Genome-wide Analyses of Transcriptional Regulation Mediated by Estrogen Receptor alpha in Human Breast Cancer Cell Line MCF-7

INAUGURAL - DISSERTATION

to obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry and Pharmacy

of Freie Universität Berlin

By

Wei Sun 孙魏

from Jiangsu, China

March, 2012

This work has been done from October 2008 to October 2011

under the guidance of Dr. Wei Chen

from the Max Delbrück Center for Molecular Medicine

and Prof. Dr. Petra Knaus

from the Department of Biology, Chemistry and Pharmacy

of Freie Universität Berlin.

1st Reviewer: Prof. Dr. Petra Knaus

2nd Reviewer: PD Dr. Peter N. Robinson

Date of defense: May 30, 2012

Acknowledgements

Table of Contents

Abbreviation

3C	**chromosome conformation capture**
4sU	**4-thiouridine**
ACE	**angiotensin-converting enzyme**
AF	**transcriptional activation domain**
AP-1	**activator protein 1**
AUC	**area under the ROC curve**
CARM1	**coactivator associated arginine methyltransferase 1**
CCND1	**cyclin D1**
cERE$_{1/2}$	**consensus ERE$_{1/2}$**
ChIA-PET	**chromatin interaction/paired-end tag sequencing**
ChIP-on-chip	**chromatin immunoprecipitation combined with microarray**
ChIP-Seq	**chromatin immunoprecipitation and sequencing**
CS-FBS	**charcoal-dextran stripped FCS**
CTSD	**cathepsin D**
DBD	**DNA binding domain**
DBDmut	**mutated DBD**
DMF	**dimethylformamide**
E1	**estrone**
E2	**17β-estrodial**
E3	**estriol**
ERα	**estrogen receptor α**
ERβ	**estrogen receptor β**

ERE	**estrogen response element**
ERE$_{1/2}$	**half site ERE**
ERs	**estrogen receptors**
EST	**expressed sequence tags**
FKH	**forkhead motif**
FoxA1	**forkhead box A1**
GEO	**Gene Expression Omnibus**
GPR30	**G-protein coupled estrogen receptor**
GRO-Seq	**global run-on and sequencing**
HER2/NEU	**epidermal growth factor receptor 2**
H3K4me3	**trimethylation of lysine 4 on histone H3**
H3K9ac	**acetylation of lysine 9 on histone H3**
H3K9me	**methylation of lysine 9 on histone H3**
H3K27me3	**trimethylation of lysine 27 on histone H3**
HCR	**highly conserved region**
HID	**HDAC interaction domain**
HMGB1	**high-mobility group protein B1**
LBD	**ligand binding domain**
lincRNAs	**Large intergenic noncoding RNAs**
MACS	**Model-based Analysis of ChIP-Seq**
NCoA3	**nuclear receptor coactivator 3**
RNA-Seq	**mRNA sequencing**
PAH	**paired amphipathic α-helices**
PAX2	**paired box 2 gene product**

P.I.	**proteinase inhibitors**
PR	**progesterone receptor**
PRMT	**protein arginine methyltransferase**
PWM	**position weight matrix**
qPCR	**real time quantitative PCR**
RACE	**rapid amplification of cDNA ends**
RARα1	**retinoic acid receptor α1**
RAS	**renin angiotensin system**
ROC	**receiver operating characteristic curve**
SID	**Sin3 interaction domain**
Sp1	**specificity protein 1**
SRC	**steroid receptor coactivator**
TF	**transcription factor**
TSSs	**transcription start sites**

1. Abstract

The steroid hormone estrogen plays a central role in gene regulation in hormone-responsive breast cancer cells, via estrogen receptors (ERs), of which ERα is one of the most important isoforms. ERα upon activation by E2 could rapidly bind to thousands of cognate genomic sites, recruit coregulators, chromatin-modifiers and the basal transcriptional machinery, and thereby ultimately modulate target gene expression.

To uncover the mechanisms by which ERα dictates gene expression, we profiled both the ERα binding sites and gene expression changes induced by E2 stimulation on a genome-wide scale. Here, using our own ChIP-Seq results, complemented by previous published data, we obtained a list of 4543 high-confident ERα binding sites at 45 minutes after E2 treatment, of which the majority are at non-promoter distal regions. Interestingly, in contrast to most previous studies, we observed a significant number of ERα binding events in the absence of E2. To determine the transcriptional response induced by E2 stimulation, we applied the state-of-the-art RNA-Seq method. More specifically, to separate the transcriptional regulation from post-transcriptional events, we measure the gene expression based on both bulk RNAs and RNAs newly synthesized at different time intervals after E2 treatment. In total, 722 genes were identified to be differentially regulated, which could be categorized into six groups with different response dynamics pattern. Further analysis on the relationships between ERα binding sites and differentially expressed genes demonstrated that the upregulated genes have significantly higher proportion of promoter proximal ERα binding sites than the downregulated ones. Among the upregulated genes, those with expression changes at early time points show the highest enrichment of ERα binding at the promoter proximal or distal regions. Taken together, the direct transcriptional response mediated by ERα seems largely to be rapid transcriptional induction.

2. Introduction

2.1. Estrogen, Estrogen Receptor and Breast Cancer

2.1.1. Estrogen and Breast Cancer

In women, estrogens are produced mainly in the ovaries and in the placenta during pregnancy. Basically, estrogens have three different bioactive forms in human (**Figure 1**) which dominate in different developmental stages, temporally, such as 17β-estrodial (E2) in non-pregnant women and estriol (E3) in pregnancy, estrone (E1) secreted after menopause (Zhao et al., 2010).

Estrone (E1)   Estrodial (E2)   Estriol (E3)
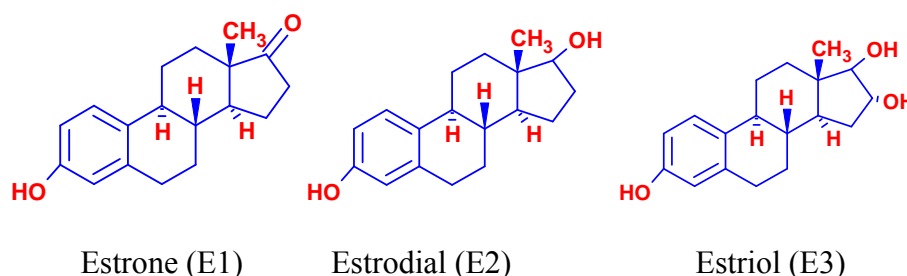
Figure 1. Chemical structural illustration of different active forms of estrogens.

Estrogens can induce a series of cellular responses within a wide variety of target tissues during physiological processes such as growth, development, and differentiation. In mammals, important estrogen actions are predominantly implicated in organs including brain, liver, vasculature, bone, and particularly mammary gland (Katzenellenbogen et al., 2000). Other than the physiological functions of estrogens in maintaining homeostasis of their target organs and tissues, they are also involved into the development and progression of breast cancer where estrogens, mitogen-like, promote cancerous cell cycle progression and cell proliferation. More than one century ago, researchers have begun to pay attention to the relationship of estrogens and breast carcinogenesis. In 1896, George Thomas Beatson for the first time described that the removal of ovaries from two young women with advanced breast cancer could remarkably induce breast cancer remission. This pioneer work had set up the link between hormones and breast cancer, and opened up a totally new field in which researchers have ever since pursued to elucidate the relationship between estrogens and breast cancer. From then on, increasingly accumulated data from clinical and animal studies have shown the casual relationship of estrogens and breast cancer, and then two hypotheses were proposed to

explain this relationship. First, estrogen-induced high cell proliferation accompanied by intensive DNA synthesis can produce high frequency of replication errors which can further disrupt vital physiological processes such as apoptosis, DNA repair. Second, the production of genotoxic byproducts by estrogen metabolism could directly damage genomic DNA, resulting in mutations in crucial genes with activation of oncogenes and repression of tumor suppressor genes (Deroo and Korach, 2006). The precise molecular mechanisms underlying the casual effects of estrogens on breast cancer, however, remain not fully understood till now.

Since breast cancer is the second most common cancer in women with an estimated 254,650 new cases of breast cancer and with 40,170 women expected to die from breast cancer in 2009 within USA only (American Cancer Society. Breast Cancer Facts & Figures 2009-2010), the prevention and clinical management of breast cancer have posed great challenges to global basic and clinical research society. To meet those challenges, the premise is that we should have a good understanding of the molecular characteristics of genesis and development of breast cancer. Pathophysiologically, breast cancer mostly originates from epithelial layer of lobules or ducts. According to tissue confinement and invasiveness extent, breast cancer can simply be defined as invasive breast cancer or in situ subtype. Based on whether the growth is dependent on estrogens, breast cancer can be defined as hormone-dependent or hormone-independent with different therapeutic regimens. The former subtype consists of approximate 80% of all cases.

The endocrine therapy, in which actions of estrogens are inhibited or blocked by small chemical molecules, has greatly benefited patients with hormone-dependent breast cancers. During the 1970s, the discovery of anti-estrogenic chemical, tamoxifen, to date the most successful targeted cancer therapeutic drug, is regarded as the cornerstone in breast cancer treatment. Tamoxifen (**Figure 2**), has been shown to physically impede the interaction of estrogen and its receptor with high efficacy, inhibiting estrogen-estrogen receptor-mediated gene transcription events that are thought to result in cell proliferation (Green and Carroll, 2007). The application of tamoxifen in prevention of breast cancer have decreased breast cancer incidence by approximately 50% (Fisher et al., 2005). Notably, in ~70% breast cancers expressing ERα, endocrine therapy with the aim to modulate the transcriptional activities of ERα using tamoxifen have greatly improved the recurrence-free survival in patients (Hortobagyi et al., 1998; Sorlie et al., 2001 and 2003). However, with great therapeutic

potentials, an average time period of only 15 months of tamoxifen usage may inevitably give rise to tamoxifen-resistance in 80% of patients with breast cancer (Clarke et al. 2001). Till now, the molecular mechanism by which breast cancers can circumvent tamoxifen-mediated proliferation inhibition is still not clear. Several lines of evidence have reported that mutations and modifications of ERα, even the amplification of 6q25 genomic regions encompassing the ERα (*ESR1*) could not account for tamoxifen resistance in some breast cancers (Hopp and Fuqua, 1998; Holst et al., 2007). The finding that ERBB2 repressed by ERα-PAX2 complex in luminal breast cancer can be reactivated by ERα-AIB1/SRC3 complex in tamoxifen resistant breast cancer has provided mechanistic insights into the molecular basis of endocrine resistance in breast cancer (Hurtato et al., 2008). Therefore, the prerequisite in further understanding of the molecular mechanisms of tamoxifen resistance seems to rely on the well appreciation of the estrogen-ERα or tamoxifen-ERα mediated gene regulation on the transcriptional level.
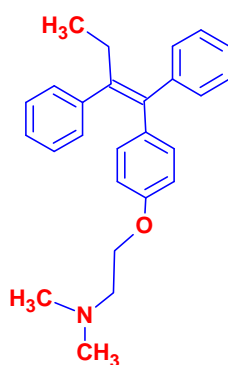
Figure 2. Chemical structural presentation of estrogen receptor antagonist (tamoxifen).

2.1.2. ERα: Structure and Functions

Estrogen actions are mainly mediated through estrogen receptors (ERs), members of protein superfamily of steroid hormone nuclear receptors (Mangelsdorf et al., 1995; Pettersson et al., 2001). ERs are ligand-activated transcription factors, and play a crucial role in both physiological and pathological processes in breast tumors. To date, there are two ER isoforms in human, ERα and ERβ, which show distinct tissue-specific expression patterns and biological functions (Paech et al., 1997). Here in this study, we mainly focused on characterizing the involvement of ERα in transcriptional regulation, which is believed to be the driving force for cell proliferation and growth of ERα positive breast cancer.

ERα was initially discovered by Elwood V. Jensen in the 1960s, and first cloned in 1986 from MCF-7 human breast cancer cells (Green et al., 1986). ERα, similar to other hormone-responsive nuclear receptors, contains five well-conserved functional domains A-F (**Figure 3**). The N-terminal of ERα is A/B region, in which transcriptional activation domain 1 (AF-1) locates. A DNA binding domain (DBD) resides in the C region and consists of two non-equivalent cysteine-rich zinc fingers (Green et al., 1988; Ruff et al., 2000). Of the two zinc fingers, the N-terminal one can bind to estrogen response elements (EREs; 5'-GGTCA-NNN-TGACC-3', where N is any nucleotide), and the second one may function in stabilizing non-specific protein-DNA interactions. Genome-wide analysis on gene regulation by mutated DNA binding domain (DBDmut) or wild-type ERα has revealed that direct ERE binding is required for most of (75%) estrogen-dependent gene regulation and 90% of hormone-dependent recruitment of ERα to genomic binding sites (Stender et al., 2010). Ligand binding domain (LBD) embeds in region E, which together with region F form the second transcriptional activation domain, AF-2. The functionality of Region D as the hinge domain is poorly characterized.
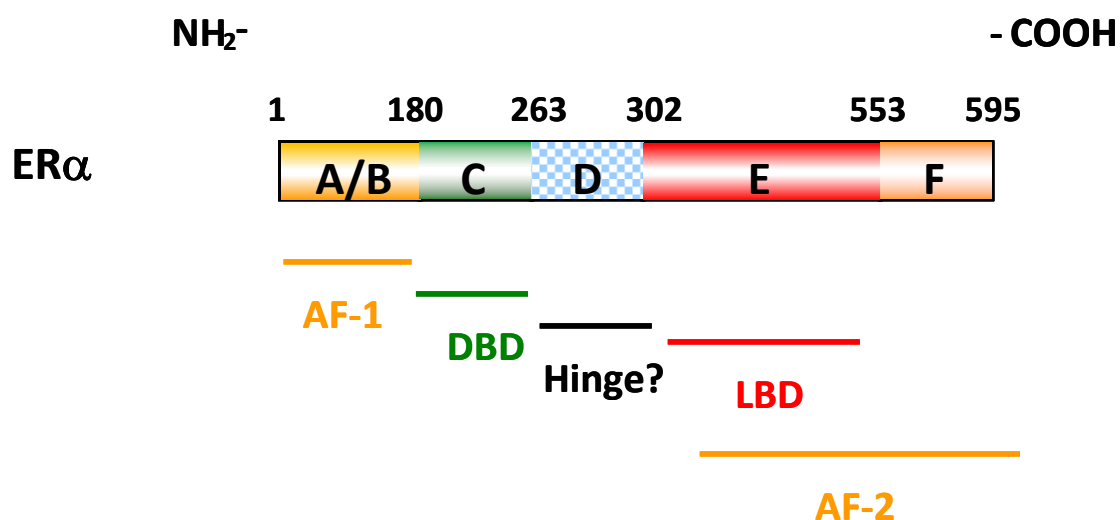
Figure 3. Schematic representations of structural and functional domains in ERα. ERα contains several conserved domains: double zinc finger DBD that mediates dimerization and DNA binding, a ligand binding domain (LBD), and two transcriptional activation functions (AF-1 and AF-2).

ERα is expressed in many human organs, e.g., breast, uterus, urogenital tract, liver, bone, gastrointestinal, cardiovascular system and brain (Pearce et al., 2004). In clinical pathology, ERα together with progesterone receptor (PR) and epidermal growth factor receptor 2

(HER2/NEU) have been used as markers to discern molecular subtypes of breast cancer and predict prognosis in patients with breast cancers. To facilitate the breast cancer research, ERα-positive breast cancer cell lines such as MCF-7, T-47D, and ZR-75-1 etc. are good model systems most widely applied in studying molecular mechanism underlying the involvement of ERα in breast cancer. Especially, the use of MCF-7 breast cancer cell line after being developed by Soule et al. in 1973 has greatly speeded up the research into breast cancer biology (Soule et al., 1973).

It is believed that ERα locates largely in the cytoplasm and translocates into the nucleus once upon stimulation (Renoir et al., 1990). In an ERα positive breast cancer cell, inactive ERα in the cytosol is activated by estrogens through binding to its LBD and then translocates into the nucleus (**Figure 4**). Upon binding, E2-ERα complex dimerizes, changes its conformation and then modulates gene expression through binding to estrogen response elements (EREs) or other *cis*-regulatory elements on target genes (Nilsson et al., 2001). EREs, which are specific, palindromic sequences, are recognized and bound by ERα for initiating subsequent target gene transcriptional regulation. Intriguingly, non-canonical EREs including ERE half sites ($ERE_{1/2}$) have once been regarded as weak binding sites for ERα, however, Joshi and colleagues reported that ERα actually binds strongly to these non-canonical EREs, and the binding can be even enhanced in the presence of high-mobility group protein B1 (HMGB1) (Joshi et al., 2011). Also, they demonstrated that minimal target site for ERα is a consensus $ERE_{1/2}$ ($cERE_{1/2}$) that holds true in many different specific cellular contexts, and that HMGB1 enhances both the binding affinity and transcriptional activity of ERα (Joshi et al., 2011). Those results supported the notion that ERα-ERE interactions have previously unexpected plasticity. Apart from directly binding on genomes, ERα can also interact with DNA via tethering to cooperating transcription factors or coregulators such as activator protein 1 (AP-1), specificity protein 1 (Sp1), NF-κB, forkhead box A1 (FoxA1) and Runx1 (Carroll et al., 2005; Stender et al., 2010). In either case, ERα both needs the recruitment of coregulators and components of the basal transcriptional machinery to either increase or decrease target gene transcription levels, ultimately eliciting a physiological response (Deroo and Korach, 2006; Glass et al., 2000).
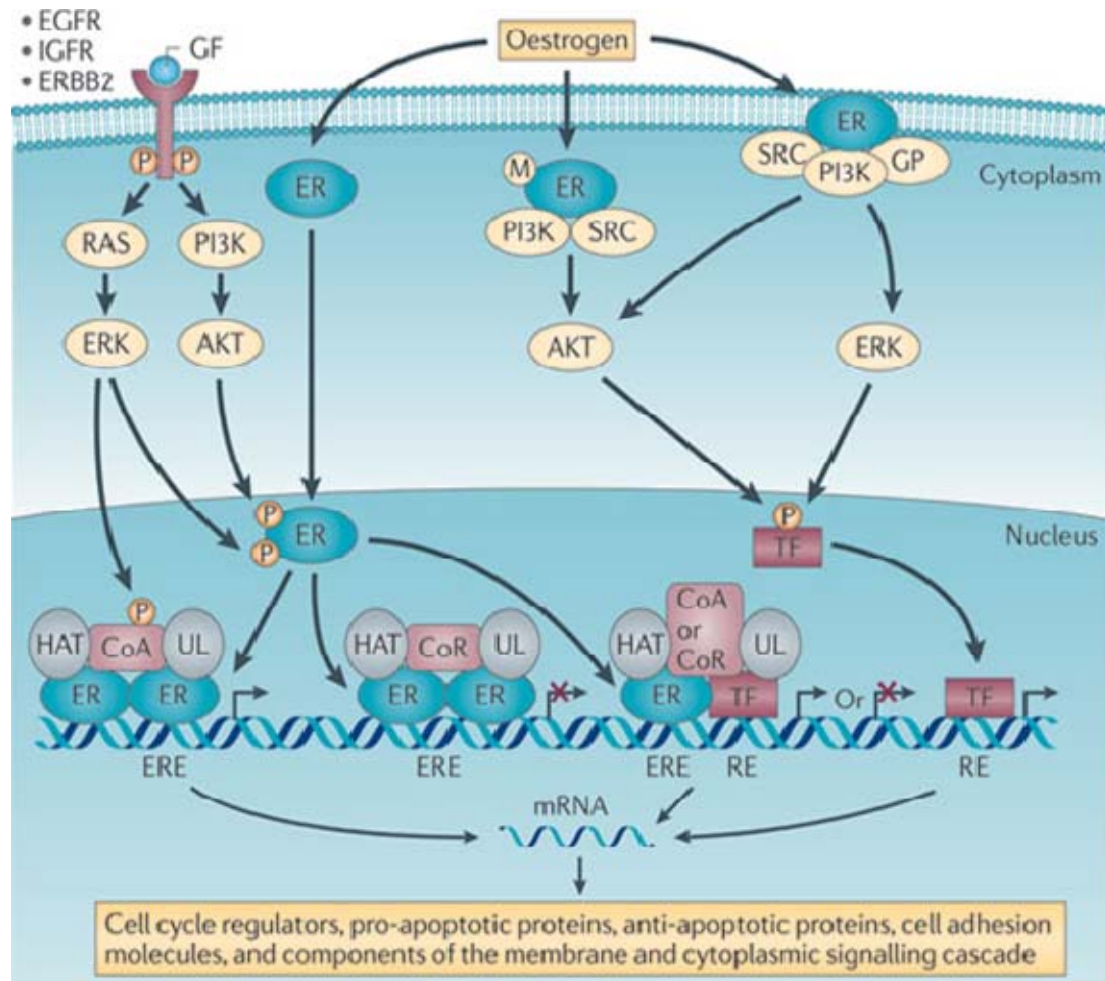
Figure 4. Molecular mechanisms of ERα action. Following the classic ERα genomic action, E2 binds to ERα, and then the E2-ERα complexes dimerize and bind to DNA at sequence-specific elements, estrogen response elements (EREs). By interacting with coactivator (CoA) or corepressor (CoR) complexes at these elements, the ERα can either activate or to repress target gene transcription. The complexes are composed of the active (phosphorylated (P)) coactivator or corepressor, histone acetyltransferase (HAT), ubiquitin ligase (UL). Alternatively, ERα can regulate gene expression via tethering to other direct DNA-binding transcription factors (TFs). As such, ERα can bind to EREs that are near the response elements (REs) of the cooperating TFs, or ERα can indirectly interact with DNA via tethering to TFs. In addition, E2-activated ERα localized in the cytoplasm or at the cell membrane (where ERα can be methylated (M)) can interact with SRC, PI3K and G proteins (GPs) and elicit non-genomic signaling. This signaling pathway mediated by protein kinase cascades results in the phosphorylation and activation of target TFs, which can thereby regulate target gene transcription through binding to RE sites. Lastly, growth factor receptors, such as epidermal growth factor receptor (EGFR), ERBB2 (also known as HER2/NEU) and insulin-like growth factor receptor (IGFR) after growth factors (GFs) stimulation can activate ERK and AKT kinases, which can then phosphorylate and activate ERα in an E2-independent manner. Figure is adapted from Thomas et al., 2011.

