regulators are proteins that interact directly with nuclear receptors to form a bridge between the receptor and basal transcriptional machinery to regulate gene transcription [6, 7]. Co-regulators can activate (co-activators) or repress (co-repressors) the transcription activity of nuclear receptors, including ERs. They can be divided into distinct classes based on their biochemical and functional activity including bridging factors, protein-modifying enzymes, protein-demodifying enzymes, chromatin remodeling complexes, and mediator complexes [7]. Co-regulators are recognized to be critical for proper function of ERs, and alterations in co-regulator function and expression are associated with cancer and other diseases. Therefore, assessment of ER co-regulator status and activity is crucial to determine role of ERs in disease progression and to predict prognosis and response to therapy.

Steroid receptor co-activator [SRC] p160 co-activators are well known factors that are recruited to ERα including SRC-1 (NCOA1), SRC-2 (NCOA2, Tif2, GRIP1) and SRC-3 (NCOA3, AIB1). The first co-regulator discovered was SRC-1[8]. Some of the other co-activators of ERα are BCAS3, BRG1, CARM1, CBP, CITED1, Cyclin D1, DBC1, E6-AP, GCN5L2, MUC1, p300, PELP1, SRA[9].

Co-repressors are proposed to provide a counterbalance to the estrogen-induced transactivation, and represent a potential mechanism employed by the cell to regulate hormonal responses. Some of the co-repressors of ERα are nuclear receptor corepressor-1 (NCOR1)[10],

silencing mediator of retinoic acid and thyroid hormone receptor (SMRT/NCOR2) [11] and SAFB1[12].

2.2.2.1 SRC-1 role in breast cancer metastasis

SRC-1 has been associated with execution of breast cancer metastasis and mediation of resistance to endocrine therapies which is increasingly prevalent among breast cancer patients. Currently, two distinct mechanisms have been elucidated for the role of SRC-1 in breast cancer metastasis [13]; SRC-1 1) impacts on the epithelial to mesenchymal transition (EMT) and epithelial depolarization via regulation of the EMT transcription factor Twist, 2) upregulates the expression of integrin α5 to promote cell migration and invasion. With respect to the development of resistance, SRC-1 overexpression was shown to convert Tamoxifen, a selective estrogen receptor modulator (SERM), from a transcriptional repressor to a transcriptional activator in breast cancer [14]. Tamoxifen is a common anti-estrogen therapy prescribed for pre-menopausal ERα-positive breast cancers.

2.2.3 Pioneer factors

Pioneer factors are a special class of transcription factors. They physically interact with condensed chromatin to facilitate the binding of additional transcription factors. Several TFs have been shown to act as pioneer factors for ERα including FoxA1, PBX1, AP2γ and GATA [15-18]. These factors possibly recruit chromatin modifiers and generate a local environment that is more accessible for ERα binding and help rapid transcriptional response of ERα [19, 20].

FoxA1 (Hepatocyte Nuclear Factor 3α) is an ER pioneer factor that promotes ER binding to chromatin. FoxA1 was shown to be recruited to sites of H3K4me1 and H2K4me2 (please indicate which histones these are, and which amino acid the methylation is) and initiated chromatin remodeling events [20]. Furthermore, genome-wide mapping of FoxA1 and ER revealed that half of all ER binding sites overlap with FoxA1 binding regions in the genome [16, 20]. The pre-bound FoxA1 appears to be required for facilitating ERα recruitment and modifying chromatin structure at the regulatory regions of estrogen-target genes in both induced and repressed genes. Moreover, after observing motifs of AP2 transcription factors are enriched within ER-binding motifs, AP2γ (encoded by *TFAP2C*) was identified as a putative pioneer factor for ER [17]. ChIP-seq study showed AP2γ overlaps with approximately 50% of all ERα binding events in the MCF-7 breast cancer genome and the majority of these shared regions also overlap with FoxA1[17]. In addition, recently PBX1 (Pre-B-cell leukemia homeobox 1) is shown to be a putative pioneer factor [18]. Approximately half of the ERα binding events were overlapped with PBX1 in the MCF-7 breast cancer genome. However, whether AP2 γ or PBX1 can directly bind to condensed chromatin independently of other factors is not proven yet[21].

2.2.4 Histone modifications

The fundamental unit of chromatin is the nucleosome. Each basic unit consists of DNA wound around an octamer of four core histones (H3, H4, H2A, and H2B). The N- and C- terminal tails of histones are subject to several post-translational modifications including acetylation, methylation, phosphorylation, sumoylation, ubiquitination, ADP ribosylation, deimination, and proline isomeration[22]. Some histone modifications alter the packing of chromatin by opening

or closing the DNA through changes in electrostatic charge or inter-nucleosomal contacts. These modifications control the access of transcription factors to DNA[23]. Moreover, some of these modifications promote the recruitment of chromatin binding proteins[23].

Histone modifications also play a significant role in ERα-mediated transcription as well as in cancer progression [24]. For instance, acetylation and deacetylation of conserved lysine residues present in histone tails have been suggested as a mechanism by which ERα modifies chromatin structure[25]. Briefly, acetylation of lysine residues results in a neutralization of the net positive charge, resulting in a net negative charge thereby causing decreased histone-DNA interaction[23]. This effectively opens up the chromatin and generally associates with active transcription.

The methylation of histones is also an important regulatory signal in ERα-mediated gene expression [26]. Methylation does not affect the histone-DNA interactions but these modifications regulate transcription by recruiting distinct effector proteins that alter the chromatin environment in favor of either activation or repression. Different methylation states, i.e. unmethylated, mono-, di-, or trimethylated (Kme1, Kme2, and Kme3), recognize different effector proteins with unique enzymatic activities and thereby differentially influence transcriptional regulation[23]. For example, H3K4 methylation is linked with activation, while H3K9 methylation correlates with repression[24].

2.3 BASIC MECHANISM OF ESTROGEN RECEPTOR REGULATION OF GENE EXPRESSION

2.3.1 Transcription activation

ERα regulation of gene expression is ligand-dependent [27, 28]. Estrogen binding to the ERα causes a conformational change. This allows the estrogen-ERα complex to bind to specific regulatory elements in the target genes. The estrogen response element (ERE) is a 13 nucleotide inverted palindrome and is known as the best characterized ERα regulatory element. The E2-ERα complex binds directly to the ERE as a dimer. Besides, estrogen-ERα complex can bind indirectly to chromatin by protein-protein interactions with transcription factors such as AP-1 and NFκB [29, 30]. This non-classical mechanism is often described astranscription cross-talk. The estrogen-ERα complex can also bind to DNA adjacent to TFs such as FoxA1 and Sp1, which stabilize ERα binding and promote the assembly of the transcription complex. After the estrogen-ERα complex is tethered to the regulatory element, it can recruit co-regulatory proteins. The transcriptional activation of ERα target genes involve the interaction of estrogen-ERα complex and co-activators with mediator proteins and basal transcription factors. The structure of chromatin is changed through histone acetylation and other modifications, and then RNA Pol II initiates gene transcription.

2.3.2 Transcription repression

The mechanisms of ERα-mediated transcriptional repression and regulatory elements are less unknown but co-repressors are involved in transcriptional repression. A handful of studies have addressed potential mechanisms of estrogen repression. Existing ERα-mediated gene repression through genomic actions can be divided into three main categories. Repression occurs when ERα; 1) binds directly to DNA, recruits co-repressors and HDACs and displaces RNA Pol II which result in more condensed chromatin configuration [31-33]; 2) competes with other transcription factors for co-regulators resulting in reciprocal repression[34]; 3) inhibits transcription activators by decreasing the recruitment of transcription factors onto the DNA or by interfering with their gene-activating functions[35].

One way of ER-mediated repression resulting in a displacement of RNA Polymerase II may occur through the direct binding of the ERα onto the DNA and the ERα-induced formation of co-repressor complexes which coincides with a more condensed chromatin conformation. Initially, overexpression of co-regulators has shown to repress genes in the presence of ligand. For example, overexpression of SAFB1 (scaffold attachment factor B) enhanced repression of E-cadherin [33] and overexpression of SMRT and SAFB1 enhanced repression of folate receptor α [36]. Later, active recruitment of repressive complexes in the presence of ligand —for example, NCoR, (nuclear receptor corepressor), histone deacetylase 1 (HDAC1), and CtBP1 to the CCNG2 promoter [32, 37]; NCoR and SMRT (silencing mediator of RAR and TR) to the VEGFR2 promoter [38]; NCoR and TAB2 to the BMP7, ABCG2, and BCL3 promoters [39]; NCoR, NRIP1 and SMRT to *PSCA* and *SLC35A1* promoter [40]; TTF-2 (Thyroid transcription factor-2) to pS2 and cyclin D1promoter [41]—has also been shown. Genome-wide analysis of

ERα binding sites implicated the involvement of the co-repressor NRIP1 (nuclear receptor interacting protein 1) in the estrogen-mediated repression of genes such as BCAS4, IRX4, GUSB and MUC1 at late time points [42]. These genes are most likely secondary targets of ERα since they require the estrogen-induction of NRIP1 for their repression. In another study, the recruitment of ERα, HDAC7, and FoxA1 to the RPRM promoter was associated with dissociation of RNAPII from the RPRM promoter and repression of the gene [31]. Therefore, HDAC7-mediated repression was suggested as a common mechanism for a subset of E2-repressed genes. Recently, recruitment of PITX1 (paired-like homeodomain transcription factor) to near ERα binding sites (enhancer or promoter regions) was shown to inhibit the transcription of ERα target genes in the presence of ligand [43].

Another possible mechanism for ERα-mediated repression is that ERα competes with other transcription factors for co-regulators resulting in reciprocal repression. For example, ATBF1 significantly inhibits the ERα function by selectively competing with AIB1 (SRC3) for binding to ERα in ERα-positive breast cancer cells in the presence of ligand [44]. Another study [34, 45] demonstrated that proto-oncogene ERBB2 repression as a result of estrogen-bound ERα and another TF (most likely activator protein [AP]-2) for the SCR-1 because overexpression of SRC-1 but not SRC-2 or SRC-3 relieves repression of ERBB2. Sometimes co-repressors can compete with co-activators at the regulatory regions. A genome-wide ERα binding study found that the paired box 2 gene product (PAX2) functions as a transcriptional repressor and competes with AIB1/SRC3 for binding and regulation of ERBB2 transcription [46].

A possible mechanism how estrogen-induced ER mediates repression is by inhibiting transcription activators. For example, estrogen induced ERα was shown to interact with HNF-4α

and alter binding of HNF-4α to the HBV enhancer. As a result the transcription of HBV genes was repressed [35].

Recently, the estrogen-mediated epigenetic repression of large chromosomal regions through DNA looping was shown. Hsu et al [47] characterized the influence of estrogen signaling on the long-range epigenetic silencing (LRES) and uncovered 11 large repressive zones including a 14-gene cluster located on 16p11.2. Looping dynamics was lost and epigenetic silencing occurred after estrogen treatment in the breast cancer cells. However, in normal cells, estrogen caused transient formation of multiple DNA loops in the 16p11.2 region by bringing 14 distant loci to focal ERα-docking sites for coordinate repression.

With the availability of genome-wide binding studies, the patterns of ERα with respect to both estrogen induced and repressed genes has been studies. Earlier, genome-wide ChIP-on-chip analyses have shown that there is an over-representation of ERα binding events near (within 50 kb) the transcription start sites of induced genes and an underrepresentation of ERα binding sites in the just early estrogen-repressed genes [42]. Later, Lin et al. [48] reported in a genome-wide ChIP-PET studies that there is a lower frequency of EREs in repressed sites and an enrichment of ERα binding events in the promoter regions of induced genes. Whereas, they found ERα binding events adjacent to repressed genes are more dispersed and are not localized to a specific region relative to the target gene. However, in a ChIP-seq study, Stender et al [49] reported that the majority of genes repressed by estrogen also require the ERα with a functional DNA binding domain. Their result emphasizes the importance of DNA binding in some repression activities of the ERα. Recently, a meta-analysis of genome-wide binding studies [50] suggested that ERα and other NRs such as RARA and RARG in MCF7 were close to both induced and repressed genes.

2.4 TISSUE SPECIFIC REGULATION OF GENES BY ESTROGEN MEDIATED ESTROGEN RECEPTORS

Studies indicate that differential expression of ERs, co-regulators, and transcription factors in various tissues, the presence of different regulatory elements for ERs and epigenetic modifications are involved tissue-specific gene regulation in response to estrogen. One mechanism for tissue-specific gene regulation of estrogen is by binding to different regulatory elements. For example, many ERα and ERβ binding sites were different in U2OS and MCF-7 breast cancer cells [51, 52]. As the binding sites are different, it is expected that the genes regulated by ERα and ERβ in response to estrogen will be different. In fact, previous microarray data demonstrated that ERα, ERβ and ERα/ERβ heterodimer regulated distinct set of genes [53, 54]. Another way that estrogen regulates distinct genes in different tissues is through the existence of tissue-specific regulatory elements in ER target genes. Most of the ER regulatory elements require transcription factors for activity, including AP1, FoxA1, and Sp1. Different estrogen target genes are activated as a result of collaboration between different transcription factors. Another possibility for the tissue-specific effects of estrogens is that the ERα and ERβ cistromes (cis-regulatory elements that ERs interacts with throughout the genome) are tissue specific. For example, studies based on tiling arrays and ChIP sequencing showed very little overlap between the binding sites for ERα in MCF-7 and U2OS cells[55, 56](Hatmerier et al).

Different ERα and ERβ cistromes in various tissues likely account for some differences in gene expression profiles observed after E2 treatment.

The differential expression of co-regulators in various tissues could also lead to tissue-specific effects. For instance, coactivator-associated arginine methyltransferase 1 (CARM1) is shown to inhibit estrogen-mediated proliferation and gene regulation when overexpressed in MCF-7 cells [57]. In the presence of CARM1, nearly 16% of genes induced by estrogen, including proliferative genes, were repressed.

Another likely possible mechanism whereby estrogen regulates different genes in tissues is through the differentiation state of the cells. During differentiation, there is different expression of transcription factors and co-regulatory proteins as well as epigenetic modifications that can determine which genes are regulated by ERs. Moreover, certain ER target genes are turned off by epigenetic changes. During cell differentiation, several epigenetic modifications occur in the chromosomes without altering the DNA sequence such as DNA methylation, histone modifications, and nucleosome positioning[58]. For example, FoxA1 recognizes H3K4me1 and H2K4me2 near an ERα binding site [55] and interacts with ERα to open up chromatin structure and facilitate the recruitment of transcription factors leading to increased transcription [55]. These findings showed that epigenetic changes are important to mark the sites where transcription factors bind and interact with ERαs. Therefore, epigenetic modifications in different target genes that occur during differentiation can determine if the gene will bind transcription factors at ERα regulatory elements.

2.5 DNA DAMAGE AND ESTROGEN

Previous studies have shown that transcriptional activation in response to stimuli, including estrogen, involves the formation of DNA damage including DNA double-strand breaks (DSBs), the recruitment of DNA repair proteins, and large-scale genome reorganization to allow movement of activated genes and regulatory loci to transcriptional hubs. The DNA damage during transcription might lead to the cancer formation. For example, unresolved DSBs can lead to cell cycle arrest, senescence and apoptosis. Moreover, if illegitimately repaired, DSBs can seed the formation of genomic rearrangements, amplifications, and deletions [59].

Initially, several estrogen metabolites have shown to can cause DNA [60]. In prior animal models, estrogen-induced direct or indirect DNA damage was observed [61]. Ju et al. [62]showed that during transcriptional activation by ERα, transient DSBs generated at regulatory elements of ERα-regulated genes. Recently, Williamson et al. also showed the formation of stable TOP2B-DNA cleavage complexes leading to DSBs and initiation of homologous recombination repair (HRR) in ERα-present breast cancer cell line [63]. From these studies as well as studies related to androgen receptor (AR) [64, 65], an emerging model suggests that DSB can be mediated by the class II topoisomerase TOP2B, which is recruited to the AR and ER regulatory sites on target genes for efficient transcriptional activation. These DSB are recognized by the DNA repair machinery causing the recruitment of repair proteins. Therefore, Haffner et al. [66] proposed a hormone cycling therapy, in hormone dependent tumors like breast and prostate cancers, to induce DSBs repetitively in combination with topoisomerase II poisons or inhibitors of the DNA repair components in order to overwhelm the cancer cells with breaks, ultimately leading to cell death.

2.6 TRANSCRIPTION FACTORIES

The 3.4 billion base pairs of the human genomes are packed in hierarchical structures in a cell nucleus of 1 um diameter. The spatial organization of chromatin and proteins in the nucleus is extremely important for regulation of gene expression and replication. These loops often bridge distant chromatin locations, even located on different chromosomes. Actively transcribed regions in the genome have been supposed to cluster in “transcription factories” [67] with concentrated RNAPII (for detailed review [68, 69]). The emerging view is that the loop is attached to a “transcription factory” through components of the transcription machinery (either polymerases or transcriptional activators/repressors [70-74]). The position of a gene within a loop might determine how often a gene is transcribed [75]. Moreover, loops are dynamic and their state can rapidly altered during activation or repression The 3.4 billion base pairs of the human genomes are packed in hierarchical structures in a cell nucleus of 1 um diameter. The spatial organization of chromatin and proteins in the nucleus is extremely important for regulation of gene expression and replication. These loops often bridge distant chromatin locations, even located on different chromosomes. Actively transcribed regions have been thought to cluster in “transcription factories” [67] with concentrated RNAPII (for detailed review [68, 69]). The emerging view is that the loop is attached to a “transcription factory” through components of the transcription machinery (either polymerases or transcriptional activators/repressors [70-74]). The position of a gene within a loop might determine how often a gene is transcribed [75]. Moreover, loops are dynamic and their state can be rapidly altered during transcriptional activation or repression.

Recently, many studies published related to the long-range enhancer-promoter interactions through chromatin looping and their possible roles in transcription regulation [16,

70, 76-83]. For instance, the first 3D interaction map of RNAPII occupied sites for five different cell-lines was recently established using chromatin interaction analysis by paired-end-tag sequencing (ChIA-PET) [70]. RNAPII is shown to be involved in the promoter–promoter interactions between proximal and distant genes. As a result, multi-gene complexes cooperatively regulate their activity. In another previous study, estrogen simulation of ERα present breast cancer cell line has shown to result in large-scale alterations in genome organization arising from movement of different target gene loci and distal regulatory elements to these transcriptional hubs, establishing cell-type specific expression patterns using ERα ChIA-PET [82]. This kind of genome-wide high-throughput analysis not only revealed role factors in the network of long-range interactions of the 3D chromatin structure, but also showed the complexity of transcriptional states of genes in response to stimuli. Moreover, nuclear structure including chromatin is altered in tumor cells. Changes in the spatio-temporal organization of nuclear structure can induce altered gene expression programs in tumor cells [51]. Therefore, further investigation of 3D chromatin structure is important to advance our knowledge of diseases.

2.7 EXPERIMENTAL TECHNIQUES TO STUDY TRANSCRIPTION MECHANISM

2.7.1 Expression microarrays for analysis gene expression

Microarray technology, first published in [84], has been widely used to explore the profile of gene mRNA expression patterns in the genome at once. Moreover, microarray technology has

enabled us to explore global features of hormone-regulated gene expression through nuclear receptors (NRs). Briefly, microarray technology is based on DNA hybridization process in which a DNA strand binds to its unique complementary strand. Briefly, DNA fragments each containing a nucleotide sequence that serve as a probe for a specific gene are immobilized in a specific surface. There exist two main types of microarrays in terms of their probe types that are immobilized in a predefined organization to a solid surface; (i) two channel cDNA (complementary DNA) arrays and (ii) single channel oligonucleotide microarrays. In cDNA microarrays the probes contain pre-synthesized sequences that are then placed on the array. These sequences can be hundreds of base pairs long. Oligonucleotide microarrays contain sequences that are directly synthesized onto the microarray. In the procedure for microarray experiment, by extracting and labeling mRNA and the hybridizing those purified mRNAs to the array, the amount of labeled sample specifically binding to each complementary probe/feature can be quantified at each probe location by measuring the fluorescent intensity. The overall process enables a genome-wide measurement of the expression level given the sample.

2.7.2 Measuring instantaneous transcriptional activity by GRO-seq: nuclear run-ons on a genomic scale

Microarray-based measurements of steady-state mRNA levels consider both RNA synthesis and degradation. However, they do not provide an accurate indication of ongoing transcription. Core et al. [85] developed global run-on sequencing (GRO-seq) to measure genome-wide instantaneous transcriptional activity. This new approach detects transcriptionally engaged Pol II and provides a “map” of the position and direction of engaged Pol II in the genome. The

application of GRO-seq helps us to measure direct transcriptional responses for a particular hormone-signaling pathway.