2.2. Genomic Actions of Estrogen-ERα Complex

2.2.1. Characterization of ERα Cognate Genomic Binding Sites

The canonical ERα DNA binding motifs, EREs, have long been observed within 1Kb of 5' flanking regions of many target genes such as *TFF1* (*pS2*), *EBAG9*, and *cathepsin D* (*CTSD*) (Augereau et al., 1994; Berry et al., 1989; Ikeda et al., 2000). Surprisingly, although the consensus ERE sequence has been delineated in 1989, there were only dozens of natural EREs for which both ERα binding affinity and transcriptional activation have been examined before the genome-wide approaches emerged. Before experimental global profiling ERα binding sites, Bourdeau and colleagues employed a computational approach to comprehensively search for EREs resided in the human genome, and finally identified 71,119 EREs (about one in every 43 kb of genomic DNA) (Bourdeau et al., 2004). Unexpectedly, the subsequent global mapping of ERα binding sites within human genome only found several thousands of cognate binding sites, far less than the computational predictions. It suggests that the selectivity and specificity of ERα binding to genomic sites are rather perplexing and must be delicately tuned by a number of regulatory mechanisms to orchestrate target gene transcriptional regulation. In addition, ERE motif search among several thousands of ERα cognate binding sites reveals that ERE motifs only partially account for, the binding, and some other motifs such as those bound by AP-1, Sp-1 and FoxA1 may also play crucial roles in guidance of ERα binding on genomic regions (Carroll et al., 2006). More recently, Hurtado and colleagues have further demonstrated that FoxA1 functions as a key determinant and a pioneer factor in modulation of gene transcription mediated by ERα in breast cancer cells (Hurtado et al., 2011).

Previous research on identifying essential elements for ERα-mediated transcriptional regulation has focused primarily on promoter proximal regions. However, using different several studies on mapping ERα cognate binding sites in MCF-7 breast cancer cells had unveiled that promoter proximal regions only account for less than 10% of all binding sites (Carroll et al., 2005 and 2006; Cicatiello et al., 2010; Hurtado et al., 2008; Lin et al., 2007). Therefore, ERα is preferentially recruited to distal *cis*-regulatory elements. Several lines of evidence have showed that some distal elements even more than 100kb far away from target genes could be brought into close proximity to promoters, thereby to modulate transcription of target genes via long-range chromatin looping (Fullwood et al., 2009).

In nucleus, chromosomes form complex three-dimensional structures, so that DNA elements even separated by long distance could interact physically with functional relevance. To detect long-range chromatin interactions between DNA elements engaged in transcriptional regulation, chromosome conformation capture (3C), has been intensively used. However, due to its limitation in throughput, it cannot be used for studying genome-wide chromatin interactions. To meet the needs, more recently, a method called "chromatin interaction analysis with paired-end tag sequencing (ChIA-PET)" has been developed for capturing global chromatin interaction networks mediated by transcription factors (Fullwood et al., 2009). Using this technology, Ruan and his colleagues comprehensively mapped the chromatin interaction mediated by ERα in MCF-7 cells, revealing that most high-confidence distal ERα binding sites were anchored at gene promoters via long-range DNA looping. They further showed that ERα modulated some gene transcription collectively through long-range chromatin interactions, setting up the role of chromatin interactions in mammalian gene transcription regulation. The resource provided by this study helped to decipher transcription regulation by ERα via distal (non-promoter) regulatory DNA elements.

2.2.2. Transcriptional Regulation by Interplays between ERα, Associated-Coregulators and Cooperating Transcription Factors (TFs), Histone Modifications

2.2.2.1. ERα, Associated Coregulators and Cooperating TFs

Estrogens drive a spectrum of gene regulatory responses which rely on the highly coordinated interactions between ERα, associated-coregulators and cooperating TFs. ERα associated-coregulators and cooperating TFs can be further divided into two functionally opposed groups, coactivators and corepressors (**Figure 4**). Coactivators are essential for gene transcriptional activations, whereas corepressors are required for dampening the transcription. There are two ways probably utilized by coactivators in prompting gene transcription. First, after interacting with E2-ERα complex, they can recruit basal transcriptional machinery, which then are responsible for subsequent target gene transcription. Another alternative way is that some coactivators harboring enzymatic activities could potentiate the activity of basal transcriptional machinery. For example, some coactivators with histone acetyltransferase activity can unwind the supercoil structure of regulatory DNA regions at the target gene promoters, resulting in a permissive chromatin context for target gene transcription (Shibata et al., 1997). Some studies also revealed that, to a certain extent, the presence and protein levels

of some coregulators can modulate and finally determine ERα-mediated transcription of target genes.

Protein-protein interaction studies have already identified dozens of ERα associated-coregulators. Among them, the best characterized are the p160 family of ERα coactivators, including steroid receptor coactivator 1 (SRC-1), SRC-2 (TIF2/GRIP1), and SRC-3 (AIB1). ChIP-cloning (Labhart et al., 2005) and recently ChIP-on-chip studies (Hurtado et al., 2008) have demonstrated a tight correlation of genomic binding between SRC coactivators and ERα. Hurtado et al. have showed that the ERα coactivator SRC-3/AIB-1 and the paired box 2 gene product (PAX2) can compete for ERα binding for regulation of ERBB2 transcription, and ER-positive luminal-like breast tumors can switch to tamoxifen-resistant breast tumors by circumventing repression of ERBB2 transcription mediated by ERα-PAX2. Several pieces of evidence have shown that the interactions between the LxxLL motif of SRC1 and the AF-2 domain of ERα can induce conformational changes of ERα, allowing for its subsequent interactions with other coregulators and downstream transcriptional machinery. One important function of the SRCs as coactivators is to recruit other chromatin modifier proteins such as CBP/p300 and pCAF, which have HAT (histone acetyltransferase) activities and can form permissive chromatin structure for transcription. Another group of factors recruited by the p160 family are the arginine methyltransferase (PRMT) family that methylates N-terminal tails of histone H3 and H4. The first discovered PRMT family member, coactivator associated arginine methyltransferase 1 (CARM1 or PRMT4), enhances ERα-mediated transcriptional activity by interacting with GRIP1, a p160 family cofactor and p300 (Chen et al., 2000).

To systematically identify coregulators involved in ERα-mediated transcription regulation network, investigators have also applied bioinformatic approaches to search for binding motifs of other TFs in ERα binding sites and identified some previously characterized TFs such as AP-1 and Sp1, as well as some *de novo* coregulators including Oct, FoxA1, GATA3, PAX2, AP-2γ**,** LEF-1 and Nkx3-1 (Carroll et al., 2005 and 2006; Green and Carroll, 2007; Holmes et al., 2008; Hurtado et al., 2008; Laganiere et al., 2005; Tanet al., 2011; Wang et al., 2007). The DNA binding motifs of those TFs were found to significantly enrich near the center of ERα binding sites. This finding is in concordance with ERE-independent ERα binding, demonstrating that ERα could bind DNA via tethering to other transcription factors,

such as AP-1, to exert transcriptional activation on genes without EREs at their regulatory regions.

The diversity of ERα coregulators in ERα mediated transcription of target genes has been elucidated in many studies. AP-1 together with ERα can regulate *pS2* gene regulation. In the promoter of *pS2,* AP-1 motif and an ERE motif were in proximity (Barkhem et al., 2002). Similarly, Sp-1 motif and ERE were both present on retinoic acid receptor α1 (*RARα1*) gene promoter (Sun et al., 1998). Oct1 co-recruited with ERα can regulate the transcription of crucial cell cycle gene *CCND1* in breast cancer cells (Carroll et al., 2006; Cicatiello et al., 2004). Hurtado et al. implicated that PAX2, as a transcriptional repressor, along with ERα were recruited to the *cis*-regulatory element of *ERBB2*, repressing *ERBB2* transcription in ERα-positive breast cancer cells (Hurtado et al., 2008). Motifs of FoxA1 were significantly enriched in ERα binding sites, which implicate its close association with ERα (Carroll et al., 2005 and 2006). More recently, Hurtado et al. further demonstrated mechanistically that FoxA1, as a key determinant, can mediate differential interactions between ERα and corresponding chromatin. As a pioneer factor for ERα tethering to chromatin, nearly all ERα-chromatin interactions and subsequent gene expression changes highly relied on the FoxA1 function. Most recently, Tan S. K. et al. demonstrated that AP-2γ, which has been implicated in breast cancer tumorigenesis, is another novel collaborative factor involved in ERα-mediated transcription. They also found that there is functional interplay between AP-2γ and FoxA1 in coordinating ERα-mediated transcriptional regulation (Tan et al., 2011). LEF-1 and Nkx3-1 motifs are also enriched in ERα binding sites in breast cancer cells. However, other than being co-recruited with ERα in response to E2 stimulation, LEF-1 and Nkx3-1 binds to genome in unstimulated circumstances, and can abolish cell proliferation by competing ERα binding to *cis*-regulatory DNA elements within genome after E2 stimulation. Therefore, LEF-1 and Nkx3-1 are antagonists of ERα-mediated transcription (Holmes et al., 2008). Collectively, those studies showed that comprehensive analysis of ERα binding sites can greatly help to unwind the perplexing interactome between ERα and its coregulators.

2.2.2.2. ERα and Histone Modifications

Genome-wide studies have shown that chromatin structure and epigenetic signatures of *cis*-regulatory DNA elements such as promoters and enhancers, even gene bodies may have a determining role in local transcriptional activity (Barski et al., 2007). It has been shown that

trimethylation of lysine 4 on histone H3 (H3K4me3) was commonly found at promoter regions (Heintzman et al., 2007). And the H3K4me3 enriched regions showed DNase I hypersensitivity, a marker of active regulatory regions for transcription (Xi et al., 2007). It is now commonly believed that H3K4me3 residing in gene promoters is a predictive of active transcription. On the contrary, trimethylation of lysine 27 on histone H3 (H3K27me3) locating within gene promoters is associated with repressed gene expression. The combination pattern of H3K4me3, H3K27me3 and some other epigenetic signatures can determine the cellular fates in lineage differentiation. For instance, the bivalent pattern of H3K4me3 and H3K27me3 exists in embryonic stem cells and loss of one of the two modifications can be observed in differentiated cells. Several studies of histone modifications in cancers have shown that histone modifications are often altered during tumorigenesis (Fraga et al., 2005; Seligson et al., 2005).

In ERα-positive breast cancer cells, epigenetic signatures such as histone acetylation, phosphorylation and methylation can be formed and erased via ERα cofactors with chromatin modifying enzymatic activities. Kwon and colleagues showed that acetylation of lysine 9 on histone H3 (H3K9ac), a marker for open chromatin structure, could be observed on the promoters and enhancers of the active transcribed pS2/TFF1and GREB1 genes upon E2 induction. On the contrary, H3K9 methylation (H3K9me) could inhibit native ERα binding to chromatin, but E2-ERα complex can recruit LSD1 to demethylate H3K9 for a permissive chromatin setting in the following transcription programs (Kwon et al., 2007).

2.2.3. Global Profiling of Gene Expression Mediated by Estrogen-ERα Complex

Comprehensive delineation of E2-regulated gene expression in breast cancer cells can reveal pivotal genes in transcriptional regulatory networks, and shed lights on potential therapeutic targets in ERα-positive breast cancers (Cicatiello et al., 2010; Frasor et al., 2003; Hah et al., 2011; Lin et al., 2007; Stender et al., 2007). Frasor et al. depicted that the ERα up-regulated a set of positive regulators of cell proliferation as well as down-regulated another set of repressor of cell proliferation, highlighting that E2 stimulates breast cancer cells through diverse metabolic and cell cycle pathways. Cicatiello et al. described the kinetics of E2-regulated gene expression in ERα-positive MCF-7 and ZR-75 cells and correlated global ERα binding sites with gene expression changes (using mRNA sequencing and microRNA sequencing technologies). In the end, 33 high concordant gene clusters from the whole set of

E2-regulated genes were defined, which included some previously well-characterized transcription factors and could be organized in a gene transcriptional regulatory cascade (Cicatiello et al., 2010).

Most of gene expressions profiling studies were performed after a relatively long E2 treatment (more than 45 min); therefore, the identified gene expression changes comprised both primary and secondary effects. Recently, Hah et al. applied global run-on and sequencing (GRO-Seq) to detect immediate effects of estrogen signaling on the transcriptome of breast cancer cells. Unexpected, a large proportion of transcripts are regulated in a rapid, transient manner after E2 treatment. To date, this resource is the most comprehensive analysis on primary and transient transcriptome changes during E2 signaling (Hah et al., 2011).

2.3. E2 Target Genes and Their Involvements in Breast Cancer

Given the important role that ERα played in pathology of breast cancer, it is expected that ERα regulated genes were often found to be involved in cell proliferation, migration and etc. In 1984, using in vitro nuclear run-on transcription assay, Brown et al. showed that the expression of *TFF1* was directly controlled by E2 at the transcriptional level (Brown et al., 1984). Later, Prest et al. reported that TFF1 can stimulate the migration of breast cancer cells and then facilitate distant metastasis (Prest et al., 2002). The regulation of cell cycle is strictly controlled by E2 in breast cancer cells. Using differential cloning, the important cell cycle regulators c-Myc and cyclin D1 (*CCND1*) were identified as E2 target genes, which directly connects cell proliferation with E2 signaling in breast cancer cells (Dubik et al., 1987; Altucci et al., 1996). E2 could also regulate cell proliferation via a variety of important cell signaling pathways. For example, the work by Lin ZH et al suggested Wnt11, known to induce cell proliferation and cell progression of prostate cancer, as an E2 target gene. Wnt11 signaling pathway is up-regulated after E2 induction and could promote breast cancer cell growth (Lin ZH et al., 2007). Important TFs are also targets of E2 signaling in breast cancer cells. For instances, AP-2α and AP-2γ are the primary isoforms of the AP-2 family, and are involved in transrepression of *ERBB2* mediated by E2 (Periss et al., 2000). Several recent studies have demonstrated the upregulation of AP-2γ and downregulation of AP-2α after E2 exposure. Of note, AP-2γ, contrary to AP-2α, modulates genes responsible for cell cycle progression, promotes breast cancer cell proliferation.

Interestingly, E2 can also regulate tumor suppressor genes. For example, ATBF1, a tumor suppressor gene, could inhibit ERα induced cell proliferation. Nevertheless, Dong X.Y. et al. have shown that ATBF1 gene expression was upregulated through ERα binding on its promoter harboring an $ERE_{1/2}$. E2 at lower levels increased ATBF1 protein levels, which in turn inhibited cell proliferation induced by E2. Therefore, an auto-regulatory feedback loop exists between ATBF1 and E2-ERα signaling, ATBF1 appeared to play an important role in maintaining breast epithelial cell homeostasis (Dong et al., 2011). Another potent tumor suppressor, the cyclin-dependent kinase inhibitor protein p21 (Waf1/Cip1), could also be induced by E2 in very short treatment time in MCF-7 cells. In a study performed by Mandal et al., they demonstrated that ERα is recruited to the $ERE_{1/2}$/AP-1 sites upstream of the p21 (Waf1/Cip1) promoter (Mandal et al., 2010).

Chromatin modifying enzymes are also shown to be regulated by E2 in breast cancer cells. For example, histone demethylase JMJD2B is highly expressed in ERα-positive primary breast tumors and is critical to breast cancer cell survival under normal or hypoxia conditions. Several studies have shown that *JMJD2B,* with a functional ERα binding site in the first intron, can be regulated by E2-ERα complex (Carroll et al., 2006; Lin et al., 2007; Yang et al., 2010).