Briefly, nuclei are isolated, purified and nuclear run-on (NRO) is used to extend nascent RNAs associated with transcriptionally engaged Pol II under conditions inhibiting new initiation. Incorporation of the ribonucleotide analog (5-bromouridine 5'-triphospahate (BrUTP)) during the NRO step allowed for isolation of the newly transcribed nascent RNAs via immunoprecipitation using an antibody specific for this analog. Following immunoprecipitation, the NRO-RNA was reverse transcribed and amplified for sequencing. After sequencing, reads were mapped to the reference human genome.

2.7.3 ChIP based methods to determine protein-DNA interactions

Direct measurement of genome wide possible regulatory factor binding sites in vivo and at different cell states has recently become possible by Chromatin immunoprecipitation (ChIP) coupled with high-throughput techniques such as ChIP-chip, ChIP-Seq, and ChIP-PET. ChIP first described by Varshavsky and colleagues [86] is a process in which isolates fragments of sequences in the genome that is bound by a specific protein, most commonly a transcription factor. Briefly, ChIP works by covalently cross linking proteins to DNA typically by treating cells with formaldehyde or another chemical reagent. An antibody specific to the protein of interest can then be used to isolate the specific DNA fragments that the protein bound. The antibody and protein are then removed from the DNA. The three technologies then differ as to how they determine the location in the genome to which the DNA fragment corresponds. In the ChIP-chip method, the DNA fragments are hybridized to a microarray. In the ChIP-PET method,

the ends of the DNA fragment are sequenced. In the ChIP-Seq method, lots of short reads from within the DNA sequence are obtained. These methods only capture linear information of protein binding sites along the chromosomes but not interactions between them.

2.7.4 Methods for detecting long-range chromatin looping

Chromosome conformation capture (3C) [87] is developed to detect the frequency of interactions between any two DNA sequences at a time in cell populations. Briefly, formaldehyde crosslinking is used to fixed chromatin. Next, the cross-linked DNA is cut with a common restriction enzyme and ligated at diluted conditions to create a new junction between two pieces of DNA (i.e. between cross-linked fragments). The PCR is then used to detect these junctions. Over the past 10 years, many 3C-derivative methods have been developed to study long-range interactions genome-wide. The chromosome conformation capture-on-chip (4C) [88] and circular chromosome conformation capture (also known as 4C) [89] methods aim at revealing a complete pattern of DNA–DNA interactions for a DNA sequence of interest. The 5C (Chromosome Conformation Capture Carbon Copy) method concurrently probes, by pairs, interactions of hundreds of different sites under study [90]. The HiC method is a genome-wide version of 3C that ensures identification of all possible DNA–DNA interactions (the ‘interactome’) for a given cell population [91] . Whereas, ChiP-loop [92] and ChIA-PET [71] methods include an additional step of antibody precipitation targeting proteins potentially mediating interactions. Briefly, by immunoprecipitation of a factor of interest along with associated DNA fragments and followed by proximity ligation of distant DNA fragments tethered together within individual chromatin complexes, special arrangement of DNA with

respect to a protein complex can be revealed. These high-throughput methods provide data on different aspects of nuclear organization and they are complementary to each other. For example, ChIA-PET provides information for interactions between genomic elements that are in contact with specific protein, whereas Hi-C detects long-range interactions regardless of whether they Fbind to a specific protein. Therefore, integrating different data types might give us more complete picture.

3.0 THIRD CHAPTER: ESTROGEN REPRESSES GENE EXPRESSION THROUGH RECONFIGURING CHROMATIN STRUCTURES

3.1 BACKGROUND

Estrogen is essential for the development and function of the female reproductive system and is a known potent mitogen in breast cancer [93, 94]. The effects of estrogen are mediated through the alpha and beta estrogen receptors (ERα and ERβ), and almost an equal number of genes can be repressed or induced by estrogen-bound ERs [42]. While there is an extensive body of research studying ERα as a transcription activator (see the review articles [27, 95]), few studies concentrate on the mechanisms of ERα-mediated transcriptional repression (i.e. [31-34, 45]). Therefore, the mechanism by which estrogen-bound ERαs repress gene expression largely remains unclear. Since there are other transcription factors that have both inductive and repressive capabilities, a better understanding of the mechanism of ERα-mediated gene repression may shed light on general mechanisms by which a common transcription regulator exerts inductive and repressive influence on distinct genes.

The emergence and application of genome-wide high-throughput technologies enable studies inspecting multiple aspects of transcription processes at a whole genome scale. For instance, microarray technology has enabled one to study global impacts of estrogen on gene expression. Moreover, by measuring instantaneous transcriptional activity through global run-

ons sequencing (GRO-seq), one can study direct transcriptional responses for a particular hormone signalling pathway. In addition, ChIP-seq enables the accurate genome-wide profiling of transcription factors, co-regulators, RNA Pol II, and histone modification markers. Lastly, ChIA-PET (chromatin interaction analysis with paired-end tag sequencing) and other techniques [88-91] capture long-range chromatin looping on a genome scale. Indeed, there is already a large collection of publically available datasets utilizing above technologies to study cellular responses to estrogen treatment [50, 82, 96-100], thus affording us a new way of understanding how ERα regulates gene expression in a holistic manner.

In this study, we sought to investigate the mechanisms of estrogen-mediated transcription regulation by integrating public genome-scale datasets collected in the absence and presence of estrogen (see Supplementary Table 1 for the complete list of datasets). Through dissecting the diverse datasets from different angles, we aim to derive snapshots of the ERα-mediated transcription machinery, particularly higher-order chromatin complexes, and rend a holistic perspective of the regulation process. Our analyses have led to many novel findings that enhance our understanding of the function of estrogen as a transcription regulator. In particular, our study has led to a novel hypothesis regarding estrogen-mediated gene repression.

Table 1 Summary of datasets used in the study

Name	Usage	Factors	Datasets	Reference
Microarray	Identification of early E2-responsive genes		GSE3834, GSE9936, GSE11324, GSE5840	[42, 101-104]
GRO-seq	Instantaneous transcription activity		GSE27463	[96]
ChIP-seq	Whole-genome mapping of protein-DNA interactions	ERα, Pol II, FoxA1, AP2γ, PBX1, GATA3, CTCF, RAD21, STAG1, SRC-1, SRC-2, SRC-3, TRIM24, c-Fos, c-Jun, p300, CBP	GSE14664, GSE26831, GSE24166, GSE23893, GSE28007, GSE23852, GSE25710, GSE25021	[17, 18, 97-100, 105-108]
	Genome-wide histone modification patterns	H3K4me1, H3K4me2, H3K4me3, H3K9me3, H3K9ac, H3K14ac,	GSE23701, GSE24166	[97, 107]
ChIA-PET	3D chromosomal structure around a protein	ERα, Pol II	GSE33664, GSE39495	[70, 82]

3.2 MATERIALS AND METHODS

3.2.1 Identify consensus E2-responsive genes

The consensus estrogen-responsive genes were identified based on a ranked-product meta-analysis across 4 independent published datasets (GSE3834, GSE9936, GSE11324 and GSE5840 – Affymetrix GeneChip Human Genome U133 Plus 2.0 platform), which investigated the effect of estrogen treatment on gene expression in MCF-7 cells at early (3–4 h) time points [101]. We further filtered out genes with small gene-wise mean and standard deviation. We further selected genes that contain a single RefSeq TSS annotation to avoid determining which TSS was responsible for the transcription.

3.2.2 ChIP-seq, GRO-seq and ChIA-PET data sets and pre-processing

ChIA-PET data for MCF-7 cells data (pre-processed) were obtained from the original published supplementary data [70, 82, 96]. We merged results of IHM001F and IHH015F large scale ChIA-PET analysis [82] using supplement files of the original work [82]. Processed Pol II ChIA-PET data were obtained from authors of the original work [70]. Pre-processed ChIP-seq Pol II, TF, co-regulator and histone mark data for MCF-7 was downloaded from Nuclear Receptor Cistrome Database, [50] where the peaks were called by MACS with *P*-value cut-off 10^{-5}. Mapped GRO-seq reads (at 0 minute and 40 minute) were downloaded from GEO (GSE27463).

3.2.3 Consensus ERα cistrome

We collected a total of four ChIP-seq data sets for ERα [97-100] which profiled MCF-7 in the absence and presence of ligand. Since there was a large variation of ERα binding sites across different studies of MCF-7, we merged overlapping binding sites in at least two studies in order to form a consensus ERα cistrome using the completeMOTIFs pipeline[109]. This approach allowed us to combine the results of several studies, and provide a global picture of ERα binding sites.

3.2.4 Consensus FoxA1 cistrome

We collected three ChIP-seq datasets for FoxA1 [17, 97, 108] in the absence of ligand, as well as two ChIP-seq datasets for FoxA1 [17, 97] treated with estrogen, derived from the MCF-7 cell line. We merged overlapping binding sites in at least two studies in order to form a consensus FoxA1 cistrome in the absence of ligand using the completeMOTIFs pipeline[109]. We took overlapping binding sites (at least 1 bp) between two studies of ligand to form FoxA1 cistrome in the presence ligand.

3.2.5 Genome Annotations

Genome annotations were downloaded from the UCSC Genome Browser (www.ucsc.org), human genome Build 36 (hg18 assembly). Gene definitions were given by the RefSeq genes track. For the analysis mentioned in the paper, we have considered only those RefSeq genes

which have one annotated TSS. When visualizing experiments with the **UCSC** Genome Browser, we used human genome Build 37 (hg19 assembly).

3.2.6 Pol II ChIP-seq meta-gene profiles

To show the average Pol II ChIP-seq profiles across genes, a “metagene” profile [110] was plotted for each group. Genes were aligned at the first and last nucleotides of the annotated transcripts and sequencing tags were scaled as follows. The total sequence tag counts were directly used for the promoter (0.5 kb upstream of the TSS to 0.5 kb downstream) and the 3'-end (0.5 kb upstream of the TES to 0.5 kb downstream) of transcripts. To account for variable gene sizes, signal between 0.5 kb downstream of the TSS to 0.5 kb upstream of the gene end was represented by 1000 values obtained by cubic spline interpolation. Then the resulting tags of a gene were scaled to 100 equally sized bins (average tags in each bin) so that all genes appear to have the same length. All profiles were plotted on a normalized read per million (RPM) basis (by dividing the raw read count by the total number of mapped reads, and multiplying the result by 1,000,000).

3.2.7 GRO-seq meta-gene profiles

To show the average GRO-seq profiles across genes, a “metagene” profile [110] was plotted. Genes were aligned at the first and last nucleotides of the annotated transcripts and sequencing tags were scaled in the same fashion as discussed in the previous subsection. All profiles were plotted on a normalized read per million (RPM) basis.

3.2.8 miRNA analysis

The target prediction analysis was performed by using ComiR [111], a newly developed algorithm that is designed to predict the targets of a set of miRNAs. ComiR incorporates the miRNA expression level in the thermodynamic binding model and thus improves the prediction of existing algorithms. It then combines the improved predictions of four target prediction tools using a support vector machine trained on Drosophila Ago1 immunoprecipitation data. We used ComiR to compute the genes probabilities associated with each single miRNAs and we considered as targets the genes with a ComiR probability score greater than 0.8.

3.2.9 Statistical Analysis

Statistical significance of difference between gene groups was assessed using t-test and chi-squared test using R.

3.3 RESULTS

3.3.1 Identification of early E2-responsive genes by meta-analysis

We used the results from a recent meta-analysis of estrogen response in MCF-7 breast cancer cells [101], which identified a set of early estrogen-responsive genes. This consensus set of 967 genes was based on a ranked-product meta-analysis across four independent published datasets

investigating the effect of estrogen at early (3–4 hr) time points [101]. The 967 genes included 562 genes that are estrogen-activated and 405 that are estrogen-repressed. To avoid confounding overlapping signatures from multiple transcription start sites (TSSs), we further selected genes that have a single TSS according to the annotation from RefSeq [112]. The resulting set was a total of 748 genes, which correspond to 429 estrogen-induced and 319 estrogen-repressed genes (Additional file 1). The signals for 210 (less than half) estrogen-induced genes were categorized as "present" at the probe level in the absence of ligand, indicating that there were basal transcription activities for these genes, and the rest of the genes were categorized as "absent". As expected, the basal expression value of estrogen-repressed genes was higher than estrogen-induced genes in the absence of estrogen. While it is understood that there are additional estrogen-responsive genes in the human genome, this set of genes can be treated as representative estrogen-responsive genes to derive insights of estrogen-mediated transcription regulation.

3.3.2 Transcription states and Pol II chromatin complexes in the absence of ligand

3.3.2.1 Pol II occupancy at the promoters of estrogen responsive genes in the absence of ligand

Recruitment of the RNA polymerase (Pol II) transcription complex to promoters by specific DNA-binding proteins, e.g., TFs, is generally recognized as a key regulatory step in selective transcription in most eukaryotes. Therefore, Pol II is a good marker for transcriptionally active promoters. First, we used the existing Pol II ChIP-seq dataset derived from MCF-7 cells [98] to

investigate Pol II occupancy near the TSS of estrogen-responsive genes in absence of ligand. We performed meta-gene analysis (see Method section for details) across the promoter, the bodies of estrogen-responsive genes, and the transcription end site (TES), more specifically from -500 bp of TSS to +500 bp of TES, focusing on differences between induced and repressed genes. As seen from Figure 1-A, there were strong peaks of Pol II binding at the promoter regions of the meta-genes in the absence of ligand. There is no significant difference between estrogen-induced and estrogen-repressed genes in terms of Pol II occupancy (*t*-test $P = 0.7044$), measured as the normalized total sequence tag counts near TSS (± 500 bp). Although we observed a decrease of mean Pol II counts in the gene body for all genes, this decrease might be due to an experimental artifact. It is known that, as Pol II progresses further into gene, it becomes hyper-phosphorylated and thus a less suitable target for the antibody [85]. The results indicate that Pol II recruitment to promoters could not be the key factor that leads to the distinct transcriptional behaviours of the estrogen-induced and estrogen-repressed genes in the absence of estrogen.

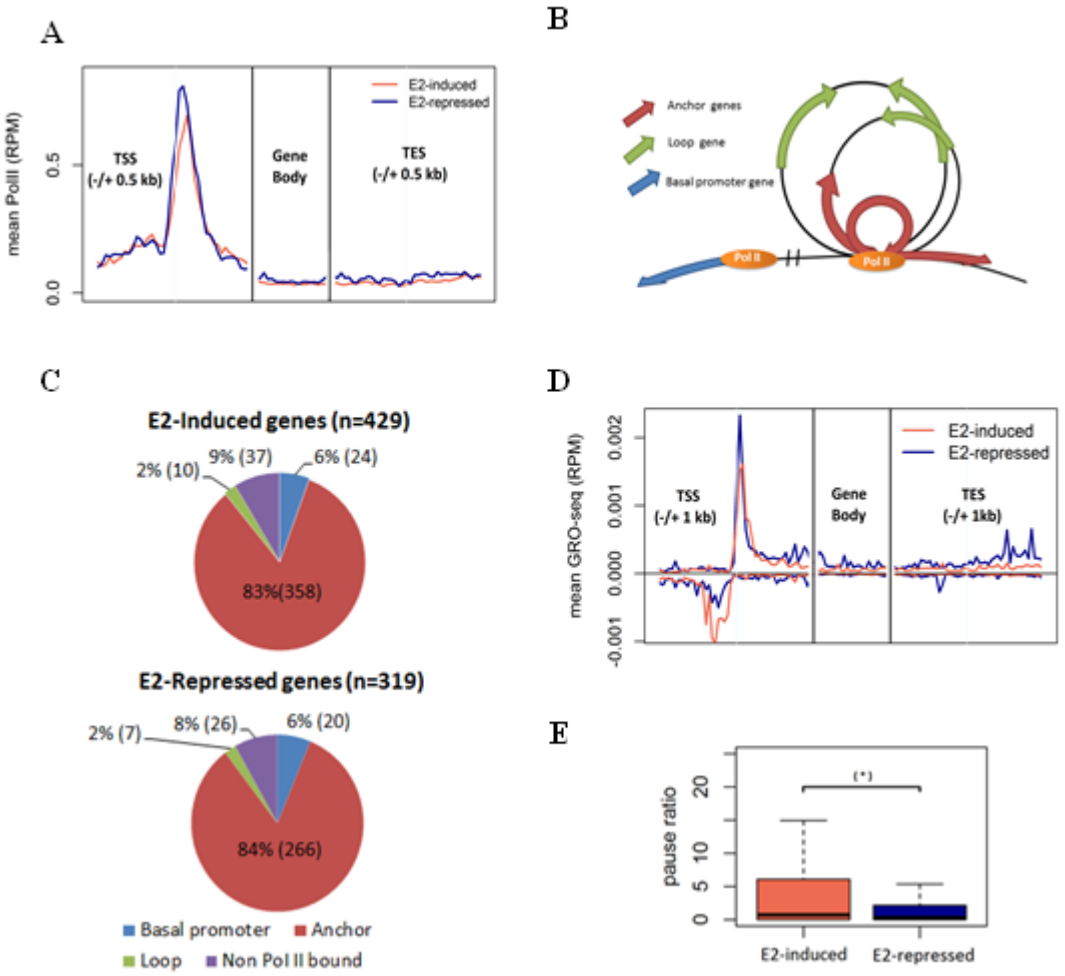

Figure 1 Transcription states and chromatin complexes in the absence of ligand

(A) Composite (meta-gene) profiles of Pol II ChIP-seq of E2-responsive genes, presented as reads per million (RPM). Profiles for promoter and 3` end were aligned at TSS and TES respectively; profiles for gene bodies were scaled. **(B)** Annotation of genes based on their relative position to the Pol chromatin complex. **(C)** Distribution of E2-responsive genes in terms of their relative position to Pol II complexes. **(D)** Composite (meta-gene) profiles of GRO-seq of Pol II bound anchor genes, presented as reads per million (RPM). Profiles for promoter and 3` end were aligned at TSS and TES respectively; profiles for gene bodies were scaled. GRO-seq reads aligned to RefSeq TSSs in both sense and antisense directions relative to the direction of gene. **(E)** Boxplots show the comparison of pause ratio (TSS/gene body) for E2-repressed genes (blue) and E2-induced genes (coral) as determined by GRO-seq in the absence ligand.

3.3.2.2 RNAPII-associated chromatin interactions in the absence of ligand

It is widely believed and supported by recent genome-scale studies [70, 82] that the formation of functional contacts between regulatory factors, e.g., TFs, and Pol II through local DNA looping introduces higher-order structures that directly impact gene expression regulation [70, 113]. The first 3D interaction map of Pol II occupied sites for five different cell-lines, including MCF-7, was recently established using ChIA-PET [70]. We used this Pol II ChIA-PET data to investigate the relationship between the spatial organization of estrogen-responsive genes and their transcription status. Briefly, ChIA-PET technology detects the arrangement of DNA with respect to a specific protein. For example, Li et al. used the anti-Pol-II antibody to immunoprecipitate Pol II complexes along with associated DNA fragments [70], followed by proximity ligation of DNA fragments tethered with the complexes to detect chromatin complexes [82]. In its simplest form, a higher-order chromatin complex includes a protein complex and tethered DNA that form a single loop. The DNA involved in such a complex can be further divided into *anchor regions* and *looping regions*, as defined by proximity of the TSS (±5 kb) to an estimated anchor point (the DNA region that directly contacts the protein complex), and by the TSS being located within a ±5 kb region flanking the ends of the complex-encompassing DNA, respectively [70]. It is not uncommon to observe multiple loops and anchor regions in a higher-order chromatin complex. Figure 1-B shows a diagram to illustrate how a gene can be categorized with respect to a Pol II-bound chromatin complex (for brevity, hereafter **referred to** as Pol II complex).

Out of 4,474 Pol II complexes, 563 complexes contained 641 (out of the total of 748, 86%) estrogen-responsive genes identified through our meta-analysis. These estrogen-responsive genes tend to disperse among different complexes, a result in agreement with a previous study

that showed that co-localization in the nucleus was not required for coordinated expression of estrogen-responsive genes [114]. Further investigation showed that the majority, 624 (out of the total of 641 inside complexes, 97%), of these genes was categorized as anchor genes. These anchor genes resided in 552 complexes. Also, 358 estrogen-induced genes (out of 429, 83%) and 266 estrogen-repressed genes (out 319, 84%) were categorized as 'anchor genes' (Figure 1-C). These results suggest that the majority of identified estrogen-responsive genes, both estrogen-induced and estrogen-repressed, were located in Pol II complexes within close vicinity to Pol II in the absence of ligand. In the rest of the analysis, we focused on these '*estrogen-responsive anchor gene*s'—for brevity, **ref**erred to as estrogen-responsive genes—to study the impact of formation or disruption of higher-order chromatin complexes on regulating expression of these genes. Taken together, these and the previous results, indicate that Pol II not only had high tendency to occupy the majority of the promoters of the estrogen-responsive genes, it also formed higher-order chromatin complexes in the absence of estrogen.