3. Aims

Estrogens via ERα has been known to promote cell proliferation and tightly related to the development of breast cancer. The cellular level of ERα commits also a strong clinic impact on the breast cancer therapy and recurrence. The ERα-mediated transcriptional regulation is thought to be initiated via binding to the E2 ligand, translocating to the cell nucleus, and subsequential binding of the E2-ERα complex to the DNA thereby up- and down-regulate its target genes. A number of cofactors as well as histone modifiers are also known to be involved in this regulation process. It is natural that the previous efforts to uncover the ERα transcriptional regulatory mechanism(s) have being focusing on ERα-DNA binding properties and gene expression changes upon E2 stimulation. The recent genome-wide studies on both aspects have largely extended our understanding of ERα transcriptional networks. The findings from different groups, however, often showed limited overlaps, even from the same cellular system (see discussion part of this thesis). Therefore, we thought to firstly create our own data on the two central aspects of ERα network, e.g. global ERα binding sites and gene expression changes, by using sequencing-based approaches, i.e. ChIP-Seq and RNA-Seq. To address the functional relationship of histone modifiers to ERα, we also performed ChIP-seq on H3K4me3, a marker at the promoters of active transcribed genes, which in return would also work as an additional evidence for novel transcripts we identified based on RNA-Seq data. In the scheme of RNA-Seq, we particularly sequenced the 4sU-labeled newly synthesized RNAs in order to exclude the influence of preexisting mRNAs and thereafter to achieve a higher sensitivity to detect ERα target genes. Moreover, with the aim to monitor the regulation dynamics of ERα, all the above-mentioned ChIP-Seq and RNA-Seq experiments were performed at multiple time points after E2 treatment. Finally, the large resources of experimental data were integrated and multiple bioinformatics tools were applied in order to answer the following questions:

1. What is the global DNA-binding property (locations, motifs, etc.) of ERα with and without E2 stimulation, and what is the binding dynamics over the extension of E2 treatment?

2. Does the binding of ERα influence the histone signatures, in particular H3K4me3, in order to fulfill immediate transcriptional regulation?

3. What are the possible regulation mechanisms of ERα binding sites, e.g. promoter- or enhancer-driven up-regulation, or cofactor-mediated down-regulation, etc.?

4. How can our results be compared to the published data, and, with the advantage of our sequencing tools, can we identify novel ERα-regulated genes (both known genes and previously un-annotated transcripts)?

With these aims, we focus our study on global binding of ERα and its early and direct gene targets; putting together, we hope to gain a deep and multi-dimensional understanding of ERα-mediated transcriptional regulation and in the meantime search for novel genes/ transcripts with functional relevance to breast cancer as well as other estrogen-responsive systems.

4. Materials and Methods

4.1. Materials

4.1.1. Chemicals and Reagents

All the chemicals are bought from ROTH (Germany) and Sigma (Germany) unless otherwise specified.

All the reagents used in this study have been described in the method section.

4.1.2. Kits

ChIP-Seq DNA Sample Prep. Kit	Illumina	Germnay
EZ-Magna ChIP™ A Chromatin Immunoprecipitation Kit	Millipore	Germany
EZ-Magna ChIP™ G Chromatin Immunoprecipitation Kit	Millipore	Germany
Histone ChIP kit	Diagenode	Belgium
NEB Next™ mRNA Sample Prep. Kit	NEB	Germany
NEB Next™ DNA Sample Prep. Kit	NEB	Germany
Phase-Lock gel kit	5 primer	Germany
QIAGEN MinElute PCR Purification and Gel Extraction Kits	Qiagen	Germany
QIAGEN QIAquick PCR Purification and Gel Extraction Kits	Qiagen	Germany
SYBR® Green PCR Master Mix	Applied Biosystems	UK
SuperScript® III First-Strand Synthesis System for RT-PCR	Invitrogen	Germany
Transcription ChIP kit	Diagenode	Belgium

4.1.3. Labware Consumables

Cell scraper	Greiner Bio-One	Germany
15cm cell culture dish	Greiner Bio-one	Germany
1.5ml centrifuge tube	Eppendorf	Germany
1.5ml low binding tube	Eppendorf	Germany
2.0ml centrifuge tube	Eppendorf	Germany
2.0ml low binding tube	Eppendorf	Germany
200μl PCR tube	Eppendorf	Germany
5ml Sarstedt serological pipette	Sarstedt	Germany
10ml Sarstedt serological pipette	Sarstedt	Germany
25ml Sarstedt serological pipette	Sarstedt	Germany
50ml Sarstedt serological pipette	Sarstedt	Germany
15ml conical falcon tube	BD Bioscience	Germany
50ml conical falcon tube	BD Bioscience	Germany

4.1.4. Instruments

BINDER CO_2 Incubators	BINDER	Germany
Bio-Rad electrophoresis system	Bio-Rad	Germany
Bio-Rad Tetrad® 2 Peltier Thermo Cycler	Bio-Rad	Germany
Centrifuge 5424	Eppendorf	Germany
Centrifuge 5417R	Eppendorf	Germany
CryoPlus Liquid Nitrogen Storage	Thermo Scientific	Germany
GeneAMP® PCR System 9700	Applied Biosystems	Germany
Gyro-rocker (SSL3)	Stuart	Germany
HERAEUS multifuge 1s-R centrifuge	Thermo Scientific	Germany
Laminar flow hood	BDK	Germany
Table mini-microcentrifuge	ROTH	Germany
Microwave	Privileg	Germany
Minispin plus	Eppendorf	Germany
PEQLAB electrophoresis system	PEQLAB	Germany
PEQLAB NanoDrop Spectrophotometer ND-1000	PEQLAB	Germany
PIPETBOY acu	Integra Biosciences	Germany
Qubit Fluorometer	Invitrogen	Germany
Stepone plus Real Time PCR System	Applied Biosystems	Germany
Thermo mixer comfort	Eppendorf	Germany
Vortex-Genie® 2 vortex mixer	Scientific Industries	Germany
Water Bath Model	GFL	Germany

4.2. Experimental methods

4.2.1. Cell Culture and Treatment

MCF-7 cells were cultured in Quantum 263 complete medium (PAA) supplemented with 100U/ml penicillin and 100mg/ml streptomycin antibiotics (Invitrogen) at 37°C with 5% CO_2. To deprive hormones, cells were maintained in RPMI 1640 medium (w/o phenol red, Invitrogen) supplemented with 1% charcoal-dextran stripped FBS (CS-FBS) (Invitrogen) for three days before treatment. With confluency at ~80%, MCF-7 cells were treated with 100nM 17β-estrodial (E2) or vehicle for indicated time for both ChIP-Seq and bulk mRNA-Seq experiments. In newly synthesized mRNA-Seq experiments, hormone deprived cells were stimulated with 100nM E2 or vehicle along with 4-thiouridine (4sU) at three time intervals (45min, 2.5hr or 5hr).

4.2.2. Chromatin Immunoprecipitation (ChIP)

Cross-link with 1% formaldehyde

isolate nuclei

lyse nuclei and sonicate chromatin

Immunoprecipitation with antibody

purify ChIP DNA

High-throughput sequencing with
Illumina Genome Analyzer

align reads to genome
and peak calling
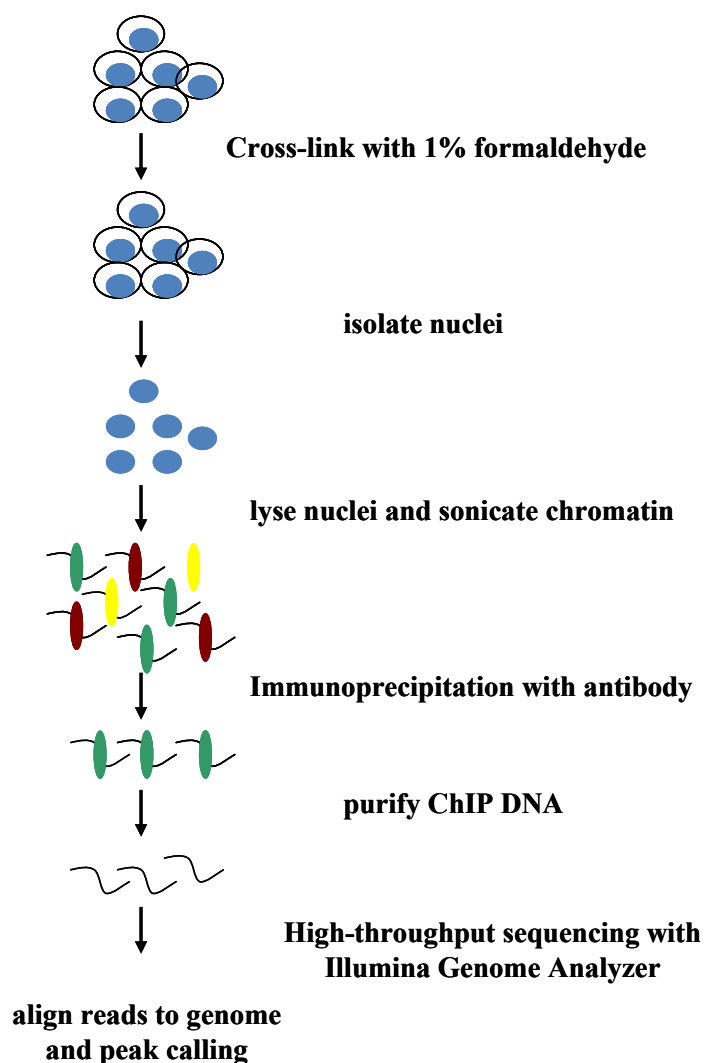
Figure 5. Schematic representation of ChIP-Seq workflow.

4.2.2.1. Crosslinking and Cell Lysis

To deprive hormones, MCF-7 human breast cancer cells were maintained in RPMI1640 medium w/o phenol red (Invitrogen) supplemented with 1% CS-FBS for 3 days before E2 treatment. After E2 or vehicle treatment at indicated time, 1% formaldehyde (Sigma Aldrich) was added into cell plates to cross-link proteins to DNA in cells and then quenched by glycine with 125mM of final concentration.

Next, cells were washed twice with ice-cold D-PBS (Gibco) freshly supplemented with 1X Protease Inhibitors cocktail (P.I., Roche), then detached from cell culture dishes by using cell scrapers and collected by centrifugation at 1,500 x *g* for 5min at 4°C. The cell pellets were

resuspended in Cell Membrane Lysis Buffer 1 (0.05M Hepes-KOH, pH7.5, 0.14M NaCl, 1mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100, freshly add 1X P.I.) and incubated at 4°C for 15min with gentle agitation. With centrifugation at 1,500 x g for 5min at 4°C, nuclei pellets were collected and then redissolved in Lysis Buffer 2 (0.2M NaCl, 1mM EDTA, 0.5mM EGTA, 10mM Tris pH8, 1X P.I. added before use) and incubated at 4°C for 15min with gentle agitation. The nuclei were spun down by centrifugation at 1,500 x g for 5min at 4°C.

4.2.2.2. Chromatin Shearing

The nuclei pellets were resuspended evenly in sonication buffer D (Diagenode) by carefully pipetting up and down for several times and sonicated for 15 cycles with 30s-on and 30s-off by using Biorupter (UCD-300, Diagenode). The whole cell lysates were centrifuged at 18,000 x g for 5min at 4°C to get rid of debris and detergents. The supernatants (30µl of chromatin solution corresponds to ~1million cells) were aspirated and ten-time diluted in ChIP dilution buffer (Millipore) freshly supplemented with 1X P.I.

4.2.2.3. Preclear and Immunoprecipitation

To preclear the chromatin with the aim to decreasing unspecific binding of chromatin to magnetic beads, 25µl of protein G magnetic beads (Invitrogen) were added to each IP reaction and incubated with agitation at 4°C for 1h. The supernatants were separated from magnetic beads by using magnetic rack (Invitrogen) and ready for immunoprecipitation. After adding the following antibodies, 5µg of anti-ERα polyclonal antibody (sc-543x, Santa Cruz), 2.5µg of anti-H3K4me3 polyclonal antibody (pAb-003-050, Diagenode), 2.5µg of anti-H3K27me3 polyclonal antibody (07-449, Millipore) and 2.5µg or 5µg of rabbit IgG polyclonal antibody (ab37415, Abcam), the immunoprecipitation reactions were performed at 4°C overnight with agitation. The next morning, after adding 20µl of protein G magnetic beads, the reactions were incubated at 4°C for 1hr.

4.2.2.4. Beads Washing

The beads were sequentially washed with the following washing buffers, once with Low Salt Immune Complex Wash Buffer (Millipore), once with High Salt Immune Complex Wash Buffer (Millipore), once with LiCl Immune Complex Wash Buffer (Millipore) and twice with TE Buffer (Millipore) at 4°C and collected by brief centrifugation. The magnetic beads comprising protein-DNA complex were pelleted by using magnetic rack.

4.2.2.5. Elution of Protein-DNA Complex

100μl of Elution Buffer (Millipore) was added to each IP reaction as well as input control with incubation at RT for 15min, and the mixture was collected by brief centrifugation. We then used magnetic rack (Invitrogen) to collect the supernatant that comprises protein-DNA complexes. We further performed a second round of elution with 100μl of Elution Buffer. The supernatant was collected as described above. In the end, the two elutes were combined into one centrifuge tube.

4.2.2.6. Reverse Cross-linking and DNA Purification

8μl of 5M NaCl was added into each IP reaction as well as input control with incubation at 65°C overnight for reverse cross-linking. The next morning, to digest RNA, 1μl of RNase A/T1 (Fermentas) per reaction was added with incubation at 37°C for 30min. Then, to digest protein, 4μl 0.5M EDTA, 8μl 1M Tris-HCl and 1μl Proteinase K (Invitrogen) were added into each reaction with incubation at 55°C for 1h. The mixture was cooled down to RT, transferred into 2ml Phase-Lock-Gel tubes (5 Primer) and supplemented with 400μl of Phenol/ Chloroform/Isopropanol (v/v/v=25:24:1, Invitrogen) to each tube, mixed well and centrifuged at 18,000 x g for 5min at RT. Then, the upper aqueous phase was transferred to a fresh, 2ml Phase-lock-Gel tube and supplemented with 400μl of Chloroform/Isoproponal (v/v=24:1, Invitrogen) per tube, centrifuged at 18,000 x g for 5min at RT. The upper aqueous phase was transferred into 1.5ml centrifuge tubes and added with 1ml of ice-cold 100% ethanol, 40μl of Sodium Acetate (pH5.0, Ambion) and 1μl of glycogen (1mg/ml, Ambion), incubated at -80°C for 30min. The DNA was collected by centrifugation at 18,000 x g for 30min at 4°C. After separating from the supernatant, the DNA pellets were washed with 500μl of ice-cold 70% ethanol, collected at 18,000 x g for 5min at 4°C. The supernatant was carefully discarded and the DNA pellet was air-dried for ~30min, resuspended in Nuclease-free water (Qiagen). Ultimately, the ChIP DNA was quantified by using Qubit™ dsDNA HS Assay Kits (Invitrogen).

4.2.3. ChIP qPCR and Data Analysis

ChIP qPCR primers designed to amplify 50 to 150bp DNA fragments from selected genomic regions were first assessed by conventional PCR using input DNA as template and then agarose gel electrophoresis. Only primers targeting specific genomic regions with a unique band in gel were retained. Primers were further evaluated by amplification efficiency (AE)

evaluation with ten-fold dilution of input DNA as qPCR template. Each PCR reaction should generate only the expected specific amplicons, as shown by the melting-temperature profiles of final products (dissociation curve). Only primers having an AE value between 0.95-1.10 were used for following qPCR experiments. Input DNA, DNA from IgG control and sample DNA were dissolved into the same volume of Nuclease-free water. The same volumes of DNA from above were used as qPCR templates, mixed with 2X SYBR$^®$ Green PCR Master Mix (Applied Biosystems) and 5µM of each primer. For each reaction, triplicates were performed. The program for qPCR in StepOnePlus real-time PCR machine (Applied Biosystems) was set as following: initial denature at 98°C for 10min followed by 40 cycles of denature at 98°C for 10s, annealing at 60°C for 10s. By visualization of amplification curves in StepOnePlus software (Applied Biosystems), baseline adjustment was made if the reaction emerged before the default baseline (cycles 3 to 15). The ΔC_t value (normalized to the input samples) was calculated for each sample: ΔC_t [C_t (sample) - C_t (input)]. Next, $\Delta\Delta C_t$ (ΔC_t (experimental sample) - ΔC_t (Myoglobin exon 2)) was calculated. Then, the fold enrichment between experimental samples and myoglobin exon 2 (negative control) was computed by using formula $2^{(-\Delta\Delta Ct)}$.

qPCR primer sequences:

GREB1-EN: forward primer, 5'-GTGTGGCAACTGGGTCATTC-3'; reverse primer, 5'-AAAGGCAGCAAACTTGTTTAGG-3', this primer set is expected to amplify a 99bp DNA fragment from ChIP DNA.

Myoglobin exon 2: forward primer, 5'-TCCTTAAGAAGGGG-CATC-3'; reverse primer, 5'-GGTACTTCACGGGGATCTTG-3', this primer set is expected to amplify an 87bp DNA fragment from ChIP DNA.

ACE-EN1: forward primer, 5'-TGTCCAC-CGTCACTCTCAC-3'; reverse primer, 5'-ACGGGCAGATCCTTGTACTG-3', this primer set is expected to amplify a 141bp DNA fragment from ChIP DNA.

ACE-EN2: forward primer, 5'-ATTTCACCTGACCCTGCTCTGC-3'; reverse primer, 5'-GTTACAGTCACCG-TGCCCTTTG-3', this primer set is expected to amplify a 148bp DNA fragment from ChIP DNA.

CYB561-EN: forward primer, 5'-CCCGCACCAGCTGTGTAGGC-3'; reverse primer, 5'-GGCTTGGTGGCGGTGTTCGA-3', this primer set is expected to amplify a 75bp DNA fragment from ChIP DNA.

4.2.4. ChIP Sequencing (ChIP-Seq)

~10ng of ChIP DNA was used for ChIP-Seq library preparation according to manufactures' instructions except for using SPRI purification kit (Beckman Coulter) instead of column purification kit (Qiagen) after each reaction. Briefly, both ends of the IP'd DNA were repaired by T4 DNA polymerase, Klenow DNA polymerase, and T4 polynucleotide kinase, An "A" base tail was added to the 3' end of the DNA fragments by using Klenow exo⁻ polymerase (3' to 5' exonuclease minus) for ligation with Illumina genomic adaptors (with a "T" overhang). DNA ligated with Illumina genomic adapters was PCR amplified for 18 cycles. The program for PCR cycles is as following:

98°C, 30sec; 98°C, 10sec; 65°C, 30sec; 72°C, 30sec; 72°C, 5min; 4°C, ∞.

18 cycles

The quantity and quality of DNA libraries were determined by Nanodrop-1000 spectrometer (Thermo Fisher) and Bioanalyzer 2100 (Agilent) respectively. In the end, the genomic adapter-ligated DNA was sequenced for 36 cycles on Illumina GAIIx following manufacture' instructions.

4.2.5. Total RNA Extraction

Total RNA was isolated from MCF-7 cells with treatment for indicated time by using TRIzol reagent (Invitrogen) according to the manufacturer' instructions. Briefly, the cells were directly lysed by adding 5ml of TRIzol reagent per 15cm cell culture dish (Greiner Bio-One). To ensure homogenization in cell lysis by TRIzol reagent, the cell lysates were passed 8-10 times through a 20G needle. The cell mixture was then split into 5 fresh, nuclease-free tubes (Ambion). After mixing vigorously with 200μl of chloroform (Invitrogen), the mixture was incubated at RT for 2min, transferred to a 2ml Phase-Lock-Gel tube (5 Prime) and centrifuged at 18,000 x g for 5min at 4°C. The upper aqueous phase containing RNA was aspirated and transferred to a fresh, nuclease-free tube. To precipitate RNA, 50μl of 5M NaCl and 500μl of isopropanol were added into each tube and incubated at RT for 10min. The precipitated RNA was pelleted by centrifugation at 18,000 x g for 30min at 4°C. The supernatant was carefully discarded and the RNA pellet was washed with 1ml of ice-cold 70% ethanol and recollected at 18,000 x g for 5min at 4°C. RNA pellets were resuspended in appropriate volume of nuclease-free water (Qiagen) and then quantified by using the NanoDrop ND-1000

Spectrophotometer (Thermo Fisher). The integrity of total RNA was assessed by Bioanalyzer 2100 (Agilent).

4.2.6. Purification of Newly Synthesized RNA with 4sU Labeling from Bulk RNA

Total RNA comprising newly synthesized RNA labeled with 4sU as well as pre-existing RNA was prepared for biotinylation. To biotinylate the 4sU-labeled RNA, 1mg/ml of EZ-Link Biotin-HPDP (Pierce) dissolved in dimethylformamide (DMF) was added to total RNA solution at a final concentration of 1μg/μl RNA. Biotinylation was carried out with incubation of the above solution for 3hr at RT. To purify biotinylated RNA from excess of Biotin-HPDP, a Phenol/Chloroform/Isopropanol (v/v/v=25:24:1) extraction was performed. After adding equal volume of Phenol/Chloroform/Isopropanol (v/v/v=25:24:1), the total RNA reaction mixture was transferred to a 2ml Phase-Lock-Gel tube, and followed were RNA precipitation and RNA pellet washing as described above. To fractionate biotinylated 4sU-labeled RNA with unlabelled RNA, streptoavidin coated magnetic beads (μMACS, Miltenyi) were used. 4sU-biotinylated RNA was first incubated with μMACS streptavidin beads for 15min at RT with rotation. Next, the separation of 4sU labeled RNA from unlabeled RNA was performed by applying the mixture onto one μColumn (Miltenyi) in the magnetic field of the μMACS separator (Miltenyi). Lastly, 4sU labeled RNA was recovered using the RNeasy MinElute kit (Qiagen) and reconstituted in appropriate volume of nuclease free water.

4.2.7. mRNA Sample Preparation and Sequencing

Total RNA from both bulk RNA and newly synthesized RNA were subjected to Illumina deep-sequencing. First of all, RNA-Seq libraries were prepared following the mRNA sequencing protocol provided by Illumina. In brief, the poly (A) RNA (poly-A+ RNA) was isolated from 1~2μg of total RNA by two rounds of hybridization to Dynal Oligo (dT) beads (Invitrogen). Then, the purified poly-A+ RNA was fragmented by using 5 X fragmentation buffer (200mM Tris acetate, pH8.2, 500mM potassium acetate and 150mM magnesium acetate) with heating at 94°C for 3.5min in a thermo cycler (BioRad). The fragmented poly-A+ RNA was precipitated by adding two volume of 100% ethanol, and subsequently purified using equal volume of SPRI beads following SPRI purification protocol. The fragmented poly-A+ RNA was reconstituted in 12μl of EB and used for first- and second-strand cDNA synthesis with random primers by using SuperscriptII reverse transcriptase and DNA polymerase I, respectively. The end-repair and A tailing steps for poly-A+ RNA converted ds-

cDNA are same as ChIP DNA library preparation as described above. Then the A tailing ds-cDNA was ligated with adaptors. After ligation, the purified cDNA using SPRI beads was amplified by 15 cycles of PCR reaction. The program for PCR cycles is as following:

98°C, 30sec; 98°C, 10sec; 65°C, 30sec; 72°C, 30sec; 72°C, 5min; 4°C, ∞.

<div align="center">15 cycles</div>

The DNA purified from PCR reaction was quantified and assessed by using Agilent Bioanalyzer 2100. Finally, the adapter ligated, PCR amplified DNA was sequenced for 36 cycles on Illumina GAIIx.

4.2.8. Paired-End RNA Sequencing Library Preparation and Sequencing

Each of the prepared RNA-Seq libraries for both bulk RNAs and newly synthesized RNAs at three time points (45min, 2.5hr and 5hr) with E2 or vehicle treatment were adjusted to the concentration of 10nM. The final library was sequenced using an Illumina HiSeq 2000 with a paired-end module with run length 2x100bp.

4.3. Computational Methods

4.3.1. Identification of ERα Binding Sites

The 36bp short reads generated by massively parallel sequencing were aligned onto the human genome assembly (hg19) with Bowtie v.0.12.0 (Langmead et al., 2009), allowing up to two mismatches. In order to avoid potential PCR amplification bias, short reads that shared the same mapping location on the same strand were discarded. The uniquely-mapped non-redundant short reads with at most two mismatches were retained for further processing. To slide Watson and Crick short reads for each protein-DNA interaction site, we used qips software (Gogol-Döring and Chen, 2010) to estimate the ChIP DNA fragment length. Next, regions enriched with short reads from ERα ChIP-Seq relative to the same regions with short reads from IgG ChIP-Seq were called as peaks by MACS software (Zhang et al., 2008). The peak calling analysis was performed independently on the replicates. To ensure the high confidence of ERα ChIP-Seq peaks in this study, peaks were further processed as following. First, the comparative analysis showed that the vast majority of ERα ChIP-Seq peaks at 2.5hr of E2 treatment also present at 45min of E2 treatment. Therefore, ERα ChIP-Seq peaks at 2.5hr of E2 treatment can be used to ensure the validity of ERα ChIP-Seq peaks identified after 45min of E2 treatment. Further, three raw datasets of ERα ChIP-Seq experiments (Cicatiello et al., 2010; Hu et al., 2009; Welboren et al., 2009) were retrieved from GEO database and peaks calling analysis for them were performed using the procedures described above. Then, ERα ChIP-Seq peaks at 45min were compared with our ERα ChIP-Seq peaks at 2.5hr as well as the three reanalyzed ERα ChIP-Seq datasets from publications. Only ERα peaks (45min) replicated in a second independent data set (present in two of our 45min data set or our ERα peak dataset at 2.5hr, or any of the three reanalyzed datasets) were retained. Further, to ensure low false negative rate in our peak identification, peaks of biological replicates kept from previous step were combined and the integrated peak datasets were used for further analysis. Finally, we identified 4543 ERα ChIP-Seq peaks at 45min of E2 treatment, and 992 ERα ChIP-Seq peaks at 45min after vehicle treatment.

4.3.2. Motif Analysis at ERα Binding Sites

The ERα ChIP regions were uniformly resized by 100bp in both directions from their peak summits. Position weight matrices (PWM) for ERα and ERβ as well as the other 130

vertebrate transcription factors were retrieved from JASPAR database (Bryne et al., 2008). Each of 200bp ERα ChIP regions was scanned with one PWM by sliding it across the whole region, and the best PWM score calculated was recorded. To estimate the background PWM distribution, the DNA sequence of each ERα ChIP region was randomly permutated by using a python script, and the scores of these permutated sequences were obtained in the way described above. To calculate the sensitivity and specificity for the ROC curve, the simulated true positive rate (sensitivity) and false positive rate (1-specificity) was computed by using PWM scores of all the ERα ChIP regions and their permutated sequences under a spectrum of PWM score thresholds. Based on this, the ROC curve was drawn for each PWM.

4.3.3. H3K4me3 ChIP-Seq Peak Detection

Short read mapping, filtering and enriched peak calling for H3K4me3 ChIP-Seq were processed by using the same bioinformatics pipeline as described in ERα ChIP-Seq peak detection. Comparative analysis of ChIP-Seq peaks between treatment conditions as well as treatment times were performed on H3K4me3 ChIP-Seq. Results (**Figure 9C**)suggested that H3K4me3 ChIP-Seq peaks showed only a very subtle change in response to E2; however, no changes were found at different treatment time points (45min and 2.5hr). To gain a comprehensive dataset on H3K4me3 occupancy in breast cancer cell line MCF-7, we then combined all ChIP-Seq peaks for H3K4me3 whatever treatment conditions (E2 or vehicle treatment) or treatment time points (45min or 2.5hr) ChIP-Seq was performed. In the end, we identified 20513 H3K4me3 ChIP-Seq peaks.

4.3.4. Determine E2-regulated Genes

The 36bp short RNA-Seq reads were mapped onto human genome assembly (hg19) by using Bowtie with at most 2 mismatches allowed. We then determined E2-regulated genes as follows.

Set up Gene List

We selected RefSeq genes residing within chromosome 1 to 22 and chromosome X and discarded all non-chromosomal contigs and the whole chromosome Y (MCF-7 is derived from female). This way we got a list of 36421 genes. Next, we merged the overlapping RefSeq genes/isoforms to a single gene cluster (hereafter referred as "genes"). This step results in 19491 genes with larger exons by merging the overlapping exons. Further, for each gene, we determined the total number of positions on both strands where reads can be

uniquely mapped to and completely fall inside of an exon and cumulated all the positions to acquire the "gene exon length". By discarding all genes with gene exon length below 50, we generated a final list of 18842 genes.

Compute Background Read Density

Firstly, we pooled all RNA-Seq files of sequencing reads into one file and then applied the qips software to find genomic regions with enriched number of sequencing reads. Based on the above analysis, we assumed regions without mapped reads to be "non-transcribed". For each data set, we determined the "background read density" as the number of reads outside of qips regions divided by total length of all genomic regions outside of qips regions.

Compute Gene Expression Levels

For each data set, by mapping reads to genome, we kept all uniquely mapped reads, i.e. also redundant reads which mapped to the same positions uniquely. For each gene, to avoid artifacts, we checked the reads mapped into the exons of the gene and counted the number of positions on both strands where reads actually map to define "bins". Then, for each bin, we determined the stack height as the number of uniquely mapped reads mapped to it. If the highest stack (bin with maximum stack height) is not higher than 2, we stopped further analysis. Otherwise, we computed the probability (p-Value) of getting the highest stack by chance when distributing all reads mapped to the gene independently to the bins (Poisson model). If the p-Value was below a threshold of 1×10^{-5} then we marked the bin as "artifact". In this case, we removed all reads in this bin and repeated the artifact identification as above. Last, we pooled all genomic positions marked as "artifacts" in all data sets to a single set of artifacts. For each data set and each gene, we determined expression level as follows. To get the read counts in the genes, we first determined the average stack height of all non-artifact bins and counted reads of the average stack height for non-artifact bins as well as reads of the average stack height for each artifact bin. We discarded all genes with read count below 10 (expression level is regarded as 0). We defined the expression level as read count divided by the gene exon length minus background read density, or 0 if this is negative.

Determine E2-Regulated Genes

We compiled the treated data set (E2+) and the corresponding untreated data set (E2-) as one pair at each time point. We discarded all genes with expression level of 0 in one or in both data sets. For each gene, we first compute the "intensity" by summing up the log2 expression levels and the "change" by calculating the difference between the log2 expression levels. We then normalized the changes of all genes relative to the intensities by applying a LOESS

(locally weighted scatterplot smoothing) regression. Next, we computed the standard deviation for each gene first sorting the genes according to their intensities and then centering a sliding window on each gene containing 1% of the genes having similar intensities. We computed the standard deviation of the changes of all genes within this window. For each gene, we defined the "Z-value" as the change of that gene divided by the standard deviation. By assigning a sign to the Z-value such that it gets positive (negative) if the gene expression level gets up (down) in the treated data set compared to the untreated data set, we acquired a "signed Z-value". Assuming a normal distribution, we computed for each gene a "p-Value" corresponding to its Z-value. For each treatment and time point, we checked the three replicates. If a gene is changed in all replicates in the same direction with a p-Value of at most 0.1, then it is called "regulated". For each gene, we computed the joined signed Z-values as the average of the signed Z-values of the three replicates. We defined a gene as "regulated", if it is regulated for at least one treatment/time point.

4.3.5. Novel E2 Regulated Intergenic Transcript Identification

We searched the combined RNA-Seq data for genomic regions enriched with small reads using qips (Gogol-Döring and Chen, 2010). We then excluded qips regions overlapped with RefSeq genes (at least one nucleotide overlapping). Then, the expression levels of qips regions at each time point with different treatment (E2 or vehicle) were calculated as the method of E2 regulated gene determination described above. There are 77 E2-regulated qips regions remaining. We then used paired-end RNA-Seq data to merge qips regions originating from the same transcripts. After this, 68 E2 regulated regions were retained. We then overlapped these E2 regulated regions with UCSC genes, Genbank EST database, Genbank mRNA database and lincRNA database from Khalil et al., 2009 and finally identified 29 novel E2-regulated previously uncharacterized regions.

5. Results

5.1. Genome-Wide Mapping of ERα Binding Sites by ChIP-Seq

ERα can exert widespread transcriptional controls on luminal breast cancer cells through binding to *cis*-regulatory DNA elements that are either proximal or distal to target genes. Hence, the genome-wide analysis of *cis*-regulatory DNA elements bound with ERα (ERα cistrome) would be a crucial step in deciphering the ERα mediated transcriptional regulatory network in ERα positive breast cancer cells. To globally profile and characterize the *cis*-regulatory DNA elements interacting with ERα, researchers have first applied the chromatin immunoprecipitation combined with microarray (ChIP-on-chip) in breast cancer cells (Carroll et al., 2005 and 2006). More recently, with the advent of massively parallel sequencing-based approach, namely ChIP-Seq, ERα cistrome has been interrogated with higher sensitivity, better resolution, and more accuracy (Welboren et al., 2009). Here, in this study, we performed a genome-wide analysis of ERα binding sites by ChIP-Seq in ERα positive MCF-7 human breast cancer cell line. We chose MCF-7, because it is the most widely used cell model for hormone-responsive luminal breast cancer. Moreover, a few previous studies on globally profiling ERα binding sites (Carroll et al., 2006; Cicatiello et al., 2010; Welboren et al., 2009) have also focused on this cell line, allowing us to compare our results with previous studies.

Briefly, in our study, MCF-7 cells were deprived of hormones for three days and then treated with 100nM E2 or vehicle. Later on, the cells were harvested at indicated time points (45min or 2.5hr). Of these two time points, 45min is the time point that has been extensively used in the previous studies of ERα binding site, because it has been shown that maximal ERα-chromatin interactions occur at this time period (Cheng et al., 2006; Shang et al., 2000). To investigate the global ERα binding dynamics, we have also performed the ERα ChIP-Seq at a later time point (2.5hr).

After crosslinking proteins with DNA by 1% formaldehyde (w/v), we used the specific antibody raised against ERα to immunoprecipitate *cis*-regulatory DNA elements interacting with the E2-ERα complex, and in parallel we utilized a non-specific antibody (rabbit IgG) in a

separate immunoprecipitation to control the experimental background. In the end, the ERα ChIP DNA as well as IgG ChIP DNA was subjected to deep sequencing.

Table 1. Summary of ERα ChIP-Seq results.

Dataset	Raw reads	Filtered reads	Mapped reads	Uniquely mapped reads	Non-redundant uniquely mapped reads
ChIP_ERα+.45m.Replicate1	33,461,154	27,562,082	26,582,439	21,900,951	21,257,225
ChIP_ERα-.45m.Replicate1	30,364,930	25,153,909	24,492,057	20,120,354	19,524,331
ChIP_ERα+.45m.Replicate2	34,570,217	28,428,276	27,493,328	22,928,683	21,360,082
ChIP_ERα-.45m.Replicate2	33,251,825	26,807,611	25,617,862	21,354,075	19,703,929
ChIP_ERα+.45m.Replicate3	19,291,696	17,677,691	17,184,281	12,890,291	11,268,384
ChIP_ERα-.45m.Replicate3	27,279,659	24,826,893	24,172,785	18,189,835	16,261,443
ChIP_IgG.45m.Replicate1	44,530,369	33,504,098	32,907,046	25,966,327	24,757,661
ChIP_IgG.45m.Replicate2	40,978,528	34,072,369	32,592,553	27,267,004	24,165,982
ChIP_ERα+.150m	15,660,237	14,337,250	13,052,334	10,678,208	10,232,175
ChIP_ERα-.150m	14,134,122	12,133,113	11,224,324	9,141,807	8,881,298
ChIP_IgG.150m	43,391,421	32,877,400	32,202,694	25,301,718	23,897,361

We have performed triplicate ChIP-Seq experiments at 45min (with E2 or vehicle treatment), and only one experiment at 2.5hr (**Table 1**). After sequencing, in total, around 90.9 and 87.3 million raw reads were generated for ERα ChIP-Seq at 45min after treating with E2 or vehicle respectively. Of these, 76.8 and 73.7 million reads with good sequencing quality (filtered reads) were retained, from which 74.3 and 71.3 million reads (mapped reads) can be mapped to genome. Among these mappable reads, 59.7 and 57.7 million reads were mapped at unique genomic positions (uniquely mapped reads). To avoid the potential PCR artifacts, only one of the reads mapped to the same position and the same strand was retained. As a result, 55.5 and 53.9 million reads (non-redundant and uniquely mapped reads) were used as the input datasets, together with background data (see below), to call ERα peaks by a tool called "Model-based Analysis of ChIP-Seq" (MACS) (Zhang et al., 2008).

At 2.5hr, in total, around 14.1 and 15.5 million raw reads were generated for ERα ChIP-Seq after E2 or vehicle treatment respectively. Of these, 12.1 and 14.3 million reads (filtered reads) were retained after quality filtering, 11.2 and 13 million reads can be mapped to genome. Among them, 9.1 and 10.7 million reads (uniquely mapped reads) can be uniquely mapped.

Finally, 8.8 and 10.2 million reads (non-redundant and uniquely mapped reads) were used, together with background data (see below), to call peaks by MACS.