3.3.2.3 The transcription activity of Pol II-bound anchor genes in the absence of ligand

To investigate the transcription activities of estrogen-responsive Pol II-bound anchor genes, we analysed the results of a GRO-seq experiment [96] performed in MCF-7 breast cancer cells, which studied the instantaneous transcriptional activities in the absence and presence of estrogen. GRO-seq detects *de novo* transcription activities of genes, thus providing a "map" of the position and direction of transcription activities. First, we performed meta-gene analysis of the transcription activities from -1kb of TSS to +1kb of TES (Figure 1-D) to investigate whether their transcription activities differ between estrogen-induced and estrogen-repressed genes (see Method section for details). The Figure 1-D shows that there were GRO-seq peaks within close

vicinity of TSS along both sense and antisense strands for estrogen-induced and estrogen-repressed genes, indicating that transcription was actively going on for both groups. Using the mean GRO-seq reads near TSS (−300 to +300 bp) on the sense strand as a statistic of average transcription rate for each gene, we found that the transcription rates of the estrogen-induced and estrogen-repressed genes at their promoters were not **statistically significantly different** (*t*-test $P = 0.5704$). On the antisense strand at TSS, we observed significantly more reads for the estrogen-induced genes in comparison to the estrogen-repressed genes (*t*-test $P<10^{-3}$). Function of antisense transcription is unknown but their existence suggests an open structure of DNA permissive to transcription activity by Pol II [85, 115, 116].

We further compared the transcription activities within the gene bodies of the estrogen-responsive genes, and we noted that estrogen-repressed genes had more reads in the body region than estrogen-induced genes (*t*-test $P<10^{-3}$). This result indicated that, although the transcription activities at TSS were similar, the transcription within the gene body might have paused for the estrogen-induced genes. We calculated the pause ratio for each gene, i.e. the ratio of the total reads in the vicinity of a TSS (± 300 bp) over the total reads in the corresponding gene body, with both numbers normalized by the length of the regions. Then we compared the pause ratio between estrogen-induced and estrogen-repressed groups. As seen in the Figure 1-E, the pause ratio of estrogen-induced genes was significantly higher than estrogen-repressed genes in the absence of ligand (*t*-test $P<10^{-2}$). The results suggest that, while estrogen-induced genes were actively transcribed at the promoter region, such transcription activity failed to elongate into gene bodies. Overall, our results indicated that estrogen-repressed genes tend to actively transcribe before estrogen, whereas estrogen-induced genes were transcribed at the TSS but not in full length.

3.3.2.4 Characterization of ERα binding sites respect to Pol II complex regions in the absence of ligand

To further investigate the role of ERα recruitment in formation of the Pol II complexes, we pooled the data from four MCF-7 ChIP-seq studies [97-100] to derive a set of 18,212 consensus ERα binding sites (see Method section for details). These previous studies revealed that there were a significant number of ERα binding sites in the absence of ligand, and the median distance between an ERα binding site and its nearest gene was more than 10 kb away. To relate the ERα binding sites with respect the Pol II complexes, we investigated whether some observed ERα binding sites are located within or are close to the Pol II complex regions. Recall that our analysis in the previous sections has identified that a total of 552 Pol II complexes contained 624 estrogen-responsive anchor genes. Among these complexes, a total of 434 (79%) contained ERα binding sites, and the majority of them (413) had ERα binding within anchor regions. Table 1 shows the distribution of complexes containing estrogen-induced and estrogen-repressed genes and their relationship to ERα binding sites. We further performed Jaccard test [117] to assess whether the overlap of these ERα binding sites with Pol II complexes is beyond random chances, and the results indicate that ERα is significantly enriched in the Pol II complexes ($P < 0.01$). The results show that even in the absence of ligand, ERαs were actively involved in the majority of the Pol II complexes that contain estrogen-responsive genes, both estrogen-induced and estrogen-repressed ones. The results also indicate that, while most ERαs are linearly remote to their target genes, their involvement in Pol II complexes brought them close to the genes in the 3D space.

3.3.2.5 Characterization of pioneer transcription factor binding sites and histone marks in the absence of ligand

Pioneer factors are a special class of transcription factors that interact with compacted chromatin to facilitate the binding of additional factors [118]. Pioneer factors have been shown to recruit chromatin modifiers and generate a local environment that is more accessible for ERα binding and help rapid transcriptional response of ERα [19, 20]. We next investigated whether a well-known ERα-pioneer-factor, FoxA1, as well as other putative ERα-pioneer factors, including PBX1, AP2γ and GATA, are also enriched in the Pol II complexes to facilitate ERα binding in these regions. We identified the binding sites for these factors from the ChIP-seq data of these factors in MCF-7 breast cancer cells [17, 18, 97, 105, 108]. These datasets include 24,250 PBX1 [18], 30,976 AP2γ [17] and 20,704 GATA3 [105] binding sites without estrogen treatment. We also pooled 12,531 FoxA1 binding sites from three different studies [17, 97, 108] in order to identify consensus binding sites by merging overlapping binding sites in at least two studies (see Method section for details). A total of 2,120 FoxA1, 7,345 AP2γ, 3,223 GATA3 and 4,548 PBX1 binding sites overlapped with the 552 Pol II complexes containing estrogen-responsive genes (for the number of overlapping regions (at least 1 bp) between pioneer factor binding sites inside the Pol II complex region see Supplementary Figure 1). Their distributions in the complexes containing estrogen-induced and estrogen-repressed genes are also shown in Table 1. Jaccard test [117] showed that these binding sites were significantly enriched ($P < 0.01$) in the anchor region of these complexes. We further looked for overlap between ERα binding events and pioneer factor binding events inside the Pol II complex in the absence of ligand. A majority of ERα binding events (3045 out of 3950, 77%) inside Pol II complexes containing estrogen-responsive genes overlapped with at least one pioneer factor binding event in the

absence of ligand (see Supplementary Figure-2). The results indicate that these factors were providing the foundation for ERα-DNA interactions inside the complexes, which their co-occurrence with ERα binding sites indicate that the latters are likely to be functional.

To determine other chromatin features associated with Pol II complexes, specifically promoter regions of estrogen-reponsive genes, we examined the relationship of the genome-wide distributions of histone marks and TSS of estrogen-responsive genes (± 1kb) using the data from MCF-7 breast cancer cells [97, 107]. We found that the histone marks indicating active transcription (H3K4me1, H3K4me3, H3K9ac and H3K14ac) were spatially correlated with promoters of estrogen-induced and estrogen-repressed genes; the marks lie closer than expected, in terms of genomic distance, to these genes. Interestingly, H3K9me3 and H3K27me3, which are typically found at inactive or closed chromatin, were not spatially correlated with promoters of neither estrogen-induced nor estrogen-repressed genes. To gauge significance, we used the GenometriCorr package [117], which uses permutation to create a distribution of genomic distances that would be expected if the marks were uncorrelated. Since the distances that we observed lie outside this distribution, we concluded that the active histone marks were indeed spatially correlated with the promoters of the estrogen-induced and estrogen-repressed genes. (Supplementary Table 2 shows the result of the correlation test between promoter regions and histone marks). Overall, these results indicate that in the absence of the ligand, most estrogen-responsive genes assumed a higher-order chromatin configuration that involved Pol II, ERα, ERα-pioneer factors and active histone marks.

Table 2 The distribution of Pol II complexes where E2-induced and E2-repressed genes reside and their relationship to ERα and pioneer factor binding sites inside the anchor region of complexes

Number	Complexes containing E2-induced genes (%)	Complexes containing E2-repressed genes (%)
All Pol II complex	344	238
ERα	248 (72.1%)	194 (81.5%)
FoxA1	199 (57.8%)	187 (78.6%)
AP2γ	286 (83.1%)	222 (93.3%)
PBX1	271 (78.8%)	211 (88.7%)
GATA3	235 (68.3%)	193 (81.1%)

Table 3 Summary of association between histone marks and E2-responsive gene promoters

	Correlation			
	E2-		E2+	
Histone marks	E2-induced	E2-repressed	E2-induced	E2-repressed
H3K9ac	0.60**	0.50**	0.65**	0.51**
H3K14ac	0.46**	0.40**	0.53**	0.48**
H3K4me1	0.15**	0.10**	0.23**	0.21**
H3K4me2	-0.06*	0.02	-0.02	-0.01
H3K4me3	0.75**	0.64**	0.68**	0.65**
H3K9me3	-0.01	-0.01	-0.04	0.00
H3K27me3	-0.03	0.03	0.00	0.00

P-value < 0.002 **, 0.05<P-value ≤0.002 *

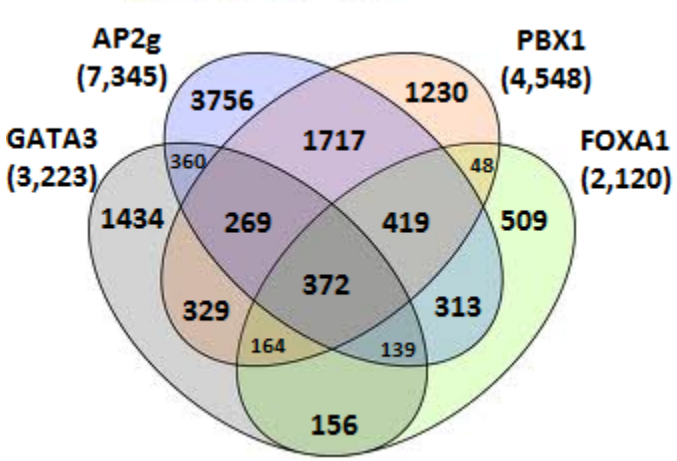

Figure 2 Number of overlapping pioneer factor binding sites inside Pol II complexes in the absence of ligand

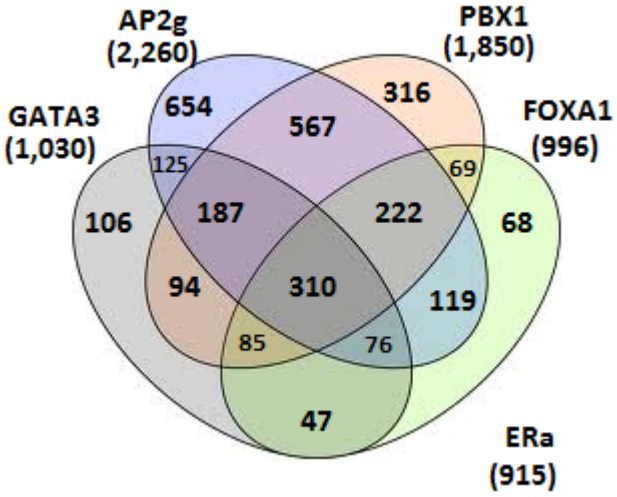

Figure 3 Number of overlapping pioneer factor binding sites with ERα inside Pol II complexes in the absence of ligand

3.3.3 Impact of estrogen treatment on transcription activity of E2-responsive genes

To understand the effect of estrogen on the transcription status and the higher-order chromatin structure of Pol II-bound anchor genes, we also studied Pol II ChIP-seq, GRO-seq, ERα ChIP-seq, TF/co-regulator ChIP-seq, histone mark/variant ChIP-seq, and ChIA-PET derived using ERα antibody in the presence of estrogen treatment. However, due to the lack Pol II ChIA-PET data after estrogen treatment, we try to infer the states of original Pol II complexes based on integrative analysis of the results from the above data types, particularly the ERα ChIA-PET data.

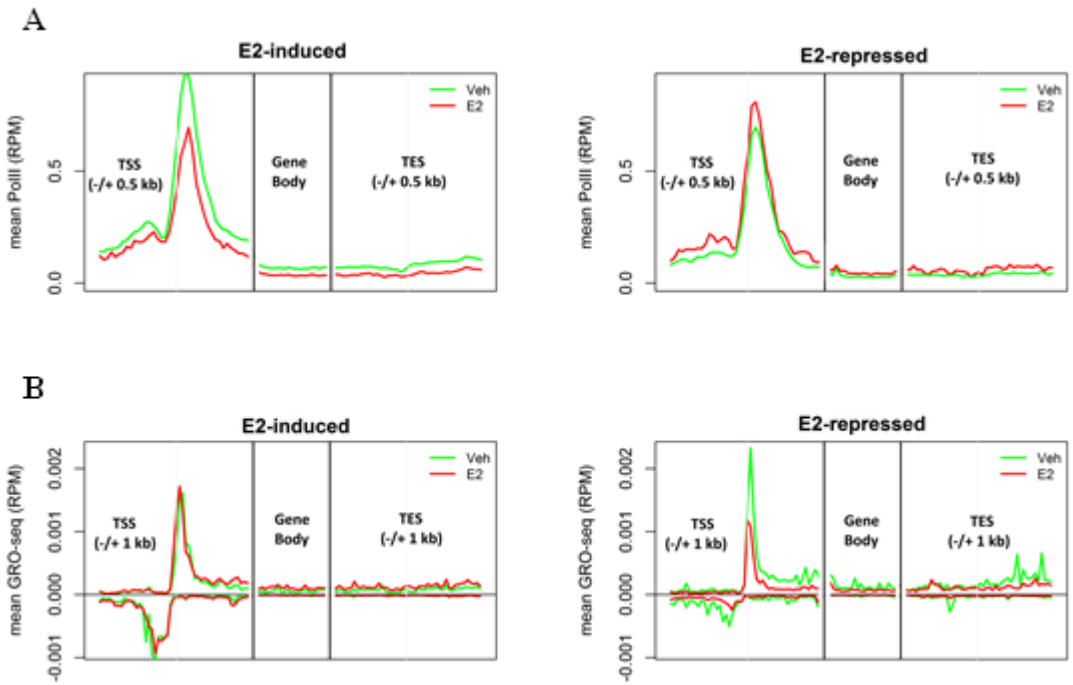

Figure 4 Comparison of transcription states of E2-induced and E2-repressed genes.

(A) Comparison of meta-gene profiles of Pol II ChIP-seq of E2-induced and E2-repressed genes, presented as reads per million (RPM) aligned sequences per gene per nucleotide (density) versus position relative to the TSS in the absence (green) and in the presence of ligand (red). **(B)**

Comparison of *de novo* transcription of E2-induced and E2-repressed genes in the absence (green) and in the presence of ligand (red) determined by GRO-seq.

3.3.3.1 The transcription activity of Pol II-bound anchor genes in the presence of ligand

After estrogen treatment, Pol II occupancy measured by anti-Pol II ChIP-seq [98] at the promoter (± 500 bp near TSS) showed a statistically significant increase (paired *t*-test $P<10^{-19}$) and decrease (paired *t*-test $P<10^{-10}$) for the estrogen-induced and estrogen-repressed genes respectively. While paired *t*-tests clearly detect the trends of changes, the meta-gene analysis showed only a 38.8% increase and 21% decrease of the peak areas for the estrogen-induced and estrogen-repressed genes respectively (Figure 2-A). The change in the number of Pol II tags at the promoter region reflects the affinity of Pol II DNA interaction at the site, but does not necessarily directly entail changes in the transcription rate.

We further studied the impact of estrogen treatment on *de novo* transcript rate using the GRO-seq data[96], and the results are shown in Figure 2-B. For the estrogen-induced genes, the transcription rate at promoter regions did not change significantly (paired *t*-test $P = 0.16$), but the transcription rate in the gene body (sense strand) increased significantly (paired *t*-test $P<10^{-19}$), and the pause ratio was significantly decreased (paired *t*-test $P = 0.0015$). We did not observe changes in the antisense transcription in the absence and presence of ligand (paired *t*-test $P = 0.85$) for estrogen-induced genes. The results indicate that estrogen mainly acts to enhance the elongation of transcription of these genes.

For the estrogen-repressed genes, the transcription level decreased in both the promoter (paired *t*-test $P = 0.012$) and gene body regions (paired *t*-test $P<10^{-6}$). The pause ratio of estrogen-repressed genes between the conditions in the absence and presence of ligand was not

statistically different (paired *t*-test $P = 0.21$). Antisense transcription significantly decreased in the presence ligand for estrogen-repressed genes (paired *t*-test $P<10^{-5}$). The results suggest that estrogen treatment suppressed transcription initialization and elongation of these genes in a concordant manner.

3.3.3.2 Characterization of histone marks with respect to promoters of E2-responsive genes in the presence of ligand

We analyzed the impact of estrogen treatment on histone modification at the promoters of estrogen-responsive genes [97] to investigate if histone modifications mediate the effect of estrogen on these genes. We found that the histone marks indicating active transcription (H3K4me1, H3K4me3, H3K9ac and H3K14ac) were spatially correlated with promoters estrogen-induced and estrogen-repressed genes; the marks lie closer than expected, in terms of genomic distance, to these genes. Interestingly, they were not spatially correlated for either H3K4me2 (typically found in the promoter or gene body) or for H3K9me3 and H3K27me3 (typically found at inactive or closed chromatin). The results are shown in Supplementary Table 2. To gauge significance, we used the GenometriCorr package [30], which uses permutation to create a distribution of genomic distances that would be expected if the marks were uncorrelated. Since the distances that we observed lie outside this distribution, we concluded that the active histone marks were indeed spatially correlated with the promoters of the estrogen-induced and estrogen-repressed genes in the presence of estrogen. The results indicate that histone modifications at the promoters of estrogen-responsive genes rendered chromatin accessible to DNA binding factors, and the lack of repressive histone markers indicate that estrogen-mediated gene repression is not through histone modification.

3.3.4 High-order configuration changes upon estrogen treatment

We identified 30,115 consensus ERα binding sites after estrogen treatment from the four ChIP-seq studies [97-100]. We first investigated if ERα binding sites tend to be enriched in the original Pol II complex regions after estrogen treatment as they did in the absence of estrogen. The results showed a total of 6,090 ERα binding sites overlapped with the locations of the Pol II complexes (552) containing estrogen-responsive genes, among which 3,035 were new binding sites when compared to ERα binding data without estrogen treatment. Jaccard test [117] indicate that the ERα binding sites are significantly enriched ($P<0.01$) in the region of these complexes, indicating that the regions of original Pol II complexes containing estrogen-responsive genes were “hot” regions of ERα binding. The presence of these new ERα binding events potentially would change the overall configuration of chromatin complexes that originally existed before estrogen presence, particularly if a new ERα binding event leads to new chromatin complex formation. Therefore, we turned to study the higher-order chromatin structures formed after estrogen treatment.

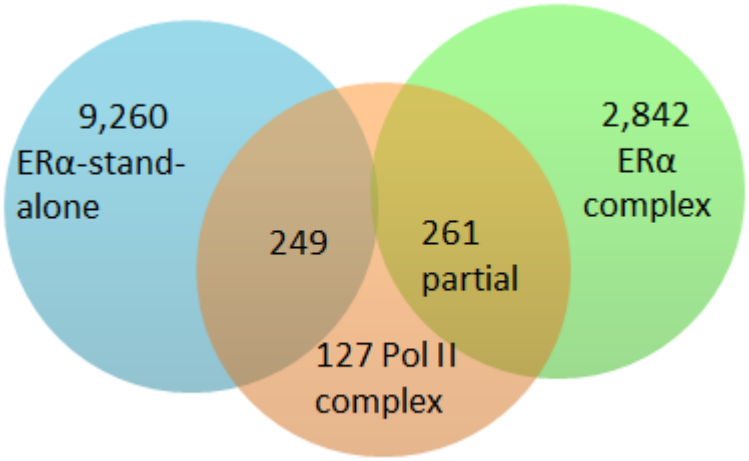

Figure 5 Venn diagram illustrating the overlap between the Pol II complexes containing E2-responsive genes formed in the absence of ligand and the ERα complexes formed in the presence of ligand

3.3.4.1 ERα binding sites tend to be enriched in the original Pol II complex regions

We identified 30,115 consensus ERα binding sites after estrogen treatment from the four ChIP-seq studies [97-100]. We first investigated if ERα binding sites tend to be enriched in the original Pol II complex regions after estrogen treatment as they did in the absence of estrogen. The results showed a total of 6,090 ERα binding sites overlapped with the locations of the Pol II complexes (552) containing estrogen-responsive genes, among which 3,035 were new binding sites when compared to ERα binding data without estrogen treatment. Jaccard test [117] indicate that the ERα binding sites are significantly enriched (P<0.01) in the region of these complexes, indicating that the regions of original Pol II complexes containing estrogen-responsive genes were “hot” regions of ERα binding. The presence of these new ERα binding events potentially would change the overall configuration of chromatin complexes that originally existed before estrogen presence, particularly if a new ERα binding event leads to new chromatin complex

formation. Therefore, we turned to study the higher-order chromatin structures formed after estrogen treatment.

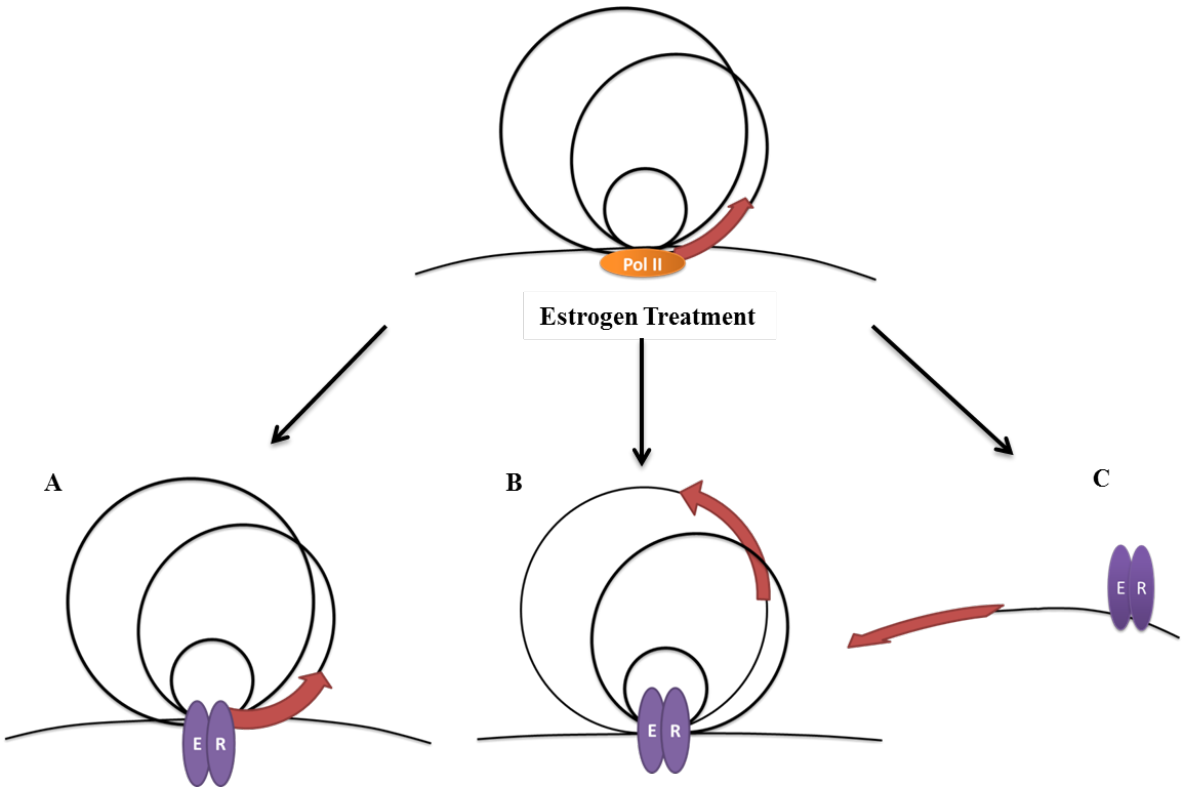

Figure 6 The position transition patterns that E2-responsive genes with respect to Pol II and ERα complexes

(A) anchor-to-anchor, (B) anchor-to-loop, (C) anchor-to-stand-alone. Green arrow shows the gene.

3.3.4.2 ERα-associated chromatin complex in the presence of ligand

We analyzed the published ChIA-PET data from a study of ERα-bound chromatin complexes [82] in MFC-7 cells treated with estrogen. This study reported ERα binding events that led to the formation of higher-order chromatin complexes (4,293), as well as those that did not form

complexes, which were referred to as stand-alone ERα binding sites (10,729). We then investigated the relationship of these ERα complex events with respect to original Pol II complexes containing estrogen-responsive genes, and the results are shown in Figure 3. There is a significant overlap between the DNA regions encompassing the original Pol II complexes and those encompassing ERα complexes or stand-alone ERα binding sites; 261 of the original Pol II complexes partially overlapped (at least 3,000 bp) with ERα complexes and 249 of them overlapped with stand-alone ERα binding sites. Some Pol II complexes with DNA loops occupying large genomic regions overlapped with more than one ERα complex, e.g., with one ERα complex region at one end as well as with a stand-alone ERα binding sites at the other end. Our analyses showed that a total of 358 out of 624 estrogen-responsive genes were found to be associated with ERα complexes after estrogen treatment [82]. For the remaining estrogen-responsive genes, the lack of an ERα involvement in the ChIA-PET data [82] may be attributed to the following factors: 1) the sensitivity of the ChIA-PET experiment, 2) the difference in experimental conditions and procedures, 3) low affinity secondary binding of ERα to chromatin [29, 30].

Since the majority of estrogen-responsive genes were involved in Pol II complexes and with ERα complexes before and after estrogen treatment respectively, we focused on the transcriptional behaviour of these estrogen-responsive genes from a perspective of the interplay between Pol II and ERα chromatin complexes. We categorized the estrogen-responsive genes based on their relationship with respect to Pol II and ERα complexes into 3 types, as shown in Figure 4. More specifically, they are: 1) *anchor-to-anchor*: a gene that was within a Pol II complex anchor region that was also an anchor gene with respect to an ERα complex (Figure 4-A). Since the position of chromatin complexes precipitated by anti-Pol II and anti-ER antibodies

overlapped, a possible interpretation is that both ERα and Pol II are involved in a common complex before and after estrogen treatment. 2) *anchor-to-loop*: an anchor gene with respect to a Pol II complex that became a loop gene with respect to an ERα complex (Figure 4-B). A possible interpretation would be that a new ERα complex is formed which likely has disrupted the original Pol II complex, because it is unlikely to have two distinct chromatin complexes encompass each other within a relatively short range of chromatin. 3) *anchor-to-stand-alone*: an anchor gene with respect to a Pol II complex that became a stand-alone ERα gene (on the basis that gene promoters were within ±20 kb of non-interacting ERα binding sites [82]) (Figure 4-C). A possible interpretation is that the original Pol II complex is disrupted, because the current ERα binding sites overlaps with the original Pol II complex site and yet no complex involving ERα is found.

In Figure 5, we render the relative positions and the relationships of the Pol II and ERα complexes of 3 genes to support the above reasoning. Figure 5-A shows an example of an *anchor-to-anchor* gene, *MYB*, where the Pol II and ERα complex anchored at the same regions, and the ERα binding sites overlapped with the anchor regions in the absence and presence of estrogen, indicating the anti-Pol-II and anti-ERα antibodies had pulled down a common complex. An example of an *anchor-to-loop* gene, *CALM1*, is shown in Figure 5-B. Before estrogen treatment, *CALM1* has an ERα binding event in the Pol II anchor region. When treated with estrogen, an ERα complex formed that included the DNA regions that subsumed the original Pol II complex, and the original ERα binding site in the Pol II complex disappeared. Therefore, it is reasonable to assume that the original Pol II chromatin complex was disrupted. Finally, an example of an *anchor-to-stand-alone* gene, *TLE1*, is shown in Figure 5-C. After estrogen treatment, there were distinct ERα binding patterns, and importantly the ERα binding

sites overlapping with the Pol II complex anchor is not associated with an ERα complex, indicating the original Pol II complex was disrupted. These types of reasoning enabled us to infer the impact of estrogen treatment on the Pol II complexes that contain estrogen-responsive genes prior to estrogen treatment, in other words, what happened to the Pol II complex after estrogen treatment.

We first compared the pattern of positional transitions of the estrogen-induced and estrogen-repressed genes respectively to investigate whether the two types of genes behaved differently (Figure 6-A). The distribution of estrogen-induced and estrogen-repressed genes in terms of their positional transition is significantly different (Chi-square test $P<10^{-11}$). Based on the distribution of genes among the categories, we inferred that the original Pol II complex containing 86% (124/144) estrogen-repressed genes were disrupted after estrogen treatment. In comparison, the original Pol II complexes containing 49% (104/213) estrogen-induced genes were inferred to be disrupted. The results also show that a significantly smaller number of estrogen-repressed genes are located in the anchor region of ERα complexes in comparison to the induced ones.

We further studied a subset of estrogen-responsive genes that had ERα in the anchor region of the original Pol II complexes in the absence of ligand (Figure 6-B) and analysed the pattern of their relative positional transitions. For this subset of genes, ERα might participate in the Pol II complex in the absence of ligand, and we are interested in inferring the impact of estrogen binding to these ERαs on the Pol II complexes. Among 113 estrogen-induced genes, 74 (65%) were categorized as *anchor-to-anchor* genes; 10 (9%) were categorized as *anchor-to-loop* genes; and 29 (26%) were categorized as *anchor-to-stand-alone* genes. Whereas from 68 estrogen-repressed genes, 13 (19%) were categorized as *anchor-to-anchor genes*; 20 (29%) were

categorized as *anchor-to-loop* genes; and 35 (52%) were categorized as *anchor-to-stand-alone* genes. The proportions of *anchor-to-anchor* genes for estrogen-induced and estrogen-repressed gene were significantly different (Chi-square test $P < 10^{-6}$). The data indicate that, with a high likelihood, the original complexes containing the majority (55, 81%) of the estrogen-repressed genes were disrupted due to the formation of new ERα complexes or ERα binding events. On the other hand, the data indicate that the original Pol II complex containing the majority (74, 65%) of estrogen-induced genes might have remained, since these genes were categorized as *anchor-to-anchor* with respect to both original Pol II complexes and ERα complexes.

In summary, a large proportion of estrogen-repressed genes assumed a higher-order chromatin configuration in the absence of ligand, which was later disrupted after estrogen treatment. The results lead to the hypothesis that original higher-order Pol II complexes provide ideal transcription environment for these genes. When treated with estrogen, binding of estrogen to the ERαs with these complexes or formation of new ERα chromatin complexes disrupted the original transcription active chromatin structures, thus leading to the repression of these genes.

Figure 7 Examples of positional transition of Pol II and ERα ChIA-PET complexes

The figure shows the positions and relationships of genes (RefSeq), Pol II complex anchor regions (Pol II Int), ERα complex anchor regions (ERα Int), ERα ChIP-seq binding sites in the absence and presence of estrogen for three genes: A) *anchor-to-anchor*, gene: MYB; B) *anchor-to-loop*, gene: CALM1; C) *anchor-to-stand-alone*, gene: TLE1. The estrogen treatment conditions are color-coded with green (absence) and red (presence). The black arrow indicates Pol II anchor region in the absence of ligand.

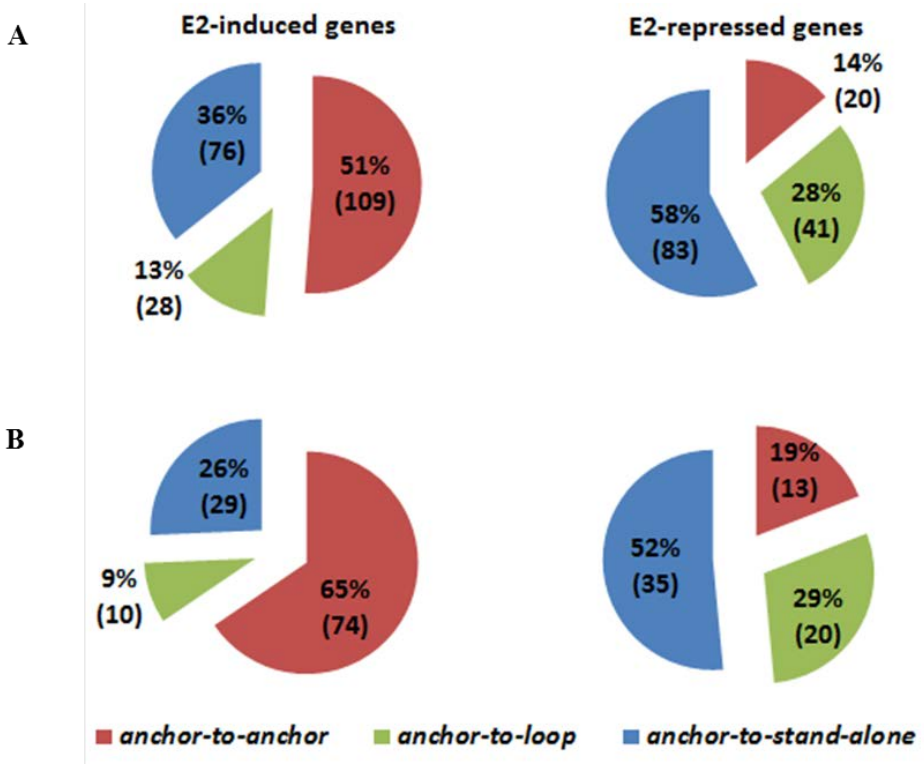

Figure 8 The distribution related to the positional transition patterns of the E2-induced and E2-repressed genes

(A) Number of gene in each positional transitional pattern group (*anchor-to-anchor*, *anchor-to-loop*, *anchor-to-stand-alone*) for E2-induced and E2-repressed genes. (B) Positional transition pattern distribution of E2-responsive genes that had ERα binding in the anchor region of the original Pol II complexes in the absence of ligand.

3.3.4.3 The transcription activity of Pol II complex associated genes in the presence of ligand

In order to investigate whether the pattern of positional transition of the estrogen-responsive genes (induced vs repressed) had any effect on the transcription activity of genes after estrogen treatment, we further compared the transcription activity of genes in each group using the

GRO-seq dataset[96]. The transcription activities of estrogen-induced genes were not statistically significantly different from each other among the *anchor-to-anchor, anchor-to-loop* and *anchor-to-stand-alone* genes. As expected, the transcription activities of all estrogen-induced genes, independent of their grouping, were increased after estrogen treatment when compared to those in the absence of ligand (paired *t*-test for *anchor-to-anchor* genes: $P<10^{-12}$, *anchor-to-loop* genes: $P = 0.0002$, *anchor-to-stand-alone* genes: $P = 0.0007$). The result indicates that, for estrogen-induced genes, position transition is not a significant factor influencing their transcription.

Interestingly, in the presence of estrogen, the transcription activity of estrogen-repressed *anchor-to-anchor* genes was significantly higher (*t*-test $P = 0.04$) than the pooled *anchor-to-loop* genes and *anchor-to-stand-alone* genes, aka, genes with disrupted original Pol II complexes. We further noted that, while the transcription activities of *anchor-to-loop* and *anchor-to-stand-alone* genes were decreased after estrogen treatment (paired *t*-test for *anchor-to-loop*: $P = 0.000614$, *anchor-to-stand-alone*: $P = 0.0024$), the transcription activity of estrogen-repressed *anchor-to-anchor* genes was not significantly changed (paired *t*-test for *anchor-to-anchor*: $P = 0.24$). In a further comparison between estrogen-induced and estrogen-repressed *anchor-to-anchor* genes, the transcription activities were not statistically significantly different (*t*-test $P = 0.45$). The results indicate that, for the estrogen-repressed genes, whether retaining or disrupting the original Pol II complexes has a significant impact on the transcription activities of these genes.

The discrepancies between the *de novo* transcription activity and steady-state mRNA levels of the estrogen-repressed *anchor-to-anchor* genes might be due to active posttranscriptional regulations of specific transcripts by mechanisms such as microRNAs (miRNAs). Several genome-wide profiling studies have characterized many estrogen-dependent

miRNAs, whose transcription is induced by estrogen in breast cancer cell lines [119, 120]. These miRNAs, including miR-22, let-7, miR-221/222, miR-18a/19b/20b and miR17-5p, were shown to negatively modulate the ERα-regulated genes after estrogen stimulation. We observed that out of 20 estrogen-repressed *anchor-to-anchor* genes, 13 of them were targets of these miRNAs. This result may explain the discrepancy between the observed high transcription activities in GRO-seq and the decreased mRNA levels in the microarray experiments for these genes.

3.3.4.4 Other TF/co-regulators associated with transition groups

We further examined the distribution of a set of well-known ERα-partner TFs and co-regulator binding sites derived from ChIP-seq experiments [17, 97, 99, 100, 105-107] from MCF-7 breast cancer cells treated estrogen; the goal is to inspect if distributions of these factors are specific to certain transition groups. After collecting publically available data for 14 different factors, we examined their enrichment within ± 20 kb of estrogen-responsive genes' TSSs, including transcription factors: FOXA1, AP2γ, GATA3, CTCF, STAG1, RAD21, cJun, cFos, and co-regulators; CBP, p300, SRC1, SRC2, SRC3, TRIM24. Significant differences between the number of genes having one particular factor and not having another particular factor in each of the six groups were obtained using Chi-square tests. *P*-values were reported with Bonferroni correction. Significant residual values (>2 or -2<) were shown inside the parenthesis. The results are shown Supplementary Table 3.

Interestingly, SRC-1, SRC-2, SRC-3 and FoxA1 binding events were enriched in *the anchor-to-anchor* group among the estrogen-induced genes. This observation agrees with the knowledge that SRCs function as co-activators for nuclear receptors. For the other factors, we did not find a significant association with respect to the sub-groups. Currently, we do not have

genome-wide co-repressor binding data. Potentially, genome-wide co-repressor data (such as NCoR, SMRT, NRIP1, LCoR and REA) can help to understand ligand-dependent transcriptional repression by ERα.

Table 4 The number of genes having TF and co-regulators binding sites within ±20 kb from their TSSs in each transition group

The table shows the distribution of binding sites among the 6 gene groups. Statistical significance was determined using chi-square tests. *P*-values were reported with Bonferroni correction. A number that is significantly deviated from the expected value (residual values greater than +2 and less than -2) are indicated with '+' and '-' respectively.

	E2-induced genes			E2-repressed genes			
TF/ co-regulator	anchor-to-anchor (n=109)	anchor-to-loop (n=28)	anchor-to-SA (n=78)	anchor-to-anchor (n=20)	anchor-to-loop (n=41)	anchor-to-SA (n=83)	Adj. *P*
SRC-1	45 (+)	0 (-)	5	1	0 (-)	0 (-)	$<10^{-17}$
P300	93	6 (-)	52	19	25	68	$<10^{-8}$
SRC-3	69 (+)	2 (-)	21	8	7 (-)	24	$<10^{-7}$
SRC-2	74 (+)	5 (-)	29	14	9 (-)	26	$<10^{-7}$
CBP	92	8 (-)	51	18	23	60	$<10^{-5}$
FoxA1	65 (+)	4 (-)	18 (-)	9	12	27	$<10^{-4}$
AP2γ	94	15	55	17	36	67	0.025
TRIM24	75	16	36	11	31	61	0.058
CTCF	89	16	53	18	36	59	0.126
Fos	42	5	17	9	9	26	0.649
RAD21	100	20	63	19	37	70	0.731
GATA3	44	7	27	11	17	37	1.000
STAG1	100	22	69	20	38	72	1.000
c-Jun	19	5	11	5	4	10	1.000

3.4 DISCUSSION

In this study, we present an integrated data analysis to derive a holistic perspective of the transcription machineries at estrogen-responsive genes, as shown in Figure 7. Systematic dissection of this data collection enabled us to ask specific questions, make inferences, and reveal different mechanisms of estrogen-mediated transcription regulation.

In the absence of the ligand, most of estrogen-responsive genes assumed a higher-order chromatin configuration that involved Pol II, ERα, ERα-pioneer factors and active histone modifications. This leads to the hypothesis that estrogen is not critical for assembling of these transcription machineries at these genes but rather (is needed?) for regulating the states of these machineries. Without the ligand, estrogen-induced genes showed active transcription at promoters but failed to elongate into gene bodies. The observed transcription pause at these sites may be due to: 1) the estrogen-free ERs in these complexes recruited co-repressors, e.g., N-CoR or SRMT [40], which prevented transcription from progressing into gene bodies; alternatively, the estrogen-free ERαs failed to recruit co-activators, e.g., SRC-1 [40], to the transcription sites.

After estrogen treatment, a large proportion of the Pol II chromatin complexes were disrupted. The results lead to the hypothesis that original higher-order Pol II complexes provided an ideal transcription environment for these genes, and disruption of these structures impaired their transcription, thus making them estrogen-repressed genes. Disruption of these complexes may be due to: 1) confirmation changes of the estrogen-bound ERαs in these complexes reduced affinity of the complexes to DNA; 2) new high affinity binding events or formation of novel complexes by estrogen-bound ERαs exerted physical torque to the original complexes, leading to their disruption; or 3) a combination of the above mechanisms.

Interestingly, estrogen-repressed *anchor-to-anchor* genes, of which original Pol II complexes were likely retained, had higher transcription activities in the presence of estrogen, indicating that the binding of estrogen to an ERα does not necessary encode repression action, an observation that corroborates the aforementioned hypothesis. In these *anchor-to-anchor* genes, the repressed expression levels were likely due to posttranscriptional regulation by transcript-specific mechanism such as microRNAs.