We have performed duplicate IgG ChIP-Seq experiments at 45min and one IgG ChIP-Seq experiment at 2.5hr. After sequencing, 85.5 and 43.4 million raw reads were generated for IgG ChIP-Seq at 45min and 2.5hr respectively. Of these, 67.6 and 32.9 million reads (filtered reads) were kept after quality filtering, 65.5 and 32.2 million reads (mapped reads) can be aligned to genome. Among them, 53.2 and 25.3 million reads (uniquely mapped reads) can be uniquely mapped. Last, 48.9 and 23.9 million reads (non-redundant and uniquely mapped reads) were used as background data for MACS to call ERα binding sites (see above).
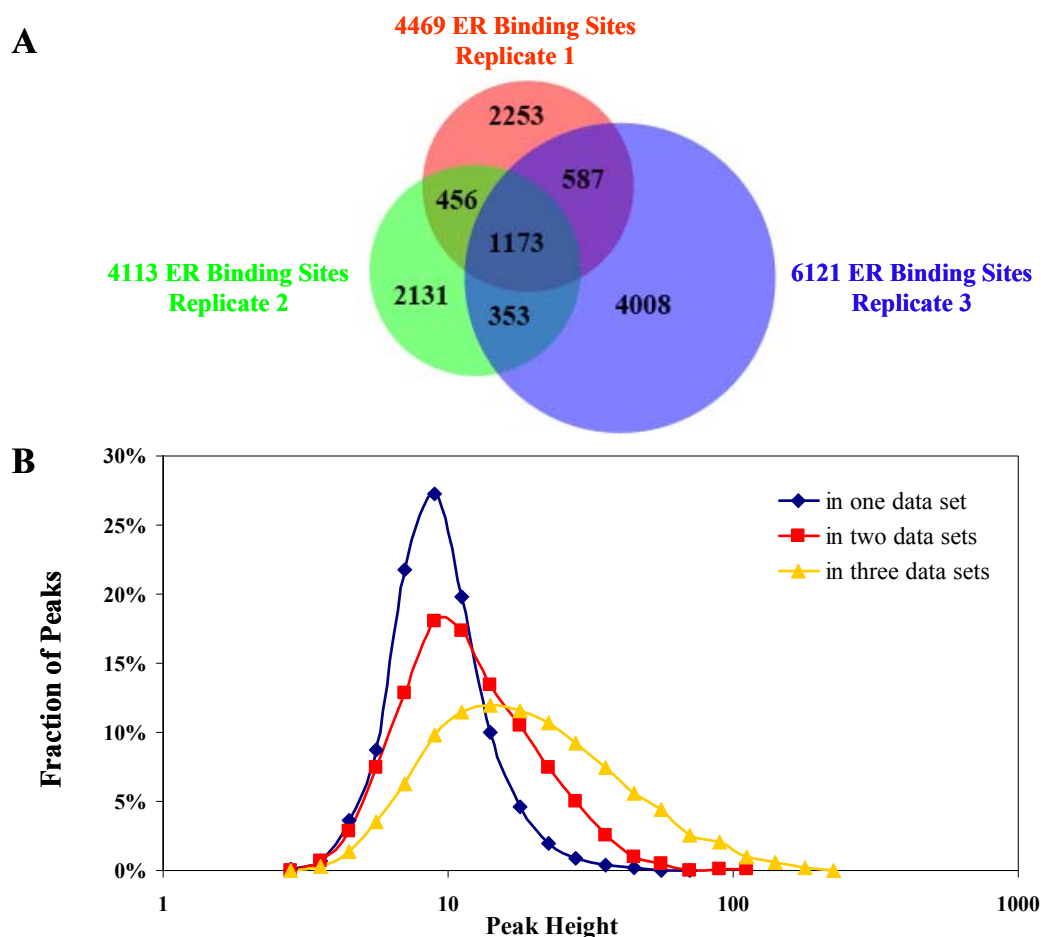
Figure 6. Overlap analysis of ERα binding sites (E2, 45min) in triplicates and peak height analysis of three distinct classes of ERα binding sites. (A) The Venn diagram shows a limited overlap of ERα binding sites among the three replicates. (B) Peak height was compared between the three classes of ERα binding sites. In general, the binding sites common to all three replicates (**yellow**) contained higher peaks than those identified in two replicates (**red**), which in turn have higher peaks than the binding sites exclusively found only in one experiment (**blue**).

At 45min after E2 treatment, we have identified 4469, 4113 and 6121 ERα binding sites for the three replicates respectively. As shown in **figure 6A**, only 1173 sites were shared among the three replicates, whereas 1396 sites were common to any two of the three replicates, 2253, 2131 and 4008 sites were only found in the 1st, 2nd and 3rd replicates respectively. The rather limited overlap between the datasets could in principle be due to a low quality of the results. However, we believe that the observed differences at least in part arise from the nature of ERα-DNA interaction, i.e. stronger or more stable binding is captured more often in reproducible manner than weaker or transient binding. To confirm this hypothesis, we plotted the distribution of peak heights for the binding sites identified only in one replicate, in two replicates or in all three replicates, separately. As shown in **figure 6B**, in general, the binding sites common to all three replicates contained higher peaks than those identified in two replicates, which in turn have higher peaks than the binding sites exclusively found only in one experiment. These results confirm that the overlapped findings often reflect strong or stable interaction whereas most of weak or transient binding might be detected in an irreproducible manner. In order to characterize as many as possible ERα binding sites and at the same time to avoid the artifacts, we collected and reanalyzed another three published ERα ChIP-Seq datasets (Cicatiello et al., 2010; Hu et al., 2010; Welboren et al., 2009). In addition to our binding sites identified in at least two replicates, we also retained our binding sites identified once, but present in at least one of the three published datasets. As a result, a final list of 4543 high-confident ERα binding sites (E2, 45min) as well as 992 ERα binding sites (vehicle, 45min) was identified and used for the following analysis.
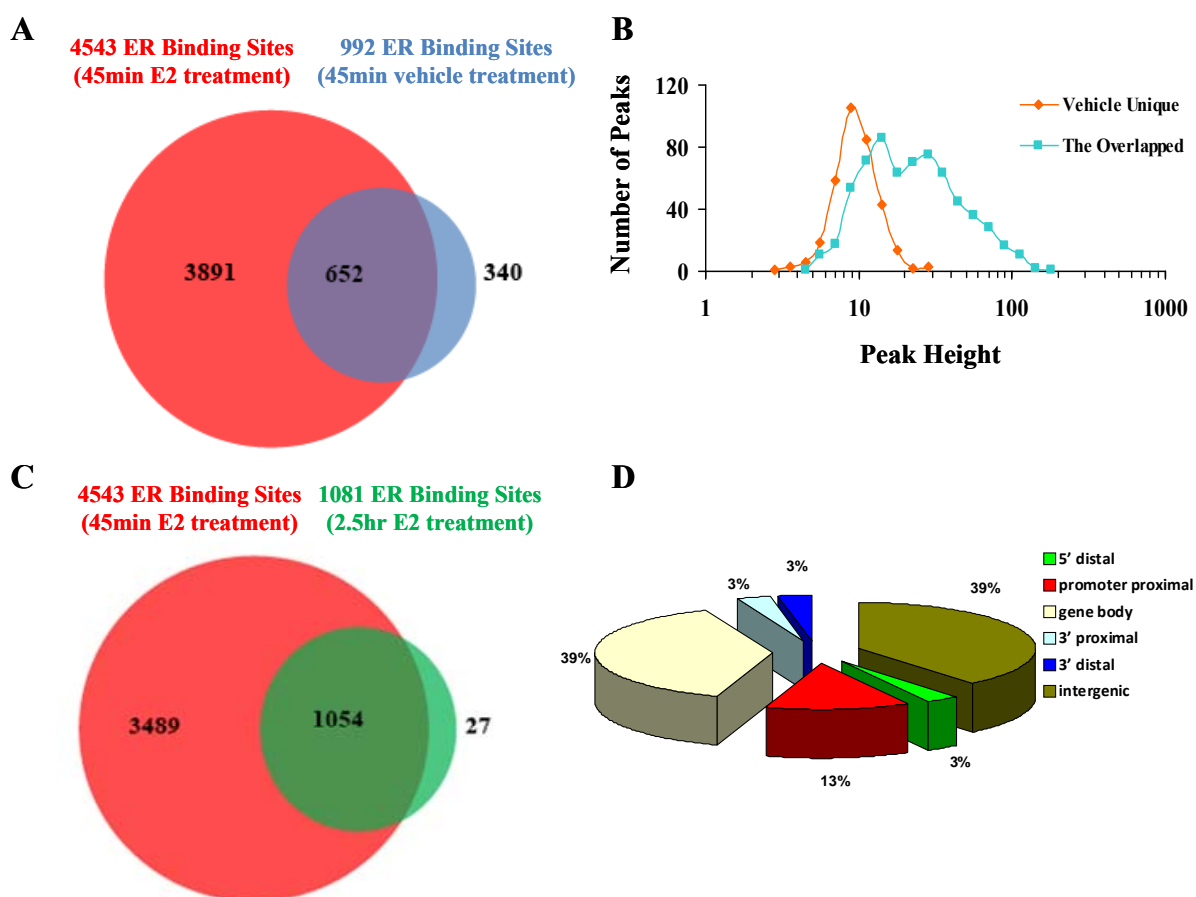
Figure 7. An overview of the characteristics of ERα binding sites. (A) Illustrated is the overlap of ERα binding sites at 45min between E2 and vehicle treatment. The majority of ERα binding sites with vehicle treatment are also present in E2 treatment dataset. (B) Peak height was compared between two classes of ERα binding sites identified in the absence of E2. In general, the binding sites (**orange**) remaining after E2 treatment contained significantly higher peaks than those sites (**blue**) disappearing after E2 treatment. (C) Overlap between ERα binding sites identified at different time points demonstrated that the ERα binding pattern is defined already in early period of E2 treatment. (D) The genomic distribution of all the ERα binding sites is represented by the 3D pie chart.

In contrast to previous findings which showed nearly no ERα binding events in absence of E2, we have identified a significant number of ERα binding sites without E2 treatment (vehicle, 45min), although the number is four times smaller than the number of ERα binding sites determined following E2 treatment (**Figure 7A**). It is known that in the absence of E2, ERα largely stays in the cell cytoplasm and only a relatively small proportion can enter nucleus via some alternative pathways mediated by MAPK-ERK and PI3K-AKT (Tomas et al., 2011). To address whether ERα binds at different sites under the two conditions, i.e. with and without E2 stimulation, we compare the two sets of ERα binding sites. It turned out that 66% of ERα

binding sites identified in the absence of E2 treatment are also present after E2 treatment (**Figure 7A**). Interestingly, in general, the 652 binding sites remaining after E2 treatment contained significantly higher peaks than those 340 sites disappearing after E2 treatment (**Figure 7B**). This suggests, on one hand, the binding sites identified only in the absence of E2 represent mostly weak or transient ERα-DNA interactions; on the other hand, a significant number of strong binding events are still present even after hormone deprivation for three days, which might result from residual ERα remaining in cell nuclei or entering nuclei via alternative pathways.

To gain the knowledge of ERα binding dynamics in the time course of E2 treatment, we also performed a genome-wide analysis of ERα binding sites at a later time point (2.5hr). After peak calling, we have identified a total of 1081 ERα binding sites after E2 treatment for 2.5hr. Strikingly, as shown in **figure 7C**, the vast majority (97.5%) of these ERα binding sites have also been found in the 45min ERα ChIP-Seq dataset. It validated the previous notion that maximum ERα-DNA interaction happens already at 45 minutes after E2 stimulation and indicated that very few if any new interactions arise at a later time point.

To better understand the functional implications of the ERα binding events, we annotated the binding sites according to their distance to the known RefSeq genes. In total, all the 4543 ERα binding could be categorized into six subgroups, i.e., promoter proximal (between -5Kb and +5Kb from the Transcription Start Site (TSS) of RefSeq genes), gene body (5Kb downstream of TSS till the 3'end), 3' proximal (5Kb downstream of 3' end), 5' distal (between -10Kb and -5Kb from the TSS), 3' distal (between 5Kb and 10Kb downstream of 3'end) and remaining intergenic regions. As shown in **figure 7D**, we found that the majority (76%) of ERα binding sites are located in gene body (37%) and intergenic regions (39%). Whereas 15% and 3% of ERα binding sites are present within promoter proximal regions and in the vicinity of 3' polyadenylation sites, respectively, and 6% are located within 5' or 3' distal regions. Our finding that only 15% of ERα binding sites are located within promoter proximal regions is in agreement with previous studies, in which Carroll et al. have mapped 4% of ERα binding sites to 1Kb promoter proximal regions and Welboren et al. have located 7% of ERα binding sites within 5Kb of promoter proximal regions (Carroll et al., 2006; Welboren et al., 2009).

It is well known that ERα binds DNA directly often at canonical Estrogen Response Elements (EREs). To search for DNA motifs enriched in ERα binding sites identified in our study, we employed position weight matrices (PWMs) for the DNA binding motifs of ERα as well as other 131 vertebrate transcription factors from JASPAR database (http://jaspar.cgb.ki.se/) to scan through our ERα binding sites identified after treating with E2 for 45 minutes or 2.5 hours. As shown in **figure 8B** and **8C**, the area under the ROC curve (AUC) for ERα DNA binding motif as well as of ERβ DNA binding motif (**Figure 8A**) was apparently much greater than AUC for any of the other matrices, both in 45 minutes and 2.5 hours datasets. This indicated that the binding events identified in this study are significantly enriched with direct binding of ERα at their canonical recognition sites.

A

B

C
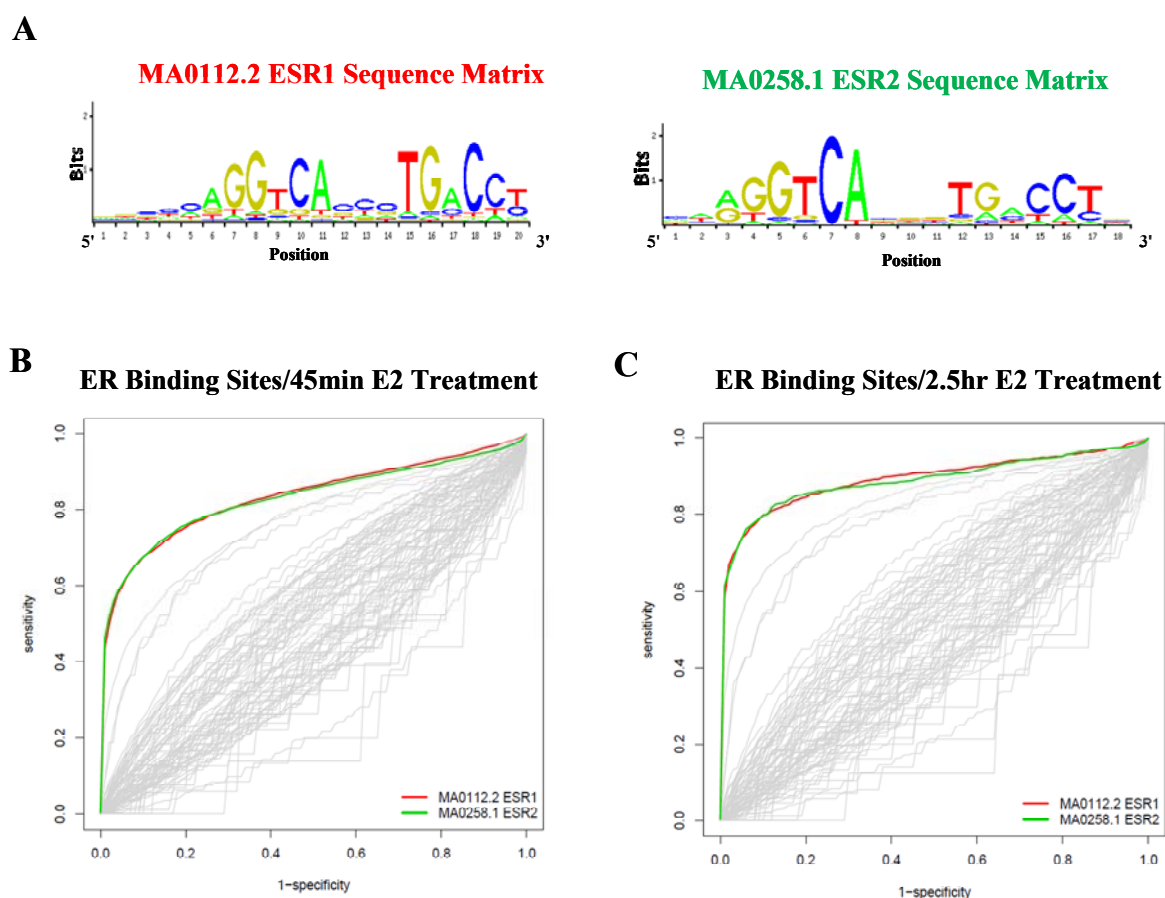
Figure 8. Motif analyses of ERα binding sites identified after 45min or 2.5hr of E2 treatment. (A) Sequence logos of ERα and ERβ binding motifs adapted from JASPAR database. Enrichment of transcription factor binding motifs at ERα binding sites identified after 45min (B) and 2.5hr (C) of E2 treatment is indicated by ROC curve. In both figure B and C, the ERE motif illustrated by ESR1 and ESR2 is more represented than other motifs.

5.2. Global Mapping of H3K4me3 Signature

Histone modifications regulate gene expression by maintaining chromatin environments and influencing DNA accessibility. Among various modifications, trimethylation at histone 3 lysine 4 (H3K4me3) is often found at promoters of actively transcribed genes (Berstein et al., 2005; Kim et al., 2005). Here we globally profiled the H3K4me3 with two motivations: 1) investigate the impact of ERα binding on the local chromatin configuration; 2) use H3K4me3 to identify novel promoters upon which ERα acts.

Table 2. Summary of H3K4me3 ChIP-Seq results.

Dataset	Raw reads	Filtered reads	Mapped reads	Uniquely mapped reads	Non-redundant uniquely mapped reads
ChIP_H3K4me3+_45m_Replicate1	17,763,974	15,773,201	15,269,874	12,203,995	11,281,743
ChIP_H3K4me3-_45m_Replicate1	18,910,900	16,566,973	16,039,521	12,878,604	12,088,907
ChIP_H3K4me3+_45m_Replicate2	17,549,663	15,739,075	15,264,108	12,200,413	11,266,127
ChIP_H3K4me3-_45m_Replicate2	19,232,127	17,087,245	16,570,277	13,299,048	12,470,894
ChIP_H3K4me3+_150m_Replicate1	27,940,930	24,573,021	23,984,981	19,7625,066	19,322,020
ChIP_H3K4me3-_150m_Replicate1	26,680,689	23,594,394	23,030,959	18,980,950	18,565,951
ChIP_H3K4me3+_150m_Replicate2	28,903,064	24,877,102	24,281,400	19,989,975	19,516,392
ChIP_H3K4me3-_150m_Replicate2	30,343,591	25,497,267	24,882,887	20,505,248	20,048,657

We have performed two biological replicates of ChIP-Seq at each time point of E2 treatment (45min and 2.5hr), and the same for vehicle treatments,(**Table 2**). After sequencing, around 35.3 and 38.1 million raw reads were generated for H3K4me3 ChIP-Seq at 45min after treating with E2 or vehicle respectively. Of these, 31.5 and 33.6 million reads with good sequencing quality (filtered reads) were retained, 30.5 and 32.6 million reads (mapped reads) can be mapped to genome. Among them, 24.4 and 26.2 million reads were mapped at unique genomic positions (uniquely mapped reads). Finally, 22.5 and 24.6 million non-redundant and uniquely mapped reads were used, together with background dataset (see **section 5.1**), as the input dataset to call H3K4me3 peaks by "Model-based Analysis of ChIP-Seq" (MACS) (Zhang et al., 2008).

At 2.5hr, in total, around 56.8 and 57 million raw reads were generated with E2 or vehicle treatment respectively. Of these, 49.4 and 49.1 million reads (filtered reads) were retained after quality filtering, 48.3 and 47.9 million reads can be mapped to genome. Among them, 39.7 and 39.5 million reads (uniquely mapped reads) can be uniquely mapped. Finally, 38.8 and 38.6 million reads (non-redundant and uniquely mapped reads) were used to call peaks. Between replicates after 45min of E2 treatment, the H3K4me3 intensities within 1kb of all TSS, including TSSs of the E2 regulated gene (see **section 5.3**) are highly reproducible (**Figure 9A**).