In comparison, the majority of estrogen-induced genes assumed an *anchor-to-anchor* pattern after estrogen treatment, and therefore retained active chromatin state for transcription at the TSS. Estrogen treatment likely facilitated transcription elongation by recruiting co-activators or releasing co-repressors. Observed enrichment of co-activators such as SRC-1, SRC-2 and SRC-3 in the promoter of these genes support this notion, and it will be interesting to further investigate if known ER co-repressors, such as N-CoR or SMRT, behave as expected. For the estrogen-induced genes that underwent chromatin structure transition, the formation of new ERα complexes might have facilitated recruiting of co-activators to enhance the transcription of these genes.

Beyond providing insights for understanding the mechanisms of estrogen-mediated gene regulation, our study leads to a general model for gene repression: any DNA-binding protein that is capable of disrupting a transcription-favoring chromatin complex can function as a transcriptional repressor. Disruption of a transcription-favoring complex can be simply achieved when the factor binds to DNA adjacent to the original complex, or forms a novel chromatin complex with a sufficiently high affinity to exert physical torque on the chromatin and disrupt the original complex. This model expands the concept of "transcriptional repressors" to include proteins that do not necessarily have any "inactivation" domain. In the same vein, any protein

that cooperates with such a factor in the process can be thought of as a “co-repressor”. This hypothesis remains to be further tested for other transcription factors, particularly other nuclear receptors such as RARA and RARG, which tend to exhibit dualism in regulating gene expression.

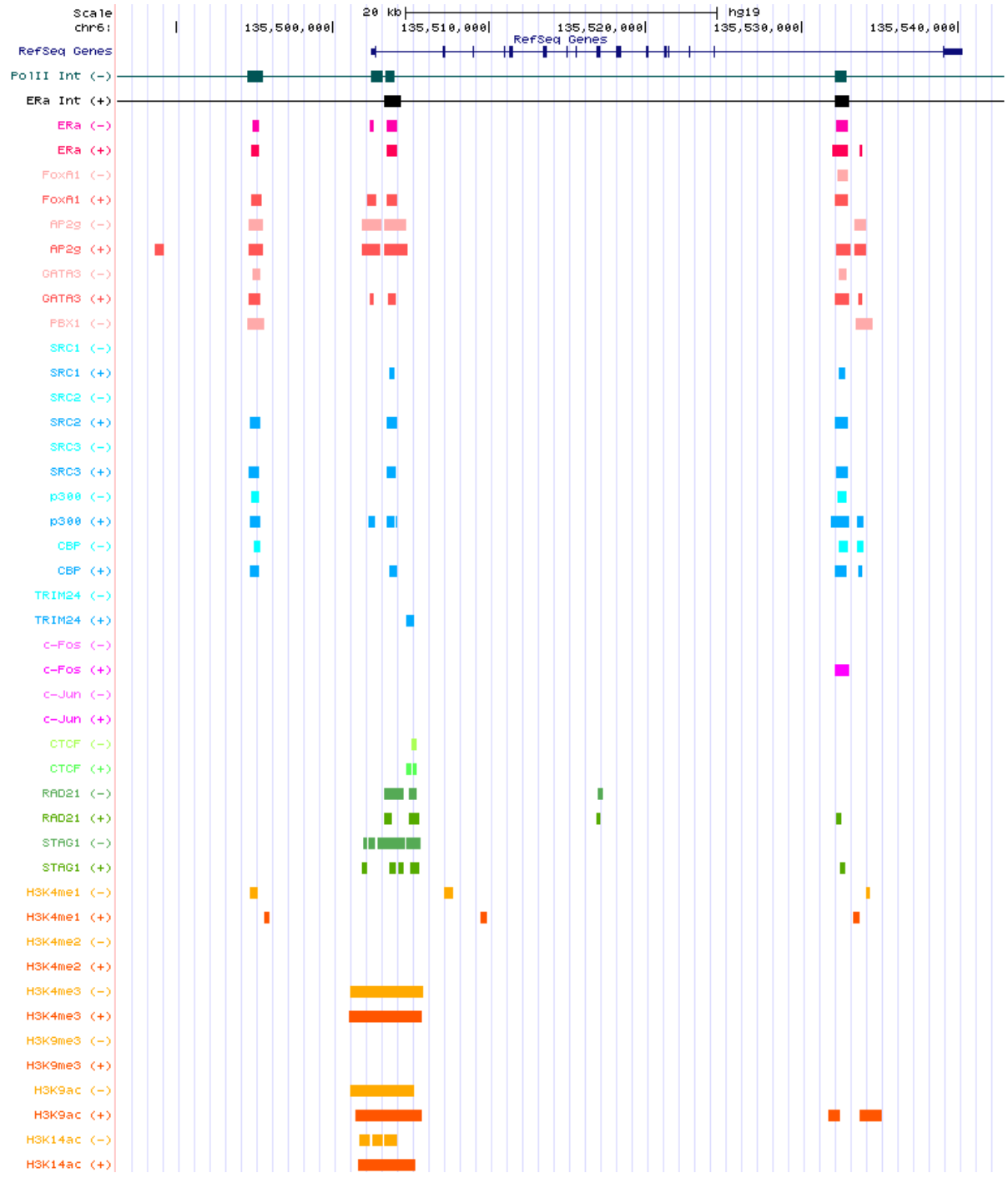

Figure 9 Pol II and ERα ChIA-PET interactions data and ChIP-seq binding data in the vicinity of *MYB* gene in the absence (-) and presence (+) of estrogen

4.0 FOURTH CHAPTER: IMPROVING CHIP-SEQ PEAK-CALLING FOR FUNCTIONAL CO-REGULATOR BINDING BY INTEGRATING MULTIPLE SOURCES OF BIOLOGICAL INFORMATION

4.1 BACKGROUND

Transcription factors (TFs) serve as the final molecules in signal transduction pathways that coordinate expression of target genes. When activated in response to upstream signals, often encoded as chemical ligands and protein modification, TFs bind to their cis-regulatory sites to exert their regulatory effects on their target genes. During the process, TFs often interact with other proteins, which further modulate the function and efficacy of TFs to achieve fine-tuned regulation of gene expression; studying such interactions and regulations is an increasingly important component of studying gene expression systems. Nuclear receptors (NRs), such as estrogen receptor α (ERα), are transcription factors that migrate to the nucleus (often as a result of binding ligand) to regulate downstream target genes. NRs play important biological roles in normal physiology and disease. In particular ERα plays an important role in both breast cancer and osteoporosis. Upon ligand binding, ERα and other NRs are bound by proteins called co-regulators that recruit transcriptional machinery and chromatin modifying enzymes. Co-regulators are therefore critical in NR activity. Understanding the composition of functional NR/co-regulator complexes in specific signaling contexts could provide a basis for the

development of novel NR- and co-regulator-targeted therapeutics. The problem addressed in this paper arose from a study of the interaction between the major ERα co-activator SRC-1 (a member of the p160 SRC family), also known as NCOA1, with ERα and the impact of such interactions gene expression [121-124].

Recently, chromatin immunoprecipitation coupled with high-throughput next-generation sequencing (ChIP-seq) has become the main technology for global characterization of the transcriptional impact of NRs and their co-regulators [125-127]. ChIP-seq involves the short-read (~30bp) sequencing of the ChIP-enriched DNA fragments. These short sequence reads (tags) are then aligned to a reference genome. Then the actual binding loci from the positional tag distributions (i.e. sequenced DNA fragments mapped onto a reference genome sequence) are determined using 'peak calling' algorithms. Numerous peak calling algorithms have recently been developed for identifying ChIP-enriched genomic regions from ChIP-seq experiments [108, 128, 129] but there is a wide range of discordance among the peak calls from different algorithms [130]. Therefore, there is a need for the methods that can integrate additional information besides ChIP-seq tags to identify functional TF binding sites. Furthermore, studying the interactions between TFs and their co-regulators through ChIP-seq technology poses an additional challenge since co-regulators do not directly bind DNA. Co-regulator ChIP-seq measures the secondary protein-DNA binding through primary TFs and leads to relatively weak sequencing signals—i.e. relatively small number of sequence tags above noise. As such, it remains a challenge for contemporary peak calling methods to detect weak secondary protein-DNA-binding signals and simultaneously maintain a high specificity.

Often, a well-designed experiment studying interaction between a TF and its co-regulator generates critical information in addition to the ChIP-seq data for the co-regulator binding. For example, ChIP-seq data reflecting the binding of the primary TF of interest to its cis-regulatory sites are often collected; the genomic sequence surrounding Chip-seq peaks are usually available, which can be used to reflect the intrinsic sequence characteristics of regulatory sites; transcriptomic data that reflect functional outcomes of the interaction of the TF and its co-regulators can also be monitored. In this study, we investigated and compared different statistical and machine learning approaches to integrate multiple types of information to overcome the difficulty of identifying functional ERα/SRC-1 interaction in presence of weak ChIP-seq signal.

4.2 METHODS

4.2.1 ChIP-seq data

U2OS cells stably expressing Flag-tagged ERα (obtained from Dale Leitman) were used for ChIP as previously published [42]. SRC-1 and ERα ChIP DNA from ethanol (vehicle) and estradiol (E2)-treated U2OS cells were amplified for Illumina sequencing. IgG ChIP DNA was also amplified for Illumina sequencing. The ChIP-seq datasets used in this study had the following number of uniquely mapped sequence tags, ChIP_ER_E2: 10,380,852, ChIP_SRC-1_E2:6,995,566 tags, ChIP_IgG: 8,641,543 tags. SRC-1 peaks were called using MACS 1.4.1 [108], BayesPeak [129], and T-PIC [128] with IgG as negative control.

4.2.2 Evaluation procedure

The selected peaks were evaluated in terms of their overlap with high-scoring sequence motifs. The motif analysis was performed using the program CLOVER [131] , with *P* value cutoff 0.005 which compares sets of DNA sequences to a library of transcription factor-binding motifs and identifies whether any of the motifs are statistically overrepresented or underrepresented in the sets. We measured enrichment of selected motifs in sets of ± 300 bp from SRC-1 ChIP-seq peak summit.

4.2.3 Computational Framework

We investigated the following computational approaches to identify potential binding sites, including unsupervised classification, supervised classification and semi-supervised classification. The task was formulated as a binary classification problem for supervised and semi-supervised framework, where each ChIP-seq peak was either ‘functional’ or ‘non-functional’. Each ChIP-seq peak was represented with a vector of binary features, where each feature was derived from one biological information source.

4.2.4 Features

We devised a total of 67 features, which can be grouped as follows. 1) Genomic information: trigrams (triplet of nucleotides) to represent intrinsic characteristics of genome sequence surround the peak summit to create a feature vector; averaged nucleosome occupancy prediction

results as another feature. 2) Primary TF binding events: the called ER ChIP-seq data peak that overlap with SRC-1/ERα ChIP-seq. 3) Functional outcome of TF activation: whether the peak is mapped to SRC-1 sensitive gene.

4.2.4.1 N-gram Presence (64 Features)

Previously, n-gram distribution of sequences have been utilized for TF binding site prediction [132]. An N-grams consists of a sequence of *n* letters, where letters are possible nucleotide (A,T,G,C) bases of DNA sequences in ChIP-seq peaks. As such, a trigram has 64 possible combinations of three nucleotides, and we constructed a vector of length of 64 elements used the Galaxy Toolkit [133], each representing the presence or absence of a give trigram in the 600bp surrounding the summit of the peak of interest.

4.2.4.2 Nucleosome Occupancy (1 Feature)

Nucleosomes are fundamental repeating unit of eukaryotic chromatin. Nucleosomes consist of 147 bp of DNA sequence wrapped around a histone core complex, and they are separated from each other by linker DNA of up to 50 bp. Recently, Tillo et al. [134] proposed that nucleosome occupancy of DNA sequence around functional human transcription factor binding sites (TFBSs) is remarkably higher. To represent the nucleosome occupancy status of the Chip-seq peaks, we use the scores from Kaplan et al.'s [135] genome wide nucleosome predictions. For each base location of human genome, Kaplan et al. provided the "average occupancy" score, which is the predicted probability for each position in the genome to be covered by any nucleosome. For each peak, we took the mean value of average occupancy score around ±50 bp (an approximate length of a nucleosome) region of the peak summit. For each candidate peak, its nucleosome

occupancy feature is represented as a binary variable, with value set equal 1 if the mean value greater than 0.75 and 0, otherwise.

4.2.4.3 Primary TF binding events (1 Feature)

For each candidate SRC-1 peak, we associate a binary variable to indicate if the peak overlaps with any ERα ChIP-seq peak. We defined that an ERα and an SRC-1 peak overlap if they share at least one base pair.

4.2.4.4 Functional outcome of TF activation (1 Feature)

We collected the gene expression data from cells that were treated with vehicle and E2 in presence and absence of anti-SRC-1 siRNA have been employed for our analysis. Differentially expressed genes between these samples were found using limma (Linear Models for Microarray Analysis) package - an implementation of the empirical Bayes linear modelling approach [136] . We identified a list of genes that were differentially expressed between the control vs anti-SRC-1 siRNA groups and labelled them as SRC-1 sensitive genes. *ChIPpeakAnno* [137] was used to map each ChIP-seq peak to a gene if possible using default setting of the program. For each candidate SRC-1 peak, we associate a binary variable to indicate if the peak is mapped to one of SRC-1-sensitive genes.

4.2.5 Machine learning approaches

For unsupervised learning, k-means clustering, training and classification procedures for supervised and semi-supervised framework are implemented using the MATLAB® (Natick,

MA). We rank key features by ROC class reparability criteria using also MATLAB®. The microarray data analysis was done with the use of the R packages from the Bioconductor project (www.bioconductor.org). We used DAVID [138] for GO analysis.

4.2.5.1 Unsupervised Clustering

We used k-means clustering (k=2) with city block distance metric to see cluster candidate peaks into two groups.

4.2.5.2 Supervised Classification

To build this type of classifiers, labelled data of both true-positive peaks and false-positive peaks were required. We experimentally validated 18 SRC-1 peak by quantitative PCR (qPCR) experiments (data not shown), which were used as positive training cases, together with a set of randomly drawn control (anti-IgG) ChIP-seq peaks as negative training cases, to train supervised classifiers. We investigated the performance of three state-of-the-art classifiers: Naive Bayes (NB) [139] implemented by the MATLAB, Support Vector Machines (SVM)[140] and Random Forest (RF) [141]. Different ratios of positive to negative cases, (1:1, 1:2 and 1:3), were considered in this study for testing, and training.

NB classifier with *Bernoulli* distribution was used where each peak represented as binary-valued feature vectors. For SVM, we studied different types of kernels and chose the polynomial kernel in this study. For training RF classifiers, we grew 50 trees. For the number of variables randomly selected at each node, we used the default value that was equal to the square root of the feature dimension.

We measured performance of classifiers with 9-fold cross-validation process and report precision, recall and accuracy values. Precision and recall were used in order to evaluate model performance of classifier. Precision was measured as the fraction of correctly predicted TP binding sites (experimentally verified) among all binding sites predicted by the classifier to be TP binding site. Recall is the fraction of the TP binding sites that are also predicted to TP. Accuracy is calculated as the fraction of correct calls (TP + TN) overall total number of predictions.

4.2.5.3 Semi-supervised Classification

Self-training is one of the common algorithms used for semi-supervised learning [142]. In self-training [143], a classifier is built from labeled instances (*L)* and used to predict the labels for instances in unlabeled set (*U)*. Then *m* instances in *U* that the current classifier has high classification confidence are labeled and moved to enlarge *L*. The whole process iterates until stopped. The stopping criterion in self-training is that, either there is no unlabeled instance left or the maximum number of iterations has been reached. Different ratios of positive to negative cases, (1:1, 1:2), were considered in this study for testing, and training The detailed algorithm is shown below.

Algorithm

Input: positively labeled data (*P*) $\{(x_i, y_i)\}_{i=1}^{lp}$, negatively labeled data (*N*) $\{(x_j, y_j)\}_{j=1}^{Ln}$, and

unlabeled data (*U*) $\{x_k\}_{j=lp+Ln+1}^{lp+Ln+u}$

1. Initially, let $L_0 = \{(x_i, y_i)\}_{i=1}^{lp} \cup \{(x_j, y_j)\}_{j=1}^{ln}$ and $U = \{x_k\}_{j=lp+ln+1}^{lp+ln+u}$ where $l_n = l_p$.

2. Set t, the iteration counter, to 0.

3. Repeat until the stopping criteria are not satisfied,

    a. Build a classifier C_t on L_t

    b. Apply C_t to the unlabeled instances in U_t to predict a label for each instance in U_t.

    c. Generate L_t^s by selecting unlabeled instances that C_t has the highest classification confidence as positive label and select randomly equal number of negatively labeled instances from N_t.

    d. Delete the selected instances positively and negatively labeled from U_t and N_t respectively

    e. $L_{t+1} = L_t + L_t^s$

    f. Increase t by 1

4. Return the final classifier and apply it to the U.

4.3 RESULTS AND DISCUSSION

The biological study underlying this paper aims to investigate the impact of ERα/SRC-1 interaction on estrogen induced gene expression in a bone cell line transfected with ERα (U2OS-ERα), which may shed light on the effect of estrogen-related bone development, bone loss, and potentially bone metastasis. We have generated ChIP-seq data using anti-ERα and anti-SRC-1

antibodies in presence and absence of estradiol (E2). To further investigate the impact of interactions between this NR//co-regulator pair, we collected expression array data from the same cell lines with a combination of E2 treatment and SRC-1 knock down. In general, the results of an SRC-1 ChIP-seq experiment would reflect secondary, indirect binding of SRC-1 to DNA through multiple NRs.. However, in this study our experimental design aims to investigate specifically estrogen-induced interactions between ERα, SRC-1, and DNA. The detailed results of the experiments are being prepared for a separate publication (Hartmaier et al., manuscript in preparation). In the current paper, we address the fundamental issue of identifying reliable and functional ERα/SRC-1 DNA binding sites.

4.3.1 Identifying SRC-1 binding sites based on anti-SRC-1 ChIP-seq

We first set out to investigate the efficacy of studying ERα/SRC-1 DNA binding sites only based on the results of ChIP-seq experiments performed using an anti-SRC-1 antibody. Potential SRC-1 binding peaks were identified using three different algorithms: MACS 1.4.1 [108], BayesPeak [129], and T-PIC [128]. Table 5 shows the number of peaks identified by the above algorithms with different cut-off thresholds and the corresponding number of genes to which the peaks are mapped.

Table 5 The number of peaks called by different algorithms and at thresholds, and corresponding number of mapped genes.

* Union and intersection of the peaks by the three methods, as shown in Figure 10.

Method	Total number of peaks	Number of genes mapped
MACS, p=1E-8	1,966	996
MACS, p=1E-5	4,678	2,054
MACS, p=1E-3	23,306	6,341
T-PIC, p=1E-3	4,453	1,676
T-PIC, p=1E-2	6,598	2,318
BayesPeak (PP=0.90)	15,622	4,495
BayesPeak (PP=0.70)	21,373	5,507
BayesPeak (PP=0.5)	27,990	6,533
Union*	38,324	8,057
Intersection*	4,811	2,029

The results of the table raise the following issues during interpretation: First, as expected, applying different cut-off thresholds to the results by a given algorithm leads to a different number of peaks being identified: there is inevitably a trade-off between the number of peaks recovered and the quality of the peaks. Second, different algorithms assess the quality of the peaks based on different assumptions and methodologies: there is no consensus on the "goodness" of quality scores of these algorithms. We further noted that different versions of a same algorithm return different quality scores. Finally, different algorithms return disjointed sets of peaks, as shown in Figure 10, indicating that distinct assumptions and approaches enable an algorithm to discover some potential peaks that evade detection by other algorithms. These issues force decisions that potentially impact the conclusions of the study such as: which algorithm performs better, what cut-off threshold for a given algorithm to pick, and how to

consolidate the results from different algorithms so that one can maximize the number of high quality peaks. Making these choices remains challenging due to the lack of consensus in the field [130, 144].

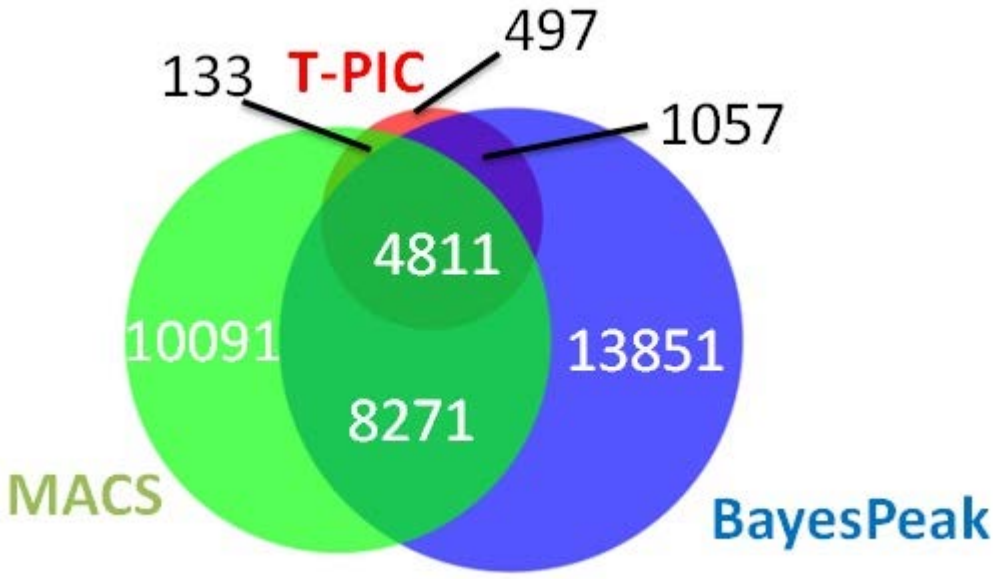

Figure 10 Peak calling by different algorithms

A Venn diagram shows the overlaps among the peaks called by MACS (*P* value cutoff of 10^{-3}), T-PIC (*P* value cutoff of 10^{-2}) and BayesPeak (*PP* cutoff of 0.5). The number of peaks are shown. The numbers of the union and intersection of the peaks and the mapped genes by the algorithms are shown in Table 5.

In order to compare our results with a recently published study by Lanz et al [145], which analyzed DNA recruitment of the co-regulators SRC-3, we studied the estrogen-induced SRC-1 peaks identified from our data using MACS algorithm with a cut-off threshold of *P* value at 10^{-10}, a threshold based on their study. Our analysis yielded a total of 1,286 peaks, which were further mapped to 684 genes. The number of peaks identified by us with the above condition is far fewer compared to their study. The discrepancy is likely to be, at least in part, due to amplification of SRC-3 in MCF-7 cells used in their study and possible differences in antibody

affinities. However, these results also raised the hypothesis that ChIP-seq signals of secondary binding at physiologic levels are usually weaker. This suggests that conventional cutoff thresholds for peak calling algorithms may be too stringent, neglecting weak peaks (peaks with relatively small number of tags) potentially resulting from real ER -1/SRC-1 DNA interactions. Therefore, additional information besides SRC-1 ChIP-seq tags should be capitalized to enhance identification of functional binding sites.

4.3.2 Integrating multiple sources of biological information for identifying SRC-1 binding sites

To corroborate the results of SRC-1 ChIP-seq, we also studied the ERα ChIP-seq data (reflecting the expected dominant SRC-1-interacting TF) and investigated peaks overlapping between the ERα and the SRC-1 ChIP-seq results. By varying the cut-off threshold of MACS, we identified different numbers of overlapping peaks between ERα and SRC-1, with the number of overlapping peaks increasing as the cut-off threshold relaxes (data not shown). The results again indicated that the conventional cut-off thresholds are failing to identify putative ERα/SRC-1 DNA binding sites (false negatives). On the other hand, simply relaxing the cut-off threshold is likely leading to increased false positive peak calls. Thus a principled method is needed to further identify functional ERα/SRC-1 DNA binding sites.

To elucidate *functional* ERα/SRC-1 DNA binding events, i.e., the binding events that influence gene expression, we generated and analyzed expression array data from cells that were treated with vehicle and E2 in the presence and absence of SRC-1 siRNA (Hartmaier et al, manuscript in preparation). The microarray data enabled us to identify 634 genes whose

expression response to estrogen treatment required SRC-1, which are hereto referred to as SRC-1-sensitive genes. When we compared the list of SRC-1-sensitive genes and the list of genes with SRC-1 binding sites derived by MACS at *P* value = 10^{-10}, we noted that only 44 genes overlapped among the lists. While the discrepancy between the number of SRC-1-sensitive genes and genes with SRC-1 peaks could be explained by other biological factors, such as that many SRC-1 peaks were not functional or secondary expression effects, it also supported our general hypothesis that ERα/SRC-1 interaction ChIP-seq signal is relatively weak and potentially true functional ERα/SRC-1 DNA binding sites were missed by the stringent setting of the peak calling algorithm.

While it may be tempting to directly combine the information from SRC-1 ChIP-seq, ERα ChIP-seq and microarray data by identifying the intersections of overlapping genes and peaks, such an approach is overly simplistic and ignores other potentially informative data, e.g., the genome-sequence characteristics of ERα/SRC-1 interaction sites and the prior information of known ERα/SRC-1 interactions. These considerations motivated us to investigate and compare different principled machine learning approaches in order to improve the sensitivity and specificity of detecting ERα/SRC-1 DNA binding sites by integrating multiple types of information.

4.3.3 An integrative approach to detect ERα/SRC-1 DNA binding sites

The overall framework and rationale of our information integration approaches are as follows. We formulated the task of identifying functional ERα/SRC-1 DNA binding sites as a classification task, in which we performed a binary classification to label a potential SRC-1

binding site derived from ChIP-seq analysis as either functional or nonfunctional. We investigated both supervised learning, which allows us to take advantage of existing knowledge of ERα/SRC-1 interactions, and unsupervised learning, which allowed us to take an unbiased approach.

The classification formulation allowed us to pool more candidate peaks identified by different peak calling algorithms at relaxed cutoff thresholds so that we did not have to rely on a single "best" algorithms and "optimal" parameterization but resorted to our classification to identify functional ERα/SRC-1 DNA binding. In this study, we collected the union of the peaks returned by all three algorithms at the cutoff threshold as follows, MACS: *P* value cutoff 10^{-3}, BayesPeak: Posterior Probability (PP) ≥ 0.5 and T-PIC: *P* value cutoff 10^{-2}. This led to a pool of 38,324 candidate peaks.

Another important advantage of the classification approach is that it allows us to integrate multiple types of biological information collected from our experiments and public databases by representing them as features for a classifier. In this way, multiple types of information contribute to the classification of potential peaks and their impact can be determined by learning algorithms. For each candidate SRC-1 peak, we constructed the following features: a vector of binary features representing presence/absence of nucleotide trigrams (triplet of nucleotides), which reflects the intrinsic characteristics of genome sequence surrounding the summit of a peak region; an average of predicted nucleosome-occupancy scores, which represents the chromatin structure characteristics around the peak summit; a binary feature reflecting if a primary binding peak, i.e., the ERα ChIP-seq peak, overlaps with the SRC-1 peak; and a binary feature representing the functional outcome of ERα/SRC-1 interaction, i.e., whether the peak is mapped to an SRC-1 sensitive gene. Detailed descriptions of features are presented in Methods section.

We evaluated the results of predictions from classification algorithms by determining if conserved ERα binding motif can be found in the classified peaks, as an indication that a peak is the result of ERα/SRC-1 DNA binding. Searching for instances of conserved TF binding motifs at the predicted binding loci is considered the most prominent verification method for validating peaks [146].

Table 6 Comparison of the performances by different machine learning algorithms

	Number of peaks	Number of peaks with ERE motif	Ratio of peaks with ERE motif match
MACS p=1E-10	1,286	941	0.73
MACS p=1E-8	1,966	1,416	0.72
MACS p=1E-5	4,678	3,077	0.66
k-means (city block)			
Cluster 1	26,211	11,943	0.46
Cluster2	12,113	3,245	0.27
supervised-NB(th=0.8,1:2)			
Positively Labeled	11,835	8,196	0.69
Negatively Labeled	26,489	6,992	0.26
supervised-SVM(kernel=polynomal,1:2)			
Positively Labeled	14,915	8,425	0.56
Negatively Labeled	23,409	6,763	0.29
supervised-RF(th=0.7,1:2)			
Positively Labeled	10,428	6,514	0.62
Negatively Labeled	27,896	8,674	0.31
semi-supervised-NB(th=0.8,1:2, I=75)			
Positively Labeled	12,597	8,458	0.67
Negatively Labeled	25,727	6,730	0.26

4.3.4 Unsupervised classification

First, we explored if the candidate peaks could be divided into two distinct groups by unsupervised learning in an unbiased manner. We applied a K-means clustering procedure to the

data and the results are listed in Table 6. We inspected the genome sequences of the peaks to assess if a conserved motif for estrogen response element (ERE) was detected by a motif classification algorithm referred to as CLOVER[131]. In cluster 1, 46% of 26,211 peaks contained the ERE motif, and, in cluster 2, 27% of 12,113 peaks contained the ERE motif. We believe this is not a good separation of the peaks in that, even though cluster 1 has more ERE-containing peaks, it is a bigger cluster and only 46% of peaks contain EREs. Thus the results would likely lead to a high false positive rate with respect to SRC-1 binding.

4.3.5 Supervised Classification

Supervised learning requires labeled data as training cases. Obtaining a training set through large-scale experimental validation of ERα/SRC-1 DNA binding is costly and difficult to perform. Therefore, we investigated whether a relatively small amount of labeled data in the supervised learning task would result in better separation compared to unsupervised clustering with our feature set. We experimentally validated 18 SRC-1 peaks as functional peaks by quantitative PCR (qPCR) experiment (data not shown). We used these peaks as positive training cases, together with a set of randomly drawn control (anti-IgG) ChIP-seq peaks as negative training cases, to train supervised classifiers. We investigated the performance of three state-of-the-art classifiers: Naive Bayes (NB), Support Vector Machines (SVM) and Random Forest (RF). For NB and Random Forest classifiers, we set the classification thresholds at 0.8 and 0.7 respectively. Since the ratio of the positive and negative training cases may have impact on classification algorithms, e.g. NB and SVM, we explored using different ratios for training,

between 1:1, 1:2, and 1:3, and classifiers were built from these training sets. Our test set consisted of all 38,324 candidate peaks. Table 6 lists the total number of peaks in each class, the number ERE-containing peaks in each class and the ratio reflecting ERE enrichment. Results for classifiers with 1:1, 1:2, and 1:3 training ratio (positive over negative) were very similar to each other (data not shown). Therefore, just results for the 1:2 ratio were shown.

We noted that the supervised classification approaches have significantly increased the number of positive peaks when compared to those derived by peak calling algorithms based on recommended cutoff thresholds. For example, NB returned 11,835 positive peaks in comparison to 1,966 and 4,678 peaks returned by MACS with cutoff *P* value set at 1E-8 and 1E-5, which reflected a 6-fold and 2.5-fold increase, respectively. Through further evaluation enrichment of ERE in the genome sequences surrounding the peaks, we found that a similar percentage of peaks contained ERE element: 69% for NB, and 72% and 66% MACS at 1E-8 and 1E-5 respectively. Thus, the results indicate that the qualities of the positive peaks returned by NB were as good as those returned by the stringent peak calling in terms of ERE enrichment.

Table 7 Comparison of different methods for identifying functional peaks

Method	Total number of peaks	Number of genes mapped	Intersection with SRC-1-dependent genes
MACS p=1E-10	1,286	684	44
MACS p=1E-8	1,966	996	57
MACS p=1E-5	4,678	2,054	123
supervised-NB(th=0.8,1:2)	11,835	3,875	238

We further inspected if the classification approaches retrieved additional *functional* peaks, i.e., the peaks that were mapped to SRC-1-sensitive genes derived from microarray

experiment. Table 7 shows the results of the SRC-1 peaks returned by different peak calling approaches that were mapped to SRC-1-sensitive genes. We noted that, by setting MACS cutoff *P* values at 1E-10, 1E-8, and 1E-5, a total of 44, 55, and 123 peaks were mapped to SRC-1-sensitive genes. On the other hand, NB has identified 238 peaks that overlap with SRC-1-sensitive genes. We also note that BayesPeak exclusively identified some of the newly "discovered" functional peaks. Similarly MACS also exclusively discovered some new peaks. These results indicate that, based on different assumptions and criteria, different peak calling algorithms are capable of identifying potential peaks to complement other peak calling algorithms. Thus it is more sensible to consider candidate peaks from more than one peak-calling algorithm as long as an objective approach can be further applied to consolidate the results.

We also performed a 9-fold cross-validation experiment to assess if the algorithm can correctly identify the experimentally validated ERα/SRC-1 training cases from the candidate peak pool. Results were listed in Table 8. NB classifier showed 86% precision, 100% recall and 96% accuracy, see Methods section for the descriptions of the metrics. This result increased our confidence that positive calls from our algorithm are likely to reflect real ERα/SRC-1 DNA binding. The results in the table indicate that the NB classifier performed better than the SVM and RF classifiers, judging from relative enrichment of ERE containing binding sites in the predicted positive peaks. Among the three classifiers tested in this study, the RF classifier performed worst. We noted that the number of features that were used by RF during the learning was much smaller than the number of features utilized by other classifiers, which may partially explain the inferior performance of this algorithm in this experiment. The SVM method also performed worse on this task than probabilistic NB. We conjecture that the reason might be that

SVM is complex algorithm with many parameters to adjust and therefore finding optimal parameters for decision boundary might be challenging for this task. We therefore concentrated on the NB classifier because it could be readily used in both supervised and semi-supervised learning environment.

Table 8 Performance of different classifiers under 9-fold cross-validation setting

Classifier	Precision	Recall	Accuracy
NB(th=0.8,1:2)	0.89	1	0.96
SVM(kernel=polynomial, 1:2)	0.89	0.96	0.94
RF(th=0.7,1:2)	0.72	1	0.91

4.3.6 Semi-supervised Classification

Our number of training cases is relatively sparse compared to a conventional machine learning setting. Semi-supervised approaches have been applied to conquer limitations of supervised and unsupervised methods when labeled data is scarce and obtaining large amounts of labeled data is expensive and time consuming [142, 143]. This is done by incrementally assigning instances, which are called with high confidence by a classifier, from unlabelled data into training cases in order to increase the number of training cases and thus enhance the generalizability of classification. Therefore, we investigated semi-supervised classification to see whether we could further increase the performance in identifying ERα/SRC-1 DNA binding.

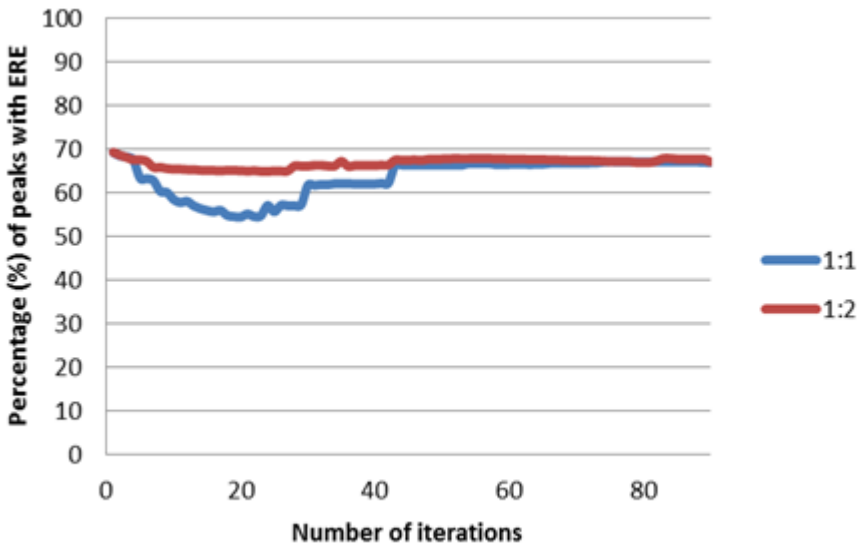

Figure 11 Self-training.

Percentage of predicted positive peaks with ERE motifs (over iterations for different TP:TN ratios for training set as indicated in the legends.

In this study, we applied a self-training algorithm [143] using NB as the base classifier because of its probabilistic outputs. We iteratively assigned the most confident positive instances called by our classifier into training cases and found that performance of the self-training became stable after 75 iterations and stopped further training. Figure 11 shows the trends of percentage of ERE-containing peaks in the positive calls in semi-supervised learning. It is interesting to note that initially as a few pseudo-positive cases were imputed into the training cases the precision of the called positive peaks decreased but later became stable after 75 iterations. A similar total number of peaks and the percentage of ERE-containing peaks were identified by our semi-supervised learning algorithm when compared to other supervised learning experiments, see Table 2. Thus the results do not show obvious advantage of semi-supervised learning over supervised learning algorithms in our experiment.

4.3.7 Identification of Informative Features

Biologically, it is of interest to identify the features that significantly contribute to the classification in that it will reveal the relationships between input features and outcome. We rank key features by ROC class reparability criteria using MATLAB (Bioinformatics Toolbox) [147], using a training dataset containing the 18 true positive peaks and 36 random non-binding sites from IgG peak calls. Following were the 15 top ranked features: *"AAC"*, *peaks-mapped-to-SRC-1-dependent-genes*, *"GCG"*, *"CGT"*, *overlapping-with-ERα-peak*, *"AAG"*, *"ACA"*, *"ACC"*, *"CGC"*, *"AGA"*, *"AGC"*, *"AGG"*, *"CGA"*, *"ACT"*, *"ATC"*, see Methods section for detailed descriptions of the features. Among the top-ranking features, we noted that the features reflecting the function outcome (*peaks-mapped-to-SRC-1-dependent-genes*) and the interaction between ER and SRC-1 (*overlapping-with-ERα-peak*) were ranked high, indicating the learning algorithm correctly recognized their importance in classification. It is interested to note that many nucleotide trigrams, which reflect the characteristics of sequences of peaks, were among the high-ranking features. We aligned the top-ranking trigrams to the ERE motif, as shown in Figure 12. Indeed, the trigrams correspond well with the important components of the ERE motif. These results indicate that the classification learning algorithm, like the motif searching algorithms (i.e. [148-151]), is able to identify highly conserved "words" that constitute one of the important motifs of the training sequences. We noted that the feature reflecting nucleosome occupancy at peak regions was ranked as 48^{th}. This may indicate either that nucleosome occupancy is a dynamic process and our static feature does not reflect the true occupancy status during the experiments or that the ERa/SRC-1 DNA binding is not heavily dependent on nucleosome occupancy.

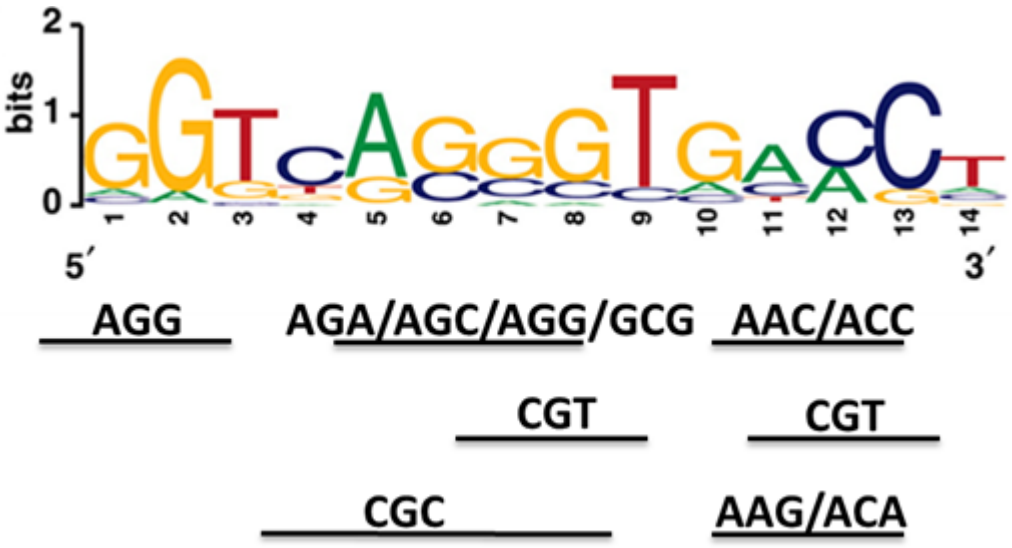

Figure 12 Overlapping top trigrams with ERE motif.

This figure shows potential matching locations of the top-ranking nucleotide trigrams identified by feature selection algorithms.

4.3.8 Biological Insights from Improved Peak Calling

We further examined the impact of the improved SRC-1 peak calling on biological insights drawn from the dataset. We conducted Gene Ontology Analysis using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) [138] on genes with an SRC-1 peak within 50kb of the TSS as determined by MACS ($P<10^{-5}$) or by our method. We observed a dramatic difference in the identification of genes enriched in specific biological processes. Specifically, our method yielded in the calling of peaks in gene sets which were highly enriched for genes involved in blood vessel development (enrichment: 4.61, Benjamani: $7.9x10^{-4}$) and actin filament-based processes (enrichment: 3.18, Benjamani: $1.3x10^{-3}$). Indeed blood vessel development has previous been implicated in bone generation [152, 153]Further, within the

genes enriched in these biological processes, we identified a number of genes with known functions in bone development. Since SRC-1 has already been implicated in E2 mediate bone maintenance, this observation provides evidence for the mechanism underlying this phenotype. In contrast, genes with SRC-1 peaks determined by MACS were not significantly associated with any biological processes.