It is known that H3K4me3 occupancy at promoter is positively correlated with gene expression level; therefore, we compared the H3K4me3 peak intensity after 2.5hr of E2 treatment with gene expression level measured by bulk mRNA-Seq also after 2.5hr of E2 treatment. To do this, we categorized genes detected by bulk mRNA-Seq into 30 bins according to their gene expression levels. For each bin, the average gene expression level was calculated. We then averaged the H3K4me3 intensity, i.e. the number of ChIP sequencing reads, within 1Kb of TSS of genes in each bin. As expected, the H3K4me3 peak intensities are significantly positively correlated with gene expression levels with the Spearman coefficient at 0.99 (**Figure 9B**).

We sought to check the impact of E2 treatment on global profile of H3K4me3 at different treatment time points. For instance, we compared the occupancy profiling of H3K4me3 between E2 and vehicle treatment at 2.5hr. To this end, we compare the H3K4me3 intensity within 1Kb of RefSeq TSSs between E2 and vehicle treatment for 2.5hr. As shown in **Figure 9C**, no significant changes for H3K4me3 occupancy were observed both at the genomic level and for the E2 regulated genes (see **section 5.3**), compared to the difference between two replicates. Furthermore, we examined whether H3K4me3 occupancy profile changed on a time scale after E2 treatment. The comparison of H3K4me3 intensity within 1Kb of all TSSs between 45min and 2.5hr after E2 treatment also showed that very few changes were detected in the time frame examined (**Figure 9D**).
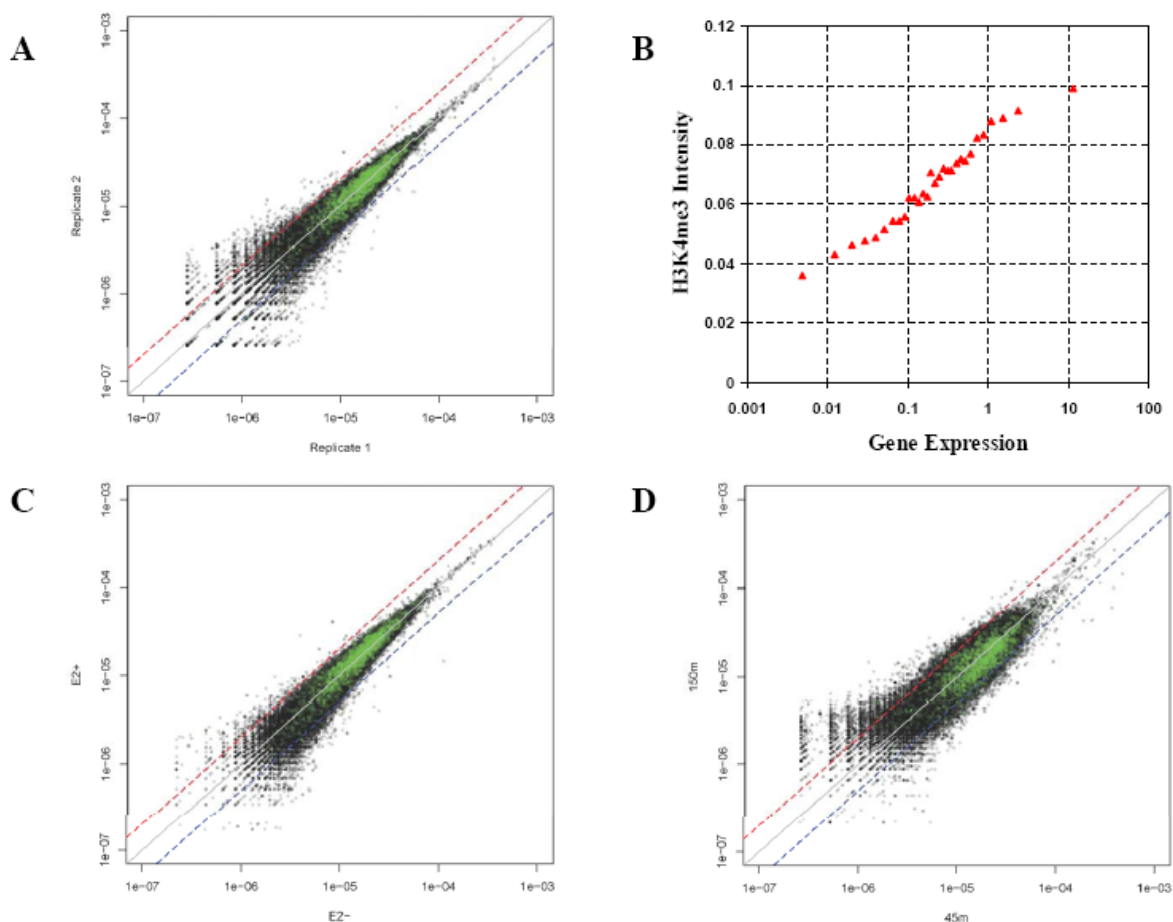
Figure 9. Characterization of H3K4me3 occupancy in MCF-7 cells. (A) The scatter plot shows the correlation of H3K4me3 intensity between two replicates after 45min of E2 treatment. TSSs of the regulated genes (see **section 5.3**) are in **green**, whereas the other TSSs are in **black**. Dots between the red and blue dashed lines are less than two-fold changes. (B) The positive correlation of H3K4me3 peak intensity, i.e. the number of ChIP sequencing reads within 1Kb of TSS, with gene expression level measured by mRNA-Seq is shown by scatter plot. (C) The scatter plot shows the high correlation of H3K4me3 peak intensity between E2 (Y axis) and vehicle treatment (X axis) at 2.5hr. (D) The scatter plot shows the high correlation of H3K4me3 peak intensity between 45min (X axis) and 2.5hr (Y axis) of E2 treatment.

Since the "active" H3K4me3 signature was commonly found at gene promoter proximal regions, we sought to check where H3K4me3 peaks identified in our study located relative to known genomic features. As shown above (**Figure 9C and 9D**), the vast majority of enriched H3K4me3 peaks we identified were present in all the datasets, regardless of E2 treatment and time points (45min or 2.5hr). Therefore, to avoid the potential false positive, in the following analysis, we kept only the peaks presented in at least two of the four datasets (45min+E2, 45min+vehicle, 2.5hr+E2, 2.5hr+vehivle) and as the result, a total of 20513 H3K4me3 peaks were retained. We then categorized these peaks according to the distance from the closest

RefSeq genes, i.e., 1) RefSeq TSS peaks (H3K4me3 peaks within 5Kb upstream or downstream of known RefSeq TSSs); 2) RefSeq gene body peaks (H3K4me3 peaks beyond 5Kb of RefSeq TSSs, but located within RefSeq genes); 3) Intergenic peaks (the remaining H3K4me3 peaks). As shown in **figure 10A**, as many as 16338 H3K4me3 peaks were around RefSeq gene TSSs. The 1488 H3K4me3 peaks located within genes probably represented the alternative promoter.

Finally, we sought to identify the promoter proximal ERα binding sites according to H3K4me3 profile. In total, 961 (~24%) ERα binding sites were found in the vicinity (5Kb upstream/downstream) of at least one H3K4me3 peaks. In comparison, 678 (~17%) ERα binding sites were found in the vicinity (5Kb upstream/downstream) of at least one known TSS of RefSeq genes. As shown in **figure 10B**, 472 ERα binding sites can be annotated as promoter proximal by both RefSeq TSSs and H3K4me3 peaks, whereas 489 and 206 ERα binding sites could be annotated as such exclusively according to H3K4me3 profile or known TSS, respectively.
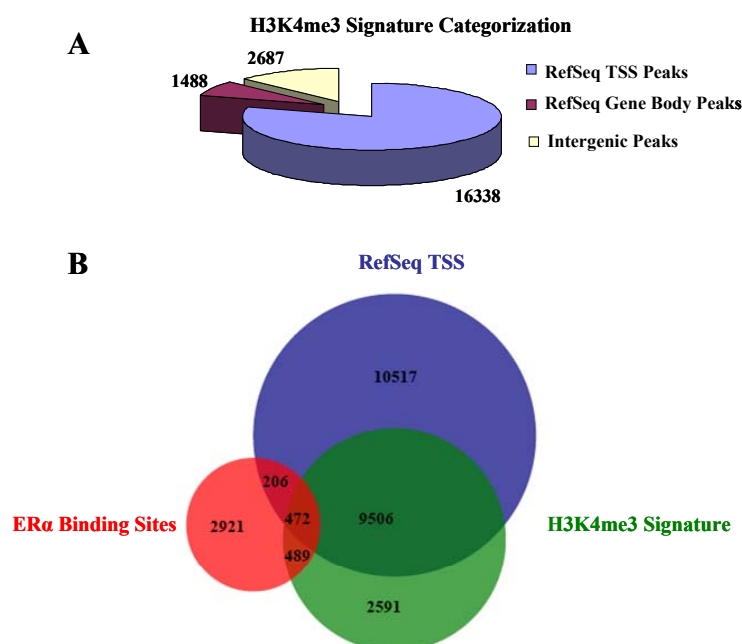
Figure 10. Characterization of H3K4me3 peaks can improve RefSeq TSS annotation. (A) Our H3K4me3 occupancy data can predict 1488 alternative promoters. (B) The Venn diagram shows that the prediction of promoter proximal ERα binding sites by RefSeq TSSs can be improved in combination with H3K4me3 occupancy data.

5.3. Global Gene Expression Profiling in MCF-7 Cells

To determine the transcriptional response induced by E2 stimulation, we performed gene expression profiling by the novel RNA-Seq method on three different time points following E2 treatment. Specifically, in order to identify the regulated genes with higher sensitivity and accuracy, we sequenced newly synthesized RNAs, in addition to the bulk RNAs. Whereas the bulk level is the combinatory effect of synthesis and turnover of both newly synthesized RNA after E2 treatment and pre-existing RNA before E2 treatment, we believe newly synthesized RNAs should provide a more direct readout for E2-driven transcriptional regulation.

Table 3. Summary of RNA-Seq results.

Dataset	Raw reads	Filtered reads	Mapped reads	Uniquely mapped reads	Non-redundant uniquely mapped reads
RNA-Seq_Bulk-.45m_Replicate1	15,772,078	13,837,253	12,508,736	9,926,870	5,602,331
RNA-Seq_Bulk+.45m_Replicate1	16,718,795	14,285,145	12,853,990	10,258,656	5,822,466
RNA-Seq_Bulk-.45m_Replicate2	49,613,041	38,089,465	34,862,620	28,292,801	14,585,376
RNA-Seq_Bulk+.45m_Replicate2	44,909,136	36,728,547	33,418,549	27,130,123	13,790,512
RNA-Seq_Bulk-.45m_Replicate3	34,778,120	28,805,697	25,701,984	20,735,028	10,300,104
RNA-Seq_Bulk+.45m_Replicate3	38,478,261	31,791,879	28,552,843	22,892,169	11,105,652
RNA-Seq_Bulk-.150m_Replicate1	27,868,072	23,972,357	21,394,247	16,851,601	8,676,209
RNA-Seq_Bulk+.150m_Replicate1	26,658,068	22,617,163	20,091,003	15,853,121	8,159,913
RNA-Seq_Bulk-.150m_Replicate2	32,877,093	28,625,803	24,630,015	20,498,062	7,858,541
RNA-Seq_Bulk+.150m_Replicate2	43,956,862	36,062,358	32,031,576	25,705,642	13,608,776
RNA-Seq_Bulk-.150m_Replicate3	43,300,980	34,752,510	30,531,707	24,203,782	12,174,210
RNA-Seq_Bulk+.150m_Replicate3	28,298,099	24,641,341	21,585,269	17,172,045	9,144,638
RNA-Seq_Bulk-.300m_Replicate1	16,988,863	14,742,949	12,682,563	9,802,960	5,476,650
RNA-Seq_Bulk+.300m_Replicate1	17,611,739	15,229,185	12,931,531	9,946,844	5,277,507
RNA-Seq_Bulk-.300m_Replicate2	70,100,153	57,573,030	50,1545,69	39,814,131	19,158,438
RNA-Seq_Bulk+.300m_Replicate2	65,157,082	54,971,189	47,505,259	37,796,749	17,675,498
RNA-Seq_Bulk-.300m_Replicate3	46,537,811	37,127,869	31,731,352	24,633,882	11,766,716
RNA-Seq_Bulk+.300m_Replicate3	46,021,323	35,470,650	30,227,804	23,446,113	10,773,527
RNA-Seq_New-.45m_Replicate1	29,684,736	25,442,621	24,085,018	19,635,313	15,097,110
RNA-Seq_New+.45m_Replicate1	29,456,188	25,509,629	24,120,530	20,2692,34	15,223,875
RNA-Seq_New-.45m_Replicate2	12,301,705	10,961,506	10,003,171	8,680,830	6,451,388
RNA-Seq_New+.45m_Replicate2	10,918,076	9,672,355	8,824,706	7,622,637	5,768,912
RNA-Seq_New-.45m_Replicate3	10,584,920	9,427,947	8,468,926	7,308,893	5,284,751
RNA-Seq_New+.45m_Replicate3	11,379,971	6,466,619	5,720,072	4,947,292	3,848,484
RNA-Seq_New-.150m_Replicate1	25,782,008	22,177,870	20,313,700	17,400,998	12,680,956
RNA-Seq_New+.150m_Replicate1	27,330,416	23,598,431	21,915,606	18,8422,75	13,235,645
RNA-Seq_New-.150m_Replicate2	11,721,621	10,516,999	9,795,270	8,488,578	6,611,862
RNA-Seq_New+.150m_Replicate2	9,021,246	7,914,091	7,358,967	6,380,545	5,130,642
RNA-Seq_New-.150m_Replicate3	40,389,199	31,631,571	29,141,164	25,134,018	16,790,835
RNA-Seq_New+.150m_Replicate3	44,139,330	35,905,092	33,010,442	28,551,698	18,430,709
RNA-Seq_New-.300m_Replicate1	21,449,714	18,007,751	15,808,612	13,496,285	8,871,606
RNA-Seq_New+.300m_Replicate1	23,347,784	19,834,397	17,575,864	15,037,394	9,610,599
RNA-Seq_New-.300m_Replicate2	8,448,112	7,115,621	6,604,358	5,691,671	4,671,999
RNA-Seq_New+.300m_Replicate2	8,920,158	7,687,929	7,212,025	6,119,567	4,969,946
RNA-Seq_New-.300m_Replicate3	30,232,380	26,665,295	24,120,631	20,608,773	11,339,210
RNA-Seq_New+.300m_Replicate3	34,139,012	29,639,079	26,839,351	22,783,241	12,097,176

In brief, at the same time of E2 stimulation we started to label the newly synthesized mRNA using the chemical 4-Thio-Uridine (4sU). At 45 minutes, 2.5 hours and 5 hours after E2 or vehicle treatment, the total (bulk) RNAs were extracted and the newly synthesized RNAs were then separated from the pre-existing RNAs by applying a chemical reaction specific to 4sU (see **Experimental Methods 4.2.6**). Both the 4sU labeled newly synthesized and the bulk mRNA samples were subjected to mRNA-Seq, and for each treatment case we performed three biological replicates (**Table 3**). For bulk mRNA-Seq, after sequencing, around 100.1 and 100.2 million raw reads were generated at 45min after treating with E2 or vehicle respectively. Of these, 82.8 and 80.7 million reads with good sequencing quality (filtered reads) were retained, 74.8 and 73.1 million reads (mapped reads) can be mapped to genome. Among them, 60.3 and 58.9 million reads were mapped at unique genomic positions (uniquely mapped reads). Finally, 30.7 and 30.5 million non-redundant and uniquely mapped reads were obtained and used for the following analysis. At 2.5hr, in total, around 98.9 and 104 million raw reads were generated with E2 or vehicle treatment respectively. Of these, 83.3 and 87.3 million reads (filtered reads) were retained after quality filtering, 73.7 and 76.6 million reads can be mapped to genome. Among them, 58.5 and 61.5 million reads (uniquely mapped reads) can be uniquely mapped. Finally, 30.9 and 28.7 million reads (non-redundant and uniquely mapped reads) were used to call transcripts. At 5hr, in total, around 128.8 and 133.6 million raw reads were generated with E2 or vehicle treatment respectively. Of these, 105.7 and 109.4 million reads (filtered reads) were retained after quality filtering, 90.6 and 94.5 million reads can be mapped to genome. Among them, 71.2 and 74.2 million reads (uniquely mapped reads) can be uniquely mapped. Finally, 33.7 and 36.4 million reads (non-redundant and uniquely mapped reads) were obtained. We used the uniquely mapped reads (including redundant and uniquely mapped reads) to estimate transcript abundance (details see **Computational Methods 4.3.4.**).

For newly synthesized mRNA-Seq, after sequencing, around 51.7 and 52.6 million raw reads were generated at 45min after treating with E2 or vehicle respectively. Of these, 41.6 and 45.8 million reads with good sequencing quality (filtered reads) were retained, 38.7 and 42.5 million reads (mapped reads) can be mapped to genome. Among them, 32.8 and 35.6 million reads were mapped at unique genomic positions (uniquely mapped reads). Finally, 24.8 and 26.8 million non-redundant and uniquely mapped reads were used to identify transcripts. At 2.5hr, in total, around 80.5 and 77.9 million raw reads were generated with E2 or vehicle

treatment respectively. Of these, 67.4 and 64.3 million reads (filtered reads) were retained after quality filtering, 62.3 and 59.2 million reads can be mapped to genome. Among them, 53.8 and 51 million reads (uniquely mapped reads) can be uniquely mapped. Finally, 36.8 and 36.1 million reads (non-redundant and uniquely mapped reads) were used to call transcripts. At 5hr, in total, around 66.4 and 60.1 million raw reads were generated with E2 or vehicle treatment respectively. Of these, 57.1 and 51.8 million reads (filtered reads) were retained after quality filtering, 51.6 and 46.5 million reads can be mapped to genome. Among them, 43.9 and 39.8 million reads (uniquely mapped reads) can be uniquely mapped. Finally, 26.7 and 24.9 million reads (non-redundant and uniquely mapped reads) were generated. As described above, we used those uniquely mapped reads (including redundant and uniquely mapped reads) to estimate transcript abundance.

In total, 1054 million sequencing reads were generated. Among these, 632 million (60%) could be mapped to unique position at human genome (Hg19) (**Table 3**). Based on these reads, the expression level of 18842 RefSeq genes were estimated by the number of reads that can be mapped on their exonic regions divided by the accumulative exon length minus background read density (See **Computational methods 4.3.4**). Between replicates, the gene expression level estimated based on our methods is highly reproducible (**Figure 11A**).
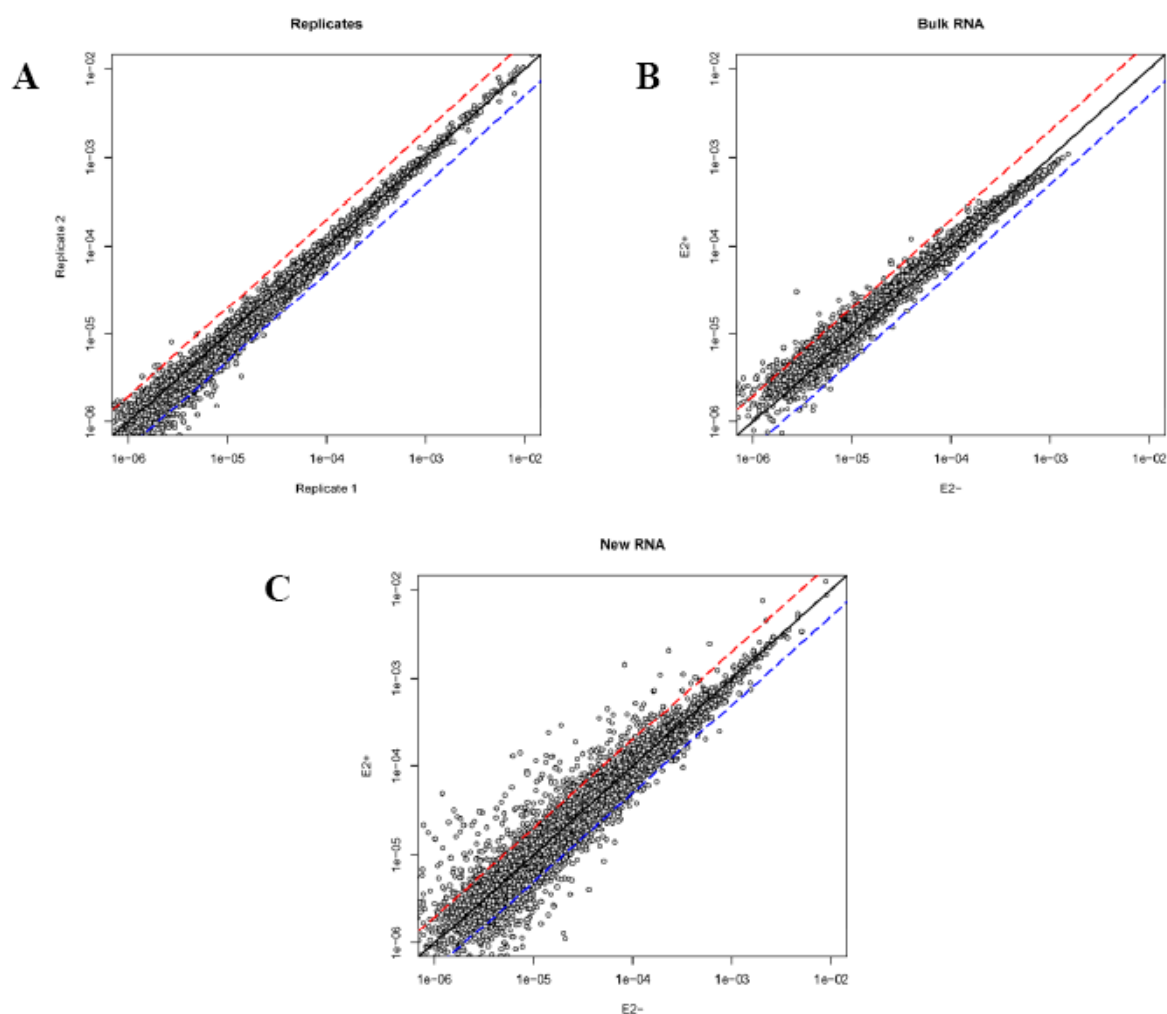
Figure 11. Overview of mRNA-Seq data after expression determination. (A) Gene expression level at 5hr after E2 treatment estimated by using newly synthesized mRNA-Seq was highly reproducible between two replicates. X and Y axes are $\log_{10}$-transformed total read counts within each transcript. Gene expression changes (E2 *vs* vehicle treatment at 2.5hr) were detected by using bulk mRNA-Seq (B) as well as by using newly synthesized mRNA-Seq (C), indicating improved detection sensitivity by newly synthesized RNA. X and Y axes represent $\log_{10}$-transformed total read counts within each transcript for vehicle and E2 treatment, respectively.

Based on bulk RNA profiling, we identified 137, 299 and 395 upregulated genes and 30, 81 and 208 downregulated genes after treating with E2 for 45min, 150min and 300min, respectively. Consistent with our expectation, profiling newly synthesized RNA allowed us to find more differentially expressed genes, particularly in earlier time points. In total, based on the results of profiling new synthesized RNA, i.e. 230, 352 and 395 genes, and 56, 149 and 208 genes were found to be up and downregulated at 45, 150 and 300min, respectively. Depending on the dynamics of their response, a total of 713 E2-regulated protein-coding

47

transcripts could be clustered into six groups with distinct expression dynamics (see **Figure 12**, appendix **Table 4** for the gene list). 471 and 242 genes were upregulated and downregulated after E2 treatment, respectively. Among the upregulated genes, 76 were upregulated at 45min but showed no changes at later time points (initial up), 167 were upregulated at 45min and thereafter (early up), and 228 were upregulated only after 2.5hr (late up). Similarly, among the downregulated genes, 34 were downregulated at 45min but showed no changes at later time points (initial down), 26 were downregulated at 45min and thereafter (early down), and 182 were downregulated only after 2.5hr (late down). The upregulation of 93 and 53 genes at 45min and 2.5hr was only apparent based on the sequencing of newly synthesized RNA (**Figure 12**), which again demonstrated that profiling newly synthesized RNA provided a higher sensitivity in detecting differential gene expression (**Figure 11B and 11C**).
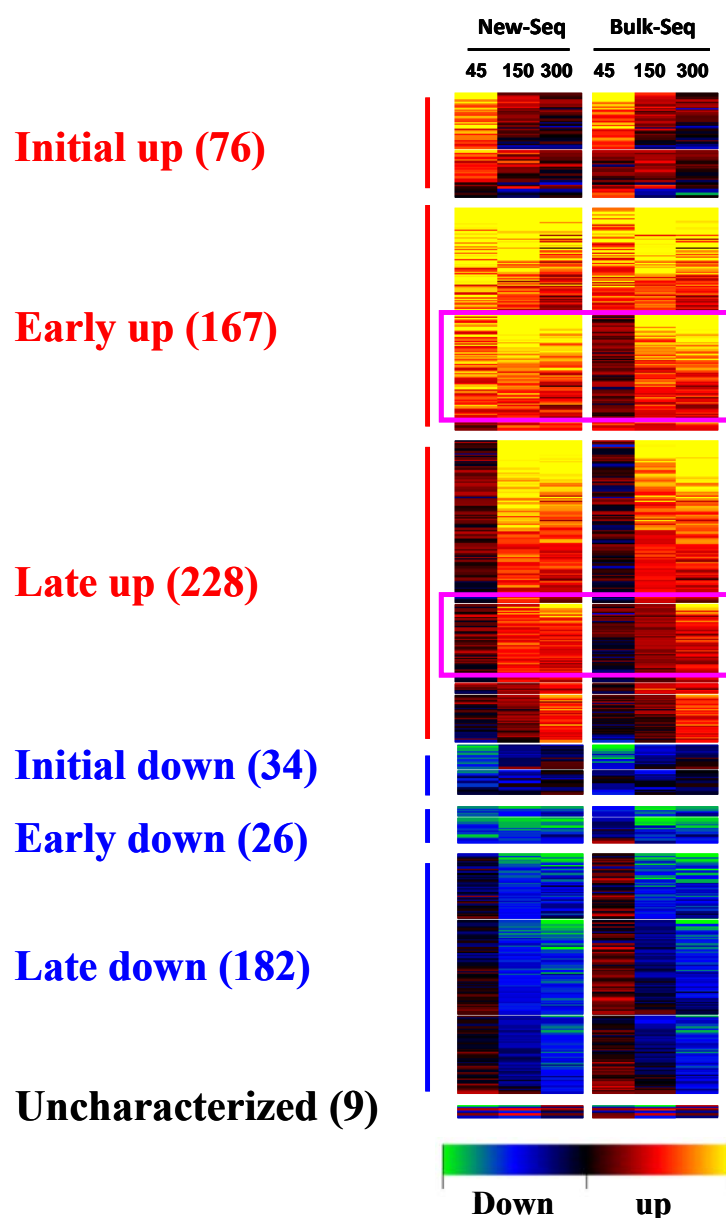
Figure 12. Heatmap of all the 722 E2 regulated genes. A total of 713 E2-regulated genes (defined as genes differentially expressed in at least one of the three time points either by using newly synthesized mRNA-Seq or by using bulk mRNA-Seq) could be clustered into six groups, depending on the dynamics of their response. Initial up group includes 76 genes upregulated at 45min but showed no changes at later time points. Early up group includes 167 genes upregulated at 45min and thereafter. Late up group includes 228 genes upregulated only after 2.5hr. Initial down group includes 34 genes downregulated at 45min but showed no changes at later time points. Early down group includes 26 genes downregulated at 45min and thereafter. Late down group includes 182 genes downregulated only after 2.5hr. The upregulation of 93 and 53 genes at 45min and 2.5hr was only apparent based on the sequencing of newly synthesized RNA (**pink** rectangle), which demonstrated that profiling newly synthesized RNA provided a higher sensitivity in detecting differential gene expression.

5.4. Identification of a Novel ERα Target Gene, ACE

Of the 722 potential ERα target genes identified in this study, many have already been found as the bona fide ERα target genes in previous studies, such as *NRIP1* (Carroll et al., 2006; Lin CY et al., 2004), *GREB1* (Lin CY et al., 2004), *TFF1, CCND1, ABCA3* and etc. Here, we focused on a novel target gene, ACE. ACE is a key enzyme of rennin agiotensin system (RAS), which is of great importance in regulating vasoconstriction. Several lines of evidence have shown that E2 treatment can regulate ACE activity by repressing its gene expression (Gallagher et al., 1999). It has been postulated that regulation of ACE by E2 treatment is mediated by the interaction of ERα with the Fos-Jun heterodimer at an AP-1 consensus binding site at the promoter (Wanda et al., 2010); however, the precise mechanism of action is still unclear.

In this study, first, we found that ACE expression was highly upregulated based on both newly synthesized and bulk mRNA-Seq results (**Figure 13A**). Indeed, it belongs to the early upregulated gene group. Second, we identified three ERα binding site that might be involved in the transcriptional regulation of ACE by ERα. Based on the genomic location of these ERα binding sites with respect to ACE gene locus, we postulated them as putative enhancers and designated them as ACE-EN1, ACE-EN2 and ACE-EN3 respectively (**Figure 13B**). To validate the binding of ERα at these three sites, we performed ERα ChIP using antibody against ERα at 45min, 2.5hr, 5hr, 10hr and 24hr with E2 treatment, and then carried out real time quantitative PCR (qPCR) to check the enrichment of the three sites together with a known binding site around the TSS of GREB1 as a positive control and a non-binding site within the second exon of myoglobin as a negative control. As shown in **Figure 13C**, the known ERα binding site GREB1-EN displayed strong binding along the time course with the maximum binding at 2.5hr. Among all the three ERα binding sites associated with ACE, only ACE-EN2 displayed a relatively strong binding with the maximum binding reaching at 10 hours after E2 treatment. We therefore hypothesized that ACE-EN2 site might be the leading *cis*-regulatory element to modulate ACE gene expression via estrogen receptor. Finally, we checked for potential physical interaction among the three sites associated with ACE in a publicly available ERα ChIA-PET data, which consists of genome wide inter- or intra-chromosomal physical interactions mediated by ERα (Fullwood et al., 2009). Indeed, we found that the three ERα binding sites have intensive intra-chromosomal interactions

mediated by ERα suggesting the involvement of all the three ERα binding sites in modulation of ACE gene expression (**Figure 13D**).
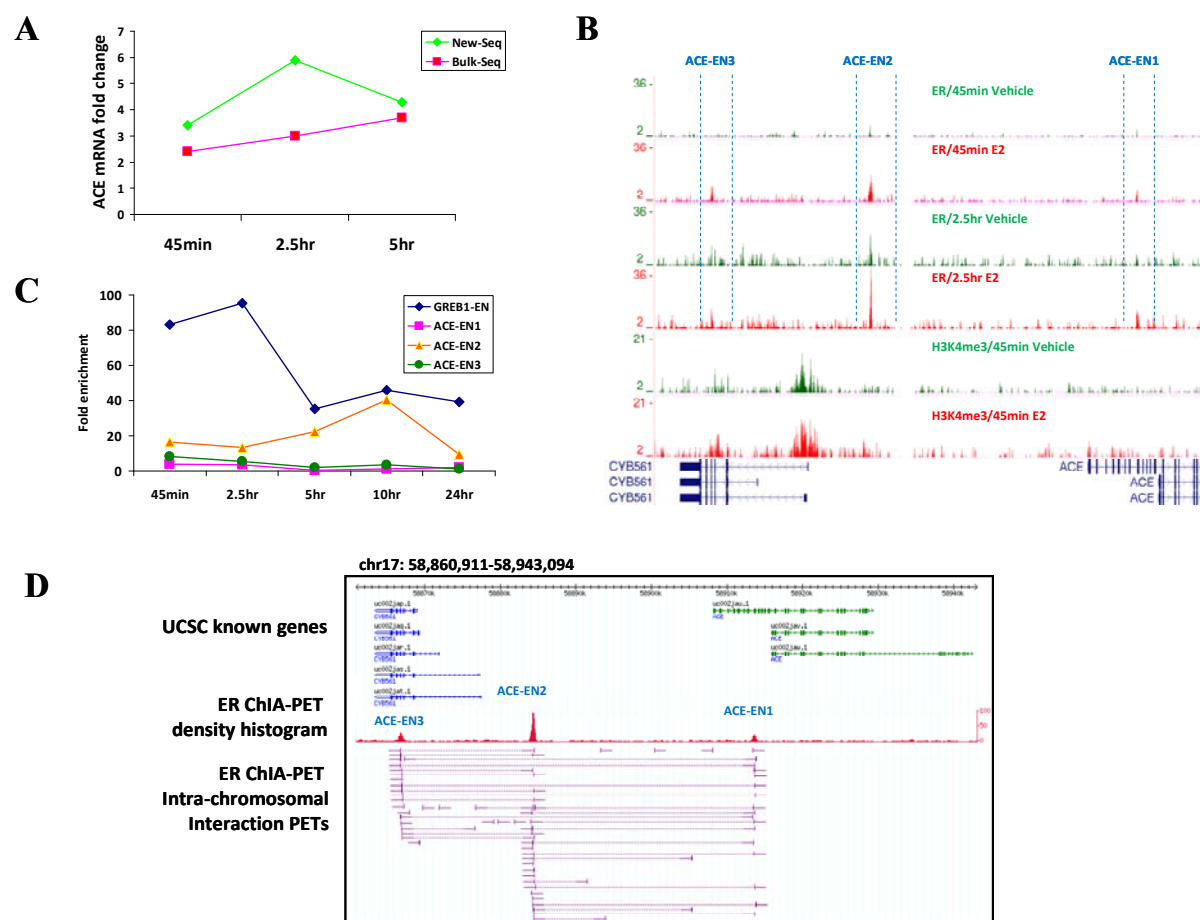
Figure 13. *ACE* **gene expression is regulated by ERα.** (A) *ACE* mRNA is highly upregulated by the E2-ERα complex in both newly synthesized and bulk mRNA-Seq data. Actually, it belongs to the early upregulated gene group with maximum upregulation at 2.5 hrs (B) UCSC genome browser screenshot of ERα binding and H3K4me3 occupancy. ERα binding sites after E2 or vehicle treatment for 45min or 2.5hr are represented as density histogram (**red** or **green**). Similarly, H3K4me3 occupancy after E2 or vehicle treatment for 2.5hr is represented as density histogram in red and green respectively. ACE-associated ERα binding sites are denoted ACE-EN1, ACE-EN2 and ACE-EN3 respectively (**blue**). (C) ERα binding occupancy at ACE-EN1, -EN2, and -EN3 loci, together with GREB1 enhancer (GREB1-EN) working as a positive binding control, over a long time course of E2 treatment is investigated by ChIP followed by real time quantitative PCR (ChIP-qPCR). Among all the three ERα binding sites associated with ACE, only ACE-EN2 displayed a relatively strong binding with the maximum binding reaching at 10 hours after E2 treatment. (D) Screenshot of ERα ChIA-PET analysis from Fullwood et al., 2009 showing ERα binding and long-range chromatin interactions between the *CYB561* and *ACE* gene loci. ERα binding sites are represented as density histograms (**red**) and long-range chromatin interactions are represented as intra-chromosomal interaction PETs (**pink**).

5.5. Identification of E2 Regulated Novel Transcripts

Recent studies have reported that thousands of large non-coding transcripts including large intergenic non-coding RNAs (lincRNAs) have been identified from the human genomes using tiling microarrays, shortgun sequencing of expressed sequence tags (ESTs) and histone modification signatures (Cabili et al., 2011; Guttman et al., 2011). These lincRNAs have been implicated in many cellular processes such as gene expression regulation and epigenetic regulation via interacting with chromatin modifier proteins (Guttman et al., 2011).

Growing evidence has shown that lincRNAs can be dysregulated in leukemia or solid tumors including breast tumors (Calin et al., 2007). For instance, the well-known lincRNA HOTAIR showed increased expression in primary as well as metastatic breast tumors, and depletion of HOTAIR inhibited the matrix invasiveness of the breast cancer cell line MCF-7 (Gupta et al., 2010). Hence, identification of novel E2 regulated intergenic transcripts previously uncharacterized in breast cancer cells is also crucial in unwinding ERα mediated transcriptional regulatory network.

To detect E2 regulated intergenic transcripts, we first mapped all the short sequencing reads that we have obtained so far in our mRNA-Seq and identified regions with expression level significant higher than background (referred as qips regions) by qips software (see **methods**). We then extracted the potential novel intergenic transcripts by excluding those overlapped with known RefSeq genes (see **methods**). After estimating the expression changes of qips regions using the same strategy in determining that of RefSeq genes, we identified 77 E2 regulated intergenic regions. Next, to join the qips regions originated potentially from the same transcripts, we exploited paired-end RNA-Seq data (**Table 4**) and merged those qips regions that can be linked with paired reads. This procedure resulted in 68 E2 regulated intergenic transcripts. Among these 68 transcripts, 14 transcripts overlapped with UCSC genes, 31 transcripts with EST and mRNA sequences deposited in Genbank, 7 transcripts with lincRNAs identified by Khalile et al (Khalile et al., 2009). The remaining 29 E2 regulated intergenic transcripts do not overlap with any annotated genomic features. (See appendix **Table 6**).

Table 4. Summary of paired-end RNA-Seq results.