4.4 CONCLUSIONS

We believe that the ability to improve ChIP-seq peak calling by utilizing available sources of biological information for indirect co-regulator binding in the presence of weak ChIP-seq signal is an important research area. Due to the intrinsic variability in the affinity of interactions between a TF and its co-regulators, it is inevitable that the ChIP-seq signal of these types of studies would span a broad spectrum and that the weak signal scenario, as in this study, would be likely to occur often. The need for methods to address this problem is acute considering the increasing number of studies using ChIP-seq to study NR and their co-regulators due to their importance in normal development and in many diseases such as breast cancers. Our work strives to explore whether the peak calling can be improved through the integration of available diverse biological sources via machine learning approaches. Our results demonstrate that it is informative to generate, collect, and integrate the following information: ChIP-seq data reflecting location of the primary interaction of the TF of interest to its cis-regulatory sites, gene expression data reflecting functional outcomes of interaction of the TF and its co-regulators, and finally genomic sequence data of the identified regions. Other types of data which is highly

likely to be useful include histone modification marks, recruitment of RNA polymerase II, and relative location of the insulator protein CTCF[154]

In summary, our results indicate that a supervised classification approach enables one to utilize even limited amounts of existing knowledge together with multiple types of biological data to enhance the sensitivity and specificity of identifying DNA binding sites for co-regulators proteins. Our feature selection experiments indicate that experimental inputs complementary to ChIP-seq are critical in identifying biological significant signals from ChIP studies with weak signals due to indirect DNA binding.

5.0 CONCLUSIONS AND FUTURE WORK

ERαs transduce estrogen response in many tissues including the breast and bone. Estrogen-induced ERαs activate and inhibit specific genes involved in cell cycle progression and cell survival. Genes regulated by estrogen are important for proliferation, differentiation, survival and particularly in cancer the stimulation of invasion, metastasis and angiogenesis. ERα action in response to E2 exposure is necessary for healthy physiology, but it is also a hallmark of malignant breast cancer.

The advent of genomic technologies for examining signal-regulated transcriptional responses and TF binding sites, including ERα, has increased our understanding of the factors that control hormone signaling and transcriptional regulation of genes. However, we are still facing bioinformatics challenges including integration of various data sources as well as dealing with noise from the experiments in order to analyze this rich genomic data sources.

In the first part of this thesis, I integrated a variety of recent genome-wide high-throughput datasets, including gene expression arrays, ChIP-seq, GRO-seq and ChIA-PET in order to derive a holistic perspective of the transcription machineries at estrogen-responsive genes, and reveal different mechanisms of estrogen-mediated transcription regulation. Our analyses have led to the following some novel findings: In the absence of the ligand, most of the estrogen-responsive genes assumed a high-order chromatin configuration that involved Pol II, ERα and ERα-pioneer

factors. Without the ligand, estrogen-induced genes showed active transcription at promoters but failed to elongate into gene bodies, and such a pause was lifted after estrogen treatment. However, the estrogen-repressed genes showed coordinated transcription at promoters and gene bodies in the absence and presence of estrogen. Through information integration, we inferred that, for estrogen-repressed genes, the majority of the high-order chromatin complexes containing actively transcribed genes were disrupted after estrogen treatment. The analyses led to the hypothesis that one mechanism for estrogen-mediated repression is through disrupting the original transcription-favoring chromatin structures.

Further, ERs interact with co-regulators to regulate gene transcription. Understanding the mechanism of action of co-regulator proteins—which do not bind DNA directly, but exert their effects by binding to transcription factors—is important for the study of normal physiology as well as diseased conditions. In the second part of this thesis, I investigated and compared different statistical and machine learning approaches including unsupervised, supervised, and semi-supervised classification (self-training) approaches to integrate multiple types of genomic and transcriptomic information derived from our experiments and public database to overcome difficulty of identifying functional DNA binding sites of the co-regulator SRC-1 in the context of estrogen response. Our results indicate that supervised learning with naïve Bayes algorithm significantly enhances peak calling of weak ChIP-seq signals and outperforms other machine learning algorithms. Our integrative approach revealed many potential ERα/SRC-1 DNA binding sites that would otherwise be missed by conventional peak calling algorithms with default settings. Our results indicate that a supervised classification approach enables one to utilize

limited amounts of prior knowledge together with multiple types of biological data to enhance the sensitivity and specificity of the identification of DNA binding sites from co-regulator proteins.

5.1 FUTURE WORK

There are several potential directions for future extensions of this research. Some of them are outlined below.

In this thesis, I combined various next-generation sequencing data related to ERα to decipher new transcriptional mechanism in MCF7 breast cancer cells. With the rapid development of sequencing technologies, ChIP-seq, GRO-seq, ChIA-PET as well as other technologies such as Hi-C, data collection is becoming more readily available for a variety of cell types. Therefore, this kind of framework could be applied to other tissue types as well as other NRs. By integrating these data sources, one could further explore a fundamental question in the gene regulation: why do the same factors (i.e. transcription factors, co-regulators) differentially regulate gene transcription in different tissues and diseases, particularly in cancer? Moreover, one could identify previously unrecognized regulatory mechanisms that contribute to gene regulation under different conditions as well as in different tissues which may provide new approaches for treatment. There is an increasing flood of genomic data from research consortiums such as TCGA (the Cancer Genome Atlas) and ENCODE (The Encyclopedia of DNA Elements). Eventually, by mining these datasets, an automatic hypothesis generation

pipeline can be developed in order to study transcription regulation under different conditions as well as different tissues.

Another future aim of my work is to improve and extend the framework developed in the second part of the thesis. Due to the intrinsic variability in the affinity of interactions between a TF and its co-regulators, it is inevitable that the ChIP-seq signal of these types of studies would span a broad spectrum and that the weak signal scenarios would occur often. The binary classification setting I provide, and most of the features I derive, can be extended to improve ChIP-seq 'peak calling' for other co-regulators as well as for transcription factors in the presence of weak signal from experiments. The framework developed in this thesis proved that it is informative to generate, collect, and integrate the following information: 1) ChIP-seq data that reflects the location of primary-interacting TFs of co-regulators; 2) gene expression data that reflects functional outcomes of interactions between TF and its co-regulators; and finally 3) the genomic sequence data of the identified binding regions. The features of this framework can be further extended by mining additional data sources including histone modification marks, Pol II binding events, relative locations of CTCF binding events, and 3D chromatin interaction data.

Another direction would be to extend this framework to the other relevant TF-DNA binding problems. For example, the majority of NR binding sites (including ERα, RARA and AR) are away from the TSS of genes. These NRs bind to the distal enhancer elements, which are the key drivers of spatiotemporal specificity in gene regulation. Although with the development of ChIP-seq technology global binding sites can be detected genome-wide, we do not know which binding sites are functional enhancers. The identification of functional-enhancer binding sites is one step in a larger effort to understand the DNA sequence features that underlie the enhancer function as well as to identify tissue-specific enhancers. The semi-supervised machine

learning framework developed in this thesis—which makes use of the results of small scale wet-lab experiments as well as other information such as sequence characteristics and available genomic information—can be applied to the identification of genome-wide “functional enhancer sites” in a similar fashion.

APPENDIX A

LIST OF PAPERS PUBLISHED

1. **Osmanbeyoglu HU**, Hartmaier RJ, Oesterreich S, Lu X (2012) Improving ChIP-seq peak-calling for functional co-regulator binding by integrating multiple sources of biological information. BMC Genomics 13 Suppl 1: S1.
2. Kohle-Ersher A, Chatterjee P, **Osmanbeyoglu HU**, Hochheiser H, Bartos C (2012) Evaluating the Barriers to Point-of-Care Documentation for Nursing Staff. Comput Inform Nurs 30: 126-133.
3. **Osmanbeyoglu HU**, Ganapathiraju MK (2011) N-gram analysis of 970 microbial organisms reveals presence of biological language models. BMC Bioinformatics 12: 12.
4. **Osmanbeyoglu HU**, Ganapathiraju MK (2011) Rapid deployment of viral-human interactome prediction for new viruses. Proc of the American Medical Informatics Association Summit on Translational Bioinformatics.
5. Chalancon G, Kosloff M, **Osmanbeyoglu HU**, Saraswathi S (2010) PLoS Computational Biology conference postcards from ISMB 2010. PLoS Comput Biol 6: e1002000.
6. **Osmanbeyoglu HU**, Wehner JA, Carbonell JG, Ganapathiraju MK (2010) Active machine learning for transmembrane helix prediction. BMC Bioinformatics 11 Suppl 1: S58.

LIST OF PAPERS UNDER SUBMISSION

1. **Osmanbeyoglu HU**, Day R, Oesterreich S, Benos PV, Lu X. Estrogen represses gene expression through reconfiguring chromatin structures.
2. Hartmaierr R, **Osmanbeyoglu HU**, Benos PV, Lu X, Oesterreich S. SRC-1 recruitment reveals functional binding sites.

BIBLIOGRAPHY

1. **Breast Cancer** [http://www.cancer.org/Cancer/BreastCancer/DetailedGuide/breast-cancer-key-statistics]
2. Beatson GT: **On the treatment of inoperable cases of carcinoma of the mamma: suggestions for a new method of treatment, with illustrative cases**. *Lancet* 1896, **148**(3802):104-107.
3. Hiraku Y, Yamashita N, Nishiguchi M, Kawanishi S: **Catechol estrogens induce oxidative DNA damage and estradiol enhances cell proliferation**. *Int J Cancer* 2001, **92**(3):333-337.
4. Riggs BL, Hartmann LC: **Selective estrogen-receptor modulators -- mechanisms of action and application to clinical practice**. *N Engl J Med* 2003, **348**(7):618-629.
5. Aranda A, Pascual A: **Nuclear hormone receptors and gene expression**. *Physiol Rev* 2001, **81**(3):1269-1304.
6. Smith CL, O'Malley BW: **Coregulator function: a key to understanding tissue specificity of selective receptor modulators**. *Endocr Rev* 2004, **25**(1):45-71.
7. Bulynko YA, O'Malley BW: **Nuclear receptor coactivators: structural and functional biochemistry**. *Biochemistry* 2011, **50**(3):313-328.
8. Onate SA, Tsai SY, Tsai MJ, O'Malley BW: **Sequence and characterization of a coactivator for the steroid hormone receptor superfamily**. *Science* 1995, **270**(5240):1354-1357.
9. Manavathi B, Dey O, Gajulapalli VN, Bhatia RS, Bugide S, Kumar R: **Derailed Estrogen Signaling and Breast Cancer: An Authentic Couple**. *Endocr Rev* 2012.
10. Lavinsky RM, Jepsen K, Heinzel T, Torchia J, Mullen TM, Schiff R, Del-Rio AL, Ricote M, Ngo S, Gemsch J *et al*: **Diverse signaling pathways modulate nuclear receptor recruitment of N-CoR and SMRT complexes**. *Proc Natl Acad Sci U S A* 1998, **95**(6):2920-2925.
11. Smith CL, Nawaz Z, O'Malley BW: **Coactivator and corepressor regulation of the agonist/antagonist activity of the mixed antiestrogen, 4-hydroxytamoxifen**. *Mol Endocrinol* 1997, **11**(6):657-666.
12. Dobrzycka KM, Townson SM, Jiang S, Oesterreich S: **Estrogen receptor corepressors -- a role in human breast cancer?** *Endocr Relat Cancer* 2003, **10**(4):517-536.
13. Walsh CA, Qin L, Tien JC, Young LS, Xu J: **The function of steroid receptor coactivator-1 in normal tissues and cancer**. *Int J Biol Sci* 2012, **8**(4):470-485.
14. Romano A, Adriaens M, Kuenen S, Delvoux B, Dunselman G, Evelo C, Groothuis P: **Identification of novel ER-alpha target genes in breast cancer cells: gene- and cell-**

selective co-regulator recruitment at target promoters determines the response to 17beta-estradiol and tamoxifen. *Mol Cell Endocrinol* 2010, **314**(1):90-100.
15. Cirillo LA, Lin FR, Cuesta I, Friedman D, Jarnik M, Zaret KS: **Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4**. *Mol Cell* 2002, **9**(2):279-289.
16. Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, Eeckhoute J, Shao W, Hestermann EV, Geistlinger TR *et al*: **Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1**. *Cell* 2005, **122**(1):33-43.
17. Tan SK, Lin ZH, Chang CW, Varang V, Chng KR, Pan YF, Yong EL, Sung WK, Cheung E: **AP-2gamma regulates oestrogen receptor-mediated long-range chromatin interaction and gene transcription**. *EMBO J* 2011, **30**(13):2569-2581.
18. Magnani L, Ballantyne EB, Zhang X, Lupien M: **PBX1 genomic pioneer function drives ERalpha signaling underlying progression in breast cancer**. *PLoS Genet* 2011, **7**(11):e1002368.
19. Lupien M, Eeckhoute J, Meyer CA, Krum SA, Rhodes DR, Liu XS, Brown M: **Coactivator function defines the active estrogen receptor alpha cistrome**. *Mol Cell Biol* 2009, **29**(12):3413-3423.
20. Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, Li W, Carroll JS, Liu XS, Brown M: **FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription**. *Cell* 2008, **132**(6):958-970.
21. Jozwik KM, Carroll JS: **Pioneer factors in hormone-dependent cancers**. *Nat Rev Cancer* 2012, **12**(6):381-385.
22. Cohen I, Poreba E, Kamieniarz K, Schneider R: **Histone modifiers in cancer: friends or foes?** *Genes Cancer* 2011, **2**(6):631-647.
23. Cloos PA, Christensen J, Agger K, Helin K: **Erasing the methyl mark: histone demethylases at the center of cellular differentiation and disease**. *Genes Dev* 2008, **22**(9):1115-1140.
24. Mann M, Cortez V, Vadlamudi RK: **Epigenetics of Estrogen Receptor Signaling: Role in Hormonal Cancer Progression and Therapy**. *Cancers (Basel)* 2011, **3**(3):1691-1707.
25. Glass CK, Rose DW, Rosenfeld MG: **Nuclear receptor coactivators**. *Curr Opin Cell Biol* 1997, **9**(2):222-232.
26. Garcia-Bassets I, Kwon YS, Telese F, Prefontaine GG, Hutt KR, Cheng CS, Ju BG, Ohgi KA, Wang J, Escoubet-Lozach L *et al*: **Histone methylation-dependent mechanisms impose ligand dependency for gene activation by nuclear receptors**. *Cell* 2007, **128**(3):505-518.
27. Leitman DC, Paruthiyil S, Yuan C, Herber CB, Olshansky M, Tagliaferri M, Cohen I, Speed TP: **Tissue-specific regulation of genes by estrogen receptors**. *Semin Reprod Med* 2012, **30**(1):14-22.
28. Heldring N, Pike A, Andersson S, Matthews J, Cheng G, Hartman J, Tujague M, Strom A, Treuter E, Warner M *et al*: **Estrogen receptors: how do they signal and what are their targets**. *Physiol Rev* 2007, **87**(3):905-931.
29. Kushner PJ, Agard DA, Greene GL, Scanlan TS, Shiau AK, Uht RM, Webb P: **Estrogen receptor pathways to AP-1**. *J Steroid Biochem Mol Biol* 2000, **74**(5):311-317.

30. McKay LI, Cidlowski JA: **Cross-talk between nuclear factor-kappa B and the steroid hormone receptors: mechanisms of mutual antagonism**. *Mol Endocrinol* 1998, **12**(1):45-56.
31. Malik S, Jiang S, Garee JP, Verdin E, Lee AV, O'Malley BW, Zhang M, Belaguli NS, Oesterreich S: **Histone deacetylase 7 and FoxA1 in estrogen-mediated repression of RPRM**. *Mol Cell Biol* 2010, **30**(2):399-412.
32. Stossi F, Likhite VS, Katzenellenbogen JA, Katzenellenbogen BS: **Estrogen-occupied estrogen receptor represses cyclin G2 gene expression and recruits a repressor complex at the cyclin G2 promoter**. *J Biol Chem* 2006, **281**(24):16272-16278.
33. Oesterreich S, Deng W, Jiang S, Cui X, Ivanova M, Schiff R, Kang K, Hadsell DL, Behrens J, Lee AV: **Estrogen-mediated down-regulation of E-cadherin in breast cancer cells**. *Cancer Res* 2003, **63**(17):5203-5208.
34. Newman SP, Bates NP, Vernimmen D, Parker MG, Hurst HC: **Cofactor competition between the ligand-bound oestrogen receptor and an intron 1 enhancer leads to oestrogen repression of ERBB2 expression in breast cancer**. *Oncogene* 2000, **19**(4):490-497.
35. Wang SH, Yeh SH, Lin WH, Yeh KH, Yuan Q, Xia NS, Chen DS, Chen PJ: **Estrogen receptor alpha represses transcription of HBV genes via interaction with hepatocyte nuclear factor 4alpha**. *Gastroenterology* 2012, **142**(4):989-998 e984.
36. Kelley KM, Rowan BG, Ratnam M: **Modulation of the folate receptor alpha gene by the estrogen receptor: mechanism and implications in tumor targeting**. *Cancer Res* 2003, **63**(11):2820-2828.
37. Stossi F, Madak-Erdogan Z, Katzenellenbogen BS: **Estrogen receptor alpha represses transcription of early target genes via p300 and CtBP1**. *Mol Cell Biol* 2009, **29**(7):1749-1759.
38. Higgins KJ, Liu S, Abdelrahim M, Vanderlaag K, Liu X, Porter W, Metz R, Safe S: **Vascular endothelial growth factor receptor-2 expression is down-regulated by 17beta-estradiol in MCF-7 breast cancer cells by estrogen receptor alpha/Sp proteins**. *Mol Endocrinol* 2008, **22**(2):388-402.
39. Zhu P, Baek SH, Bourk EM, Ohgi KA, Garcia-Bassets I, Sanjo H, Akira S, Kotol PF, Glass CK, Rosenfeld MG *et al*: **Macrophage/cancer cell interactions mediate hormone resistance by a nuclear receptor derepression pathway**. *Cell* 2006, **124**(3):615-629.
40. Merrell KW, Crofts JD, Smith RL, Sin JH, Kmetzsch KE, Merrell A, Miguel RO, Candelaria NR, Lin CY: **Differential recruitment of nuclear receptor coregulators in ligand-dependent transcriptional repression by estrogen receptor-alpha**. *Oncogene* 2011, **30**(13):1608-1614.
41. Park E, Gong EY, Romanelli MG, Lee K: **Suppression of estrogen receptor-alpha transactivation by thyroid transcription factor-2 in breast cancer cells**. *Biochem Biophys Res Commun* 2012, **421**(3):532-537.
42. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF *et al*: **Genome-wide analysis of estrogen receptor binding sites**. *Nat Genet* 2006, **38**(11):1289-1297.
43. Stender JD, Stossi F, Funk CC, Charn TH, Barnett DH, Katzenellenbogen BS: **The estrogen-regulated transcription factor PITX1 coordinates gene-specific regulation**

by estrogen receptor-alpha in breast cancer cells. *Mol Endocrinol* 2011, **25**(10):1699-1709.

44. Dong XY, Sun X, Guo P, Li Q, Sasahara M, Ishii Y, Dong JT: **ATBF1 inhibits estrogen receptor (ER) function by selectively competing with AIB1 for binding to the ER in ER-positive breast cancer cells**. *J Biol Chem* 2010, **285**(43):32801-32809.
45. Perissi V, Menini N, Cottone E, Capello D, Sacco M, Montaldo F, De Bortoli M: **AP-2 transcription factors in the regulation of ERBB2 gene transcription by oestrogen**. *Oncogene* 2000, **19**(2):280-288.
46. Hurtado A, Holmes KA, Geistlinger TR, Hutcheson IR, Nicholson RI, Brown M, Jiang J, Howat WJ, Ali S, Carroll JS: **Regulation of ERBB2 by oestrogen receptor-PAX2 determines response to tamoxifen**. *Nature* 2008, **456**(7222):663-666.
47. Hsu PY, Hsu HK, Singer GA, Yan PS, Rodriguez BA, Liu JC, Weng YI, Deatherage DE, Chen Z, Pereira JS *et al*: **Estrogen-mediated epigenetic repression of large chromosomal regions through DNA looping**. *Genome Res* 2010, **20**(6):733-744.
48. Lin CY, Vega VB, Thomsen JS, Zhang T, Kong SL, Xie M, Chiu KP, Lipovich L, Barnett DH, Stossi F *et al*: **Whole-genome cartography of estrogen receptor alpha binding sites**. *PLoS Genet* 2007, **3**(6):e87.
49. Stender JD, Kim K, Charn TH, Komm B, Chang KC, Kraus WL, Benner C, Glass CK, Katzenellenbogen BS: **Genome-wide analysis of estrogen receptor alpha DNA binding and tethering mechanisms identifies Runx1 as a novel tethering factor in receptor-mediated transcriptional activation**. *Mol Cell Biol* 2010, **30**(16):3943-3955.
50. Tang Q, Chen Y, Meyer C, Geistlinger T, Lupien M, Wang Q, Liu T, Zhang Y, Brown M, Liu XS: **A comprehensive view of nuclear receptor cancer cistromes**. *Cancer Res* 2011, **71**(22):6940-6947.
51. Hassler MR, Egger G: **Epigenomics of cancer - emerging new concepts**. *Biochimie* 2012.
52. Liu Y, Gao H, Marstrand TT, Strom A, Valen E, Sandelin A, Gustafsson JA, Dahlman-Wright K: **The genome landscape of ERalpha- and ERbeta-binding DNA regions**. *Proc Natl Acad Sci U S A* 2008, **105**(7):2604-2609.
53. Paruthiyil S, Cvoro A, Zhao X, Wu Z, Sui Y, Staub RE, Baggett S, Herber CB, Griffin C, Tagliaferri M *et al*: **Drug and cell type-specific regulation of genes with different classes of estrogen receptor beta-selective agonists**. *PLoS One* 2009, **4**(7):e6271.
54. Monroe DG, Secreto FJ, Subramaniam M, Getz BJ, Khosla S, Spelsberg TC: **Estrogen receptor alpha and beta heterodimers exert unique effects on estrogen- and tamoxifen-dependent gene expression in human U2OS osteosarcoma cells**. *Mol Endocrinol* 2005, **19**(6):1555-1568.
55. Krum SA, Miranda-Carboni GA, Lupien M, Eeckhoute J, Carroll JS, Brown M: **Unique ERalpha cistromes control cell type-specific gene regulation**. *Mol Endocrinol* 2008, **22**(11):2393-2406.
56. Miranda-Carboni GA, Guemes M, Bailey S, Anaya E, Corselli M, Peault B, Krum SA: **GATA4 regulates estrogen receptor-alpha-mediated osteoblast transcription**. *Mol Endocrinol* 2011, **25**(7):1126-1136.
57. Al-Dhaheri M, Wu J, Skliris GP, Li J, Higashimato K, Wang Y, White KP, Lambert P, Zhu Y, Murphy L *et al*: **CARM1 is an important determinant of ERalpha-dependent**

breast cancer cell differentiation and proliferation in breast cancer cells. *Cancer Res* 2011, **71**(6):2118-2128.
58. Portela A, Esteller M: **Epigenetic modifications and human disease**. *Nat Biotechnol* 2010, **28**(10):1057-1068.
59. Mani RS, Chinnaiyan AM: **Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences**. *Nat Rev Genet* 2010, **11**(12):819-829.
60. Yager JD, Liehr JG: **Molecular mechanisms of estrogen carcinogenesis**. *Annu Rev Pharmacol Toxicol* 1996, **36**:203-232.
61. Liehr JG: **Is estradiol a genotoxic mutagenic carcinogen?** *Endocr Rev* 2000, **21**(1):40-54.
62. Ju BG, Lunyak VV, Perissi V, Garcia-Bassets I, Rose DW, Glass CK, Rosenfeld MG: **A topoisomerase IIbeta-mediated dsDNA break required for regulated transcription**. *Science* 2006, **312**(5781):1798-1802.
63. Williamson LM, Lees-Miller SP: **Estrogen receptor alpha-mediated transcription induces cell cycle-dependent DNA double-strand breaks**. *Carcinogenesis* 2011, **32**(3):279-285.
64. Lin C, Yang L, Tanasa B, Hutt K, Ju BG, Ohgi K, Zhang J, Rose DW, Fu XD, Glass CK *et al*: **Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer**. *Cell* 2009, **139**(6):1069-1083.
65. Haffner MC, Aryee MJ, Toubaji A, Esopi DM, Albadine R, Gurel B, Isaacs WB, Bova GS, Liu W, Xu J *et al*: **Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements**. *Nat Genet* 2010, **42**(8):668-675.
66. Haffner MC, De Marzo AM, Meeker AK, Nelson WG, Yegnasubramanian S: **Transcription-induced DNA double strand breaks: both oncogenic force and potential therapeutic target?** *Clin Cancer Res* 2011, **17**(12):3858-3864.
67. Cook PR: **The organization of replication and transcription**. *Science* 1999, **284**(5421):1790-1795.
68. Eskiw CH, Cope NF, Clay I, Schoenfelder S, Nagano T, Fraser P: **Transcription factories and nuclear organization of the genome**. *Cold Spring Harb Symp Quant Biol* 2010, **75**:501-506.
69. Chakalova L, Fraser P: **Organization of transcription**. *Cold Spring Harb Perspect Biol* 2010, **2**(9):a000729.
70. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J *et al*: **Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation**. *Cell* 2012, **148**(1-2):84-98.
71. Fullwood MJ, Ruan Y: **ChIP-based methods for the identification of long-range chromatin interactions**. *J Cell Biochem* 2009, **107**(1):30-39.
72. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F *et al*: **CTCF-mediated functional chromatin interactome in pluripotent cells**. *Nat Genet* 2011, **43**(7):630-638.
73. Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS *et al*: **Mediator and cohesin connect gene expression and chromatin architecture**. *Nature* 2010, **467**(7314):430-435.
74. Euskirchen GM, Auerbach RK, Davidov E, Gianoulis TA, Zhong G, Rozowsky J, Bhardwaj N, Gerstein MB, Snyder M: **Diverse roles and interactions of the SWI/SNF**

chromatin remodeling complex revealed using global approaches. *PLoS Genet* 2011, **7**(3):e1002008.
75. Deng B, Melnik S, Cook PR: **Transcription factories, chromatin loops, and the dysregulation of gene expression in malignancy**. *Semin Cancer Biol* 2012.
76. Spilianakis CG, Flavell RA: **Long-range intrachromosomal interactions in the T helper type 2 cytokine locus**. *Nat Immunol* 2004, **5**(10):1017-1027.
77. Tsytsykova AV, Rajsbaum R, Falvo JV, Ligeiro F, Neely SR, Goldfeld AE: **Activation-dependent intrachromosomal interactions formed by the TNF gene promoter and two distal enhancers**. *Proc Natl Acad Sci U S A* 2007, **104**(43):16850-16855.
78. Drissen R, Palstra RJ, Gillemans N, Splinter E, Grosveld F, Philipsen S, de Laat W: **The active spatial organization of the beta-globin locus requires the transcription factor EKLF**. *Genes Dev* 2004, **18**(20):2485-2490.
79. Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, Weiss MJ, Dekker J, Blobel GA: **Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1**. *Mol Cell* 2005, **17**(3):453-462.
80. Jing H, Vakoc CR, Ying L, Mandat S, Wang H, Zheng X, Blobel GA: **Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus**. *Mol Cell* 2008, **29**(2):232-242.
81. Levasseur DN, Wang J, Dorschner MO, Stamatoyannopoulos JA, Orkin SH: **Oct4 dependence of chromatin structure within the extended Nanog locus in ES cells**. *Genes Dev* 2008, **22**(5):575-580.
82. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH *et al*: **An oestrogen-receptor-alpha-bound human chromatin interactome**. *Nature* 2009, **462**(7269):58-64.
83. Pan YF, Wansa KD, Liu MH, Zhao B, Hong SZ, Tan PY, Lim KS, Bourque G, Liu ET, Cheung E: **Regulation of estrogen receptor-mediated long range transcription via evolutionarily conserved distal response elements**. *J Biol Chem* 2008, **283**(47):32977-32988.
84. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**(5235):467-470.
85. Core LJ, Waterfall JJ, Lis JT: **Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters**. *Science* 2008, **322**(5909):1845-1848.
86. Solomon MJ, Larsen PL, Varshavsky A: **Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene**. *Cell* 1988, **53**(6):937-947.
87. Dekker J, Rippe K, Dekker M, Kleckner N: **Capturing chromosome conformation**. *Science* 2002, **295**(5558):1306-1311.
88. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W: **Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)**. *Nat Genet* 2006, **38**(11):1348-1354.
89. Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U *et al*: **Circular chromosome conformation capture**

(4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* 2006, **38**(11):1341-1347.

90. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C *et al*: **Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements**. *Genome Res* 2006, **16**(10):1299-1309.
91. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO *et al*: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome**. *Science* 2009, **326**(5950):289-293.
92. Horike S, Cai S, Miyano M, Cheng JF, Kohwi-Shigematsu T: **Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome**. *Nat Genet* 2005, **37**(1):31-40.
93. Osborne CK, Schiff R, Fuqua SA, Shou J: **Estrogen receptor: current understanding of its activation and modulation**. *Clin Cancer Res* 2001, **7**(12 Suppl):4338s-4342s; discussion 4411s-4412s.
94. Deroo BJ, Korach KS: **Estrogen receptors and human disease**. *J Clin Invest* 2006, **116**(3):561-570.
95. Welboren WJ, Sweep FC, Span PN, Stunnenberg HG: **Genomic actions of estrogen receptor alpha: what are the targets and how are they regulated?** *Endocr Relat Cancer* 2009, **16**(4):1073-1089.
96. Hah N, Danko CG, Core L, Waterfall JJ, Siepel A, Lis JT, Kraus WL: **A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells**. *Cell* 2011, **145**(4):622-634.
97. Joseph R, Orlov YL, Huss M, Sun W, Kong SL, Ukil L, Pan YF, Li G, Lim M, Thomsen JS *et al*: **Integrative model of genomic factors for determining binding site selection by estrogen receptor-alpha**. *Mol Syst Biol* 2010, **6**:456.
98. Welboren WJ, van Driel MA, Janssen-Megens EM, van Heeringen SJ, Sweep FC, Span PN, Stunnenberg HG: **ChIP-Seq of ERalpha and RNA polymerase II defines genes differentially responding to ligands**. *EMBO J* 2009, **28**(10):1418-1428.
99. Schmidt D, Schwalie PC, Ross-Innes CS, Hurtado A, Brown GD, Carroll JS, Flicek P, Odom DT: **A CTCF-independent role for cohesin in tissue-specific transcription**. *Genome Res* 2010, **20**(5):578-588.
100. Hurtado A, Holmes KA, Ross-Innes CS, Schmidt D, Carroll JS: **FOXA1 is a key determinant of estrogen receptor function and endocrine response**. *Nat Genet* 2011, **43**(1):27-33.
101. Jagannathan V, Robinson-Rechavi M: **Meta-analysis of estrogen response in MCF-7 distinguishes early target genes involved in signaling and cell proliferation from later target genes involved in cell cycle and DNA repair**. *BMC Syst Biol* 2011, **5**:138.
102. Creighton CJ, Cordero KE, Larios JM, Miller RS, Johnson MD, Chinnaiyan AM, Lippman ME, Rae JM: **Genes regulated by estrogen in breast tumor cells in vitro are similarly regulated in vivo in tumor xenografts and human breast tumors**. *Genome Biol* 2006, **7**(4):R28.
103. Chang EC, Charn TH, Park SH, Helferich WG, Komm B, Katzenellenbogen JA, Katzenellenbogen BS: **Estrogen Receptors alpha and beta as determinants of gene**

expression: influence of ligand, dose, and chromatin binding. *Mol Endocrinol* 2008, **22**(5):1032-1043.

104. Fan M, Yan PS, Hartman-Frey C, Chen L, Paik H, Oyer SL, Salisbury JD, Cheng AS, Li L, Abbosh PH *et al*: **Diverse gene expression and DNA methylation profiles correlate with differential adaptation of breast cancer cells to the antiestrogens tamoxifen and fulvestrant**. *Cancer Res* 2006, **66**(24):11954-11966.
105. Kong SL, Li G, Loh SL, Sung WK, Liu ET: **Cellular reprogramming by the conjoint action of ERalpha, FOXA1, and GATA3 to a ligand-inducible growth state**. *Mol Syst Biol* 2011, **7**:526.
106. Zwart W, Theodorou V, Kok M, Canisius S, Linn S, Carroll JS: **Oestrogen receptor-co-factor-chromatin specificity in the transcriptional regulation of breast cancer**. *EMBO J* 2011, **30**(23):4764-4776.
107. Tsai WW, Wang Z, Yiu TT, Akdemir KC, Xia W, Winter S, Tsai CY, Shi X, Schwarzer D, Plunkett W *et al*: **TRIM24 links a non-canonical histone signature to breast cancer**. *Nature* 2010, **468**(7326):927-932.
108. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W *et al*: **Model-based analysis of ChIP-Seq (MACS)**. *Genome Biol* 2008, **9**(9):R137.
109. Kuttippurathu L, Hsing M, Liu Y, Schmidt B, Maskell DL, Lee K, He A, Pu WT, Kong SW: **CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments**. *Bioinformatics* 2011, **27**(5):715-717.
110. Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, Delhomme N, Ghavi-Helm Y, Wilczynski B, Riddell A, Furlong EE: **Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development**. *Nat Genet* 2012, **44**(2):148-156.
111. Coronnello C, Hartmaier R, Arora A, Huleihel L, Pandit KV, Bais AS, Butterworth M, Kaminski N, Stormo GD, Oesterreich S *et al*: **Novel modeling of combinatorial miRNA targeting identifies SNP with potential role in bone density**. *PLoS Comput Biol* 2012.
112. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2007, **35**(Database issue):D61-65.
113. Dillon N: **Gene regulation and large-scale chromatin organization in the nucleus**. *Chromosome Res* 2006, **14**(1):117-126.
114. Kocanova S, Kerr EA, Rafique S, Boyle S, Katz E, Caze-Subra S, Bickmore WA, Bystricky K: **Activation of estrogen-responsive genes does not require their nuclear co-localization**. *PLoS Genet* 2010, **6**(4):e1000922.
115. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA: **Divergent transcription from active promoters**. *Science* 2008, **322**(5909):1849-1851.
116. Seila AC, Core LJ, Lis JT, Sharp PA: **Divergent transcription: a new feature of active promoters**. *Cell Cycle* 2009, **8**(16):2557-2564.
117. Favorov A, Mularoni L, Cope LM, Medvedeva Y, Mironov AA, Makeev VJ, Wheelan SJ: **Exploring massive, genome scale datasets with the GenometriCorr package**. *PLoS Comput Biol* 2012, **8**(5):e1002529.

118. Zaret KS, Carroll JS: **Pioneer transcription factors: establishing competence for gene expression**. *Genes Dev* 2011, **25**(21):2227-2241.
119. Castellano L, Giamas G, Jacob J, Coombes RC, Lucchesi W, Thiruchelvam P, Barton G, Jiao LR, Wait R, Waxman J *et al*: **The estrogen receptor-alpha-induced microRNA signature regulates itself and its transcriptional response**. *Proc Natl Acad Sci U S A* 2009, **106**(37):15732-15737.
120. Bhat-Nakshatri P, Wang G, Collins NR, Thomson MJ, Geistlinger TR, Carroll JS, Brown M, Hammond S, Srour EF, Liu Y *et al*: **Estradiol-regulated microRNAs control estradiol response in breast cancer cells**. *Nucleic Acids Res* 2009, **37**(14):4850-4861.
121. O'Malley BW: **Molecular biology. Little molecules with big goals**. *Science* 2006, **313**(5794):1749-1750.
122. Xu J, Qiu Y, DeMayo FJ, Tsai SY, Tsai MJ, O'Malley BW: **Partial hormone resistance in mice with disruption of the steroid receptor coactivator-1 (SRC-1) gene**. *Science* 1998, **279**(5358):1922-1925.
123. Shang Y, Brown M: **Molecular determinants for the tissue specificity of SERMs**. *Science* 2002, **295**(5564):2465-2468.
124. Lonard DM, Kumar R, O'Malley BW: **Minireview: the SRC family of coactivators: an entree to understanding a subset of polygenic diseases?** *Mol Endocrinol* 2010, **24**(2):279-285.
125. Lefterova MI, Steger DJ, Zhuo D, Qatanani M, Mullican SE, Tuteja G, Manduchi E, Grant GR, Lazar MA: **Cell-specific determinants of peroxisome proliferator-activated receptor gamma function in adipocytes and macrophages**. *Mol Cell Biol* 2010, **30**(9):2078-2089.
126. McKenna NJ: **Discovery-driven research and bioinformatics in nuclear receptor and coregulator signaling**. *Biochim Biophys Acta* 2011, **1812**(8):808-817.
127. Lanz RB, Bulynko Y, Malovannaya A, Labhart P, Wang L, Li W, Qin J, Harper M, O'Malley BW: **Global characterization of transcriptional impact of the SRC-3 coregulator**. *Mol Endocrinol* 2010, **24**(4):859-872.
128. Hower V, Evans SN, Pachter L: **Shape-based peak identification for ChIP-Seq**. *BMC Bioinformatics* 2011, **12**:15.
129. Spyrou C, Stark R, Lynch AG, Tavare S: **BayesPeak: Bayesian analysis of ChIP-seq data**. *BMC Bioinformatics* 2009, **10**:299.
130. Laajala TD, Raghav S, Tuomela S, Lahesmaa R, Aittokallio T, Elo LL: **A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments**. *BMC Genomics* 2009, **10**:618.
131. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z: **Detection of functional DNA motifs via statistical over-representation**. *Nucleic Acids Res* 2004, **32**(4):1372-1381.
132. Holloway DT, Kon M, DeLisi C: **Integrating genomic data to predict transcription factor binding**. *Genome Inform* 2005, **16**(1):83-94.
133. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences**. *Genome Biol* 2010, **11**(8):R86.
134. Tillo D, Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Field Y, Lieb JD, Widom J, Segal E, Hughes TR: **High nucleosome occupancy is encoded at human regulatory sequences**. *PLoS One* 2010, **5**(2):e9129.

135. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J *et al*: **The DNA-encoded nucleosome organization of a eukaryotic genome**. *Nature* 2009, **458**(7236):362-366.
136. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol* 2004, **3**:Article3.
137. Zhu LJ, Gazin C, Lawson ND, Pages H, Lin SM, Lapointe DS, Green MR: **ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data**. *BMC Bioinformatics* 2010, **11**:237.
138. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):P3.
139. Mitchell T: **Machine Learning**: McGraw Hill.; 1997
140. Cristianini N, Shawe-Taylor J: **An Introduction to Support Vector Machines and Other Kernel-based Learning Methods**, First Edition edn. Cambridge: Cambridge University Press; 2000.
141. Breiman L: **Random forests**. *Machine Learning* 2001, **45**(1):5-32.
142. Zhu X, Goldberg A: **Introduction to semi-supervised learning**: Morgan Claypool Publishers; 2009.
143. Yarowsky D: **Unsupervised word sense disambiguation rivaling supervised methods**. In: *the 33rd Annual Meeting of the Association for Computational Linguistics: 1995*; 1995: 185-196.
144. Wilbanks EG, Facciotti MT: **Evaluation of algorithm performance in ChIP-seq peak detection**. *PLoS One* 2010, **5**(7):e11471.
145. Bair E, Tibshirani R: **Semi-supervised methods to predict patient survival from gene expression data**. *PLoS Biol* 2004, **2**(4):E108.
146. Szalkowski AM, Schmid CD: **Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts**. *Brief Bioinform* 2010.
147. **Matlab R2011b** [http://www.mathworks.com/help/toolbox/bioinfo/ref/rankfeatures.html.]
148. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies**. *J Mol Biol* 1998, **281**(5):827-842.
149. Tompa M: **An exact method for finding short motifs in sequences, with application to the ribosome binding site problem**. *Proc Int Conf Intell Syst Mol Biol* 1999:262-271.
150. Liang S: **cWINNOWER algorithm for finding fuzzy DNA motifs**. *Proc IEEE Comput Soc Bioinform Conf* 2003, **2**:260-265.
151. Liang S, Samanta MP, Biegel BA: **cWINNOWER algorithm for finding fuzzy dna motifs**. *J Bioinform Comput Biol* 2004, **2**(1):47-60.
152. Kaigler D, Wang Z, Horger K, Mooney DJ, Krebsbach PH: **VEGF scaffolds enhance angiogenesis and bone regeneration in irradiated osseous defects**. *J Bone Miner Res* 2006, **21**(5):735-744.
153. Yao Z, Lafage-Proust MH, Plouet J, Bloomfield S, Alexandre C, Vico L: **Increase of both angiogenesis and bone mass in response to exercise depends on VEGF**. *J Bone Miner Res* 2004, **19**(9):1471-1480.

154. Phillips JE, Corces VG: **CTCF: master weaver of the genome**. *Cell* 2009, **137**(7):1194-1211.