Dataset	Raw paired reads	Filtered paired reads	Uniquely mapped F reads	Uniquely mapped R reads	Concordant pairs	Concordant pairs, ≤ 1Mb distance	Concordant pairs, ≤ 200Kb distance
Paired-End RNA-Seq	72,013,327	68,975,593	58,237,204	58,166,459	51,858,243	51,794,425	51,789,624

Here, we focused on novel intergenic transcripts potentially regulated directly by ERα, i.e. the transcripts 1) with gene expression change after E2 treatment; 2) with an ERα binding site close by (within 5Kb of window at both ends of transcripts or inside). In total, there were 14 novel intergenic transcripts having a peak within 5Kb from both ends or inside. As one example shown in **figure 14**, a novel intergenic transcript located upstream of protein coding gene ARHGAP12 showed rapid and transient increase of expression after 45min of E2 treatment and tended to diminish afterwards. The upregulation of this intergenic transcript was much more obvious in newly synthesized RNA-Seq data than that in bulk RNA-Seq data. The direct regulation of this transcript by ERα is implicated by a clear ERα binding within its locus (**Figure 14)**. Furthermore, the active transcription of this region was also supported by H3K4me3 occupancy within its locus, which might designate the transcription start site.
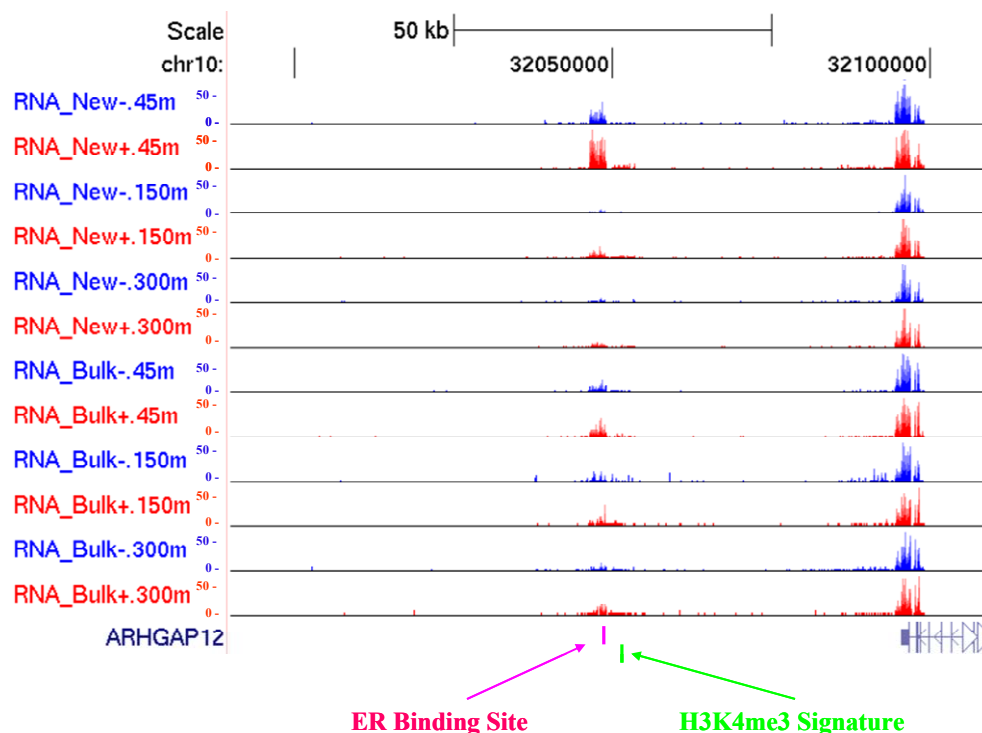
Figure 14. **An example of E2 regulated novel intergenic transcript.** A novel intergenic transcript locates upstream of *ARHGAP12* gene locus. It showed rapid and transient upregulation after 45min of E2 treatment and tended to diminish afterwards. This upregulation of this intergenic transcript was much more obvious in newly synthesized RNA-Seq data than that in bulk RNA-Seq data. The direct regulation of this transcript by ERα is implicated by a clear ERα binding within its locus. Furthermore, the active transcription of this region was also supported by H3K4me3 occupancy within its locus, which might designate the transcription start site.

6. Discussion

The effort to uncover the mechanisms by which ERα dictates gene expression has been made mainly from two aspects, which are the profiles of ERα binding sites and gene expression changes. In recent years, several extensive genome-wide studies with both aspects have greatly advanced our understanding of ERα mediated transcriptional regulatory networks, revealing a series of new features previously uncharacterized, such as the majority of ERα binding sites preferentially located within non-promoter genomic regions and distinct temporal mechanisms of E2 regulated gene expression (Carroll et al., 2006; Lin et al., 2007; Welboren et al., 2009). However, the findings by the different studies often showed only limited overlap. For example, by using different types of ChIP platforms (i.e., ChIP-on-chip, ChIP-DSL, ChIP-PET, and ChIP-Seq), several previous studies globally mapping ERα binding sites have generated a range from ~1000 to more than 10,000 sites, with limited overlap (Carroll et al., 2006; Kwon et al., 2007; Lin et al., 2007; Welboren et al., 2009). Similarly, the number of E2 responsive genes in MCF-7 cells that were identified by several research groups using microarray platform ranges from 100 up to 1500 (Cheung et al., 2010).

In this study, we demonstrated that the overlapped findings in ERα binding sites often reflect strong or stable ERα-DNA interaction whereas most of weak or transient binding might be detected in an irreproducible manner. Therefore, in order to, on one hand characterize as many as possible ERα binding sites, on the other hand to avoid the potential artifacts, we analyzed our own replicates of ERα ChIP-Seq experiments, in combination with another three published ERα ChIP-Seq datasets (Cicatiello et al., 2010; Hu et al., 2010; Welboren et al., 2009). In addition to the binding sites identified in at least two of our replicates, we also retained the binding sites identified in only one of our replicates, but present in at least one of the three published datasets. As a result, we obtained a list of 4543 high-confident ERα binding sites at 45min after E2 treatment. Interestingly, in contrast to most previous studies, we observed a significant number of ERα binding events in the absence of E2.

In order to determine the transcriptional response induced by E2 stimulation with a higher precision, we performed gene expression profiling by the state-of-art RNA-Seq method. More specifically, to separate the effects of transcriptional regulation from that of post-transcriptional events, we measure the gene expression by sequencing both the bulk RNAs

and RNAs newly synthesized at different time intervals after E2 treatment. Our expression profiling results demonstrated that profiling newly synthesized RNA provided a higher sensitivity in detecting differential gene expression, and thereby allowed more efficiently to identify the genes under direct transcription regulation. Finally, to investigate the impact of ERα binding on the local chromatin configuration, we globally profiled the H3K4me3 in MCF-7 cells before and after E2 treatment, but observed no obvious changes in H3K4me3 pattern induced by E2 stimulation.

6.1. Significant ERα-DNA Interaction in the Absence of E2

ERα binding pattern without E2 treatment might have an important role in determining the ERα mediated transcriptional regulatory network. To date, most studies in global mapping of ERα binding sites have focused only on those after E2 stimulation. The properties of ERα binding sites in the absence of E2 treatment have been seldom described and some studies on ERα binding sites even treated the profiling of binding sites without E2 stimulation as the negative control.

In contrast, our study has obtained 992 ERα binding sites in the absence of E2. By further analysis on the 992 sites, we found that 66% are also present after E2 treatment. In general, the 652 binding sites still remaining after E2 treatment contained significantly higher peaks than those 340 sites disappearing after E2 treatment. This suggests, on one hand, the binding sites identified only in the absence of E2 represent are mostly weak or transient ERα-DNA interactions; On the other hand, a significant number of strong binding events are still present even after hormone deprivation for three days, which might result from residual ERα remaining in nucleus or entering nucleus via alternative pathways. To investigate whether the two types of ERα binding were via distinct mechanisms, we search for sequencing motifs enriched by the two groups of binding sites. As shown in **figure 15A**, for the binding sites retained after E2 treatment, the most enriched sequence element was clearly the well-known ERE motif. In contrast, such motif is not enriched for the binding sites only present in absence of E2. Instead, sequences corresponding to tumor suppressor p53 and other transcription factor binding motifs were moderately enriched for this group of binding sites (**Figure 15B**). Our findings implicated the potentially different molecular mechanisms underlying the two types of ERα binding. Whereas the strong binding events that are still present even after

hormone deprivation were mostly direct ERα binding via classic ERE elements, the weak or transient ERα-DNA interactions that are only present in the absence of E2 were mostly indirect ER binding via other protein cofactors. Whether such hypothesis holds true still awaits further experimental validation, such as identification of protein interaction partners of ERα in the presence or absence of E2 stimulation.
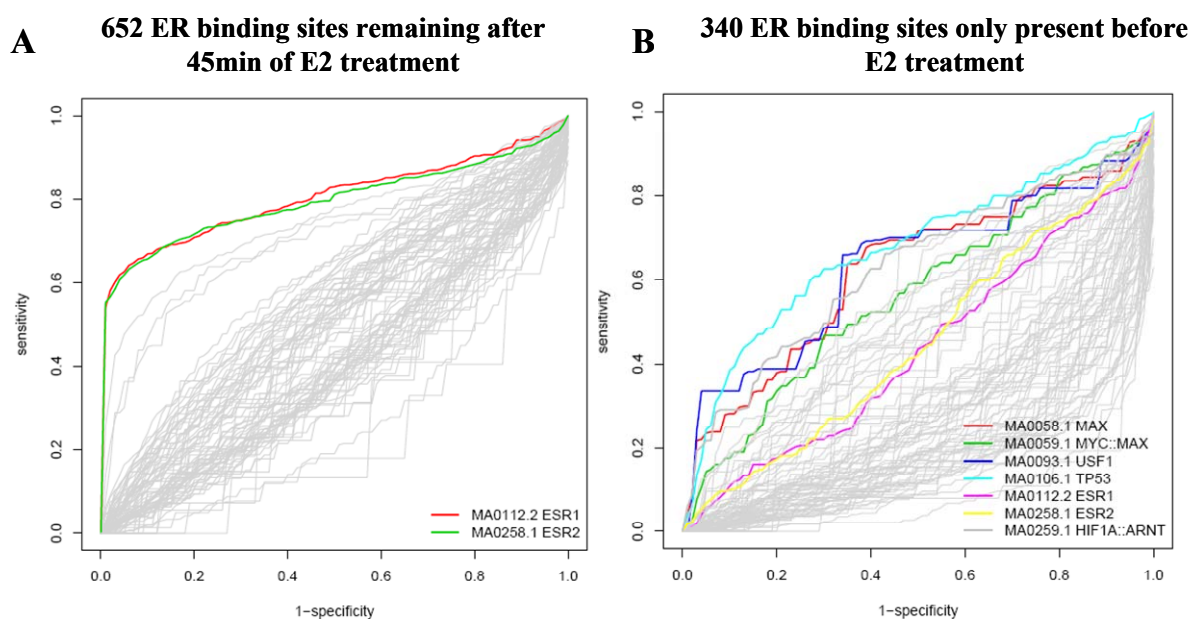
Figure 15. Motif analysis by ROC curve. Enrichment of transcription factor binding motifs at ERα binding sites consistently present before and after 45min of E2 treatment (A) and those uniquely present before 45min of E2 treatment (B). Specifically, in figure A, the most enriched motif is ERE as illustrated by ESR1 and ESR2. But in figure B, TP53 and some other transcription factors motifs (such as MYC and MAX) are more represented than ERE motif.

6.2. Maximum ERα Binding Occurs at 45 min

Previous studies have demonstrated that maximum ERα binding happens at 45min of E2 stimulation in MCF-7 cells based on analysis on limited number of sites (Carroll et al., 2005). Thereafter, numerous studies of ERα binding sites (Carroll et al., 2005 and 2006; Lin CY et al., 2007; Fullwood et al., 2009; Welboren et al., 2009) always used this time point without validating such a notion at the global level. In this study, we profiled the ERα binding sites at both 45min and 150min after E2 stimulation. As illustrated in the **figure 7C** (**result section**), nearly all the ERα binding sites identified at 150min after E2 treatment have been found at 45min after E2 treatment. We believed the vey few ERα binding sites not present in 45min datasets are those with low binding affinity and therefore could not been captured even in

different biological replicates. To prove this hypothesis, we collected three datasets of ERα binding sites that were identified in the first "45min" replicates and two "150min" replicates. After clustering the ERα binding sites into different bins according to their peak heights (**Figure 16**), we calculated for each bin the percentage of binding sites that could be found also in replicates 2 and 3 of the three ERα binding site datasets at 45min of E2 treatment. As shown in **figure 16**, nearly all binding sites present at 150min with high peak heights could be found in the "45min" dataset, whereas those exclusive "150min" ERα binding sites were often with low peak height. At the same peak height (at 10), even higher percentage of "non-replicable sites" were found between "45min" replicate experiments.
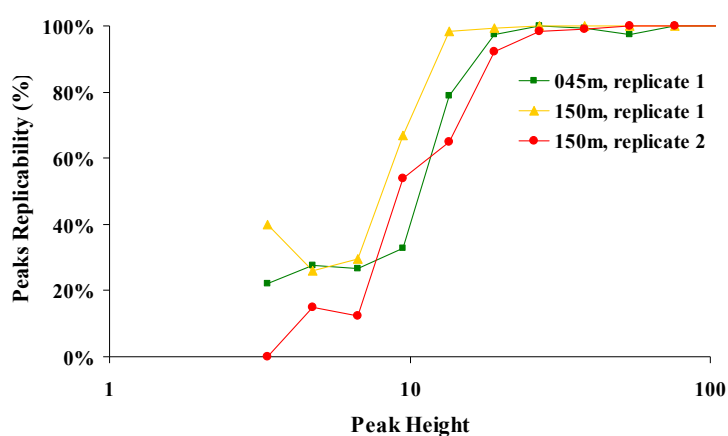
Figure 16. The overlap between ERα binding sites identified at 150 min and those at 45 min after E2 treatment. The peak height (x axis) indicates the binding affinity of ERα binding sites, while the y axis is marked as the percentage of binding sites with different peak height from "45min replicate 1" or the two "150min" datasets that also found in "45min" replicate 2 and 3. Nearly all binding sites present at 150min with high peak heights could be found in the "45min" dataset, whereas those exclusive "150min" ERα binding sites were often with low peak height. At the same peak height (at 10), even higher percentage of "non-replicable sites" were found between "45min" replicate experiments.

6.3. E2 Stimulation does not Induce H3K4m3 Pattern Changes

The important role of histone modifications in cell differentiation and cancer tumorigenesis has already been extensively studied and a variety of histone modifications have been implicated in gene expression regulation. In this study, we sought to investigate the effect of E2 stimulation or rather of ERα binding on H3K4me3 pattern. As shown in **figure 9C** (**results section**), we detected no significantly higher changes for H3K4me3 occupancy at the genomic level after E2 treatment for 45min, as compared with the difference between two replicates at 45min after E2 stimulation. Indeed, most of the changes identified in one

experiments could not be replicated in the replicates (**Figure 17**). This implicated that E2 stimulation might not induce H3K4me3 pattern changes. Alternatively, the absence of changes at the global level could also be due to the fact that E2 regulated genes only account for 3.8% of all genes expressed in MCF-7 cells. To check for the latter possibility, we restricted our comparison of H3K4me3 profile to the promoters of E2 regulated genes. As shown in **figure 9A** (see **results section**) and **figure 17**, apparently no significant changes could be found also for the genes regulated by E2 stimulation.
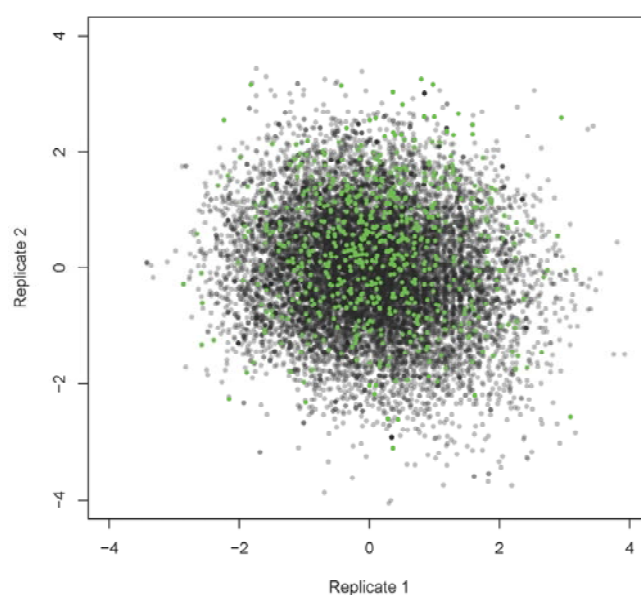
Figure 17. Comparison of the H3K4me3 occupancy changes at 45min after E2 treatment identified in two replicates. X and Y axes represented H3K4me3 occupancy changes indicated by the Z values detected in Replicate 1 and Replicate 2. The E2-regulated genes were marked in **green**.

6.4. Newly Synthesized mRNA-Seq Has a Higher Sensitivity in Detecting Primary Gene Targets

It is believed that the primary transcription events regulated by E2 would arise rapidly after E2 induction, and with extended time, the secondary transcriptional responses conferred by other factors would develop and confound the primary transcriptional effects directly induced by the E2-ERα complex. However, in a number of previous studies in E2 regulation, gene expression changes were monitored after a relatively long E2 treatment time (~3-24hr), when the changes in mRNA level were accumulated above a detectable threshold (Carroll et al.,

2006; Cicatiello et al., 2010). As a result, their lists of differentially expressed genes are supposed to be a mixture of primary and secondary transcription effects. To rule out the potential impact from the secondary effects, Lin et al. (2004) tried to define the immediate transcriptional effects of E2 signaling by inhibiting translation with cycloheximide in T-47D cells and reported only 89 genes as primary targets in E2 signaling. Although the study appeared intriguing, it did not take into account the severe side effects caused by the drug. Moreover, the mRNA levels are dependent not only on the transcription, but also on the degradation. The translation inhibitor cannot separate the regulation of the former from that of the latter.

In this study, in order to obtain a more direct readout for transcriptional regulation, we pulse labeled newly synthesized mRNA with 4sU for 45min, 2.5hr and 5hr after E2 or vehicle treatment. Although a majority of the differentially expressed genes could be identified based on both profiling of bulk RNAs and profiling of newly synthesized RNAs, the significant expression changes could be detected at earlier time points if we used the profiling of newly synthesized RNA. For example, based on newly synthesized RNA sequencing, 230 genes were identified to be upregulated at 45min after E2 stimulation, whereas based on bulk RNA sequencing, 137 genes were identified to be upregulated at the same time points. As shown in the next section, compared with those upregulated at later time points, a significant higher proportion of the genes upregulated at 45min were found to have an ERα binding sites at the promoter proximal regions, and thus are potential direct regulatory targets of ERα. Therefore, detection of expression changes at earlier time points using newly synthesized RNA sequencing could facilitate the identification of direct targets in E2 signaling.

6.5. The Relationship of ERα Binding Sites to E2 Regulated Gene Expression Changes

On one hand, not all the transcription factor binding events have a functional consequence, On the other hand, not all the genes showing expression changes are under the direct regulation of corresponding transcription factors. To detect direct target gene under ERα transcriptional regulation, we correlated the ERα binding events with E2 induced gene expression changes by compiling a list of genes with expression changes and harboring at least one ERα binding site within a distance of 5Kb away from their TSSs. As shown in

figure 18A, 21% of 713 E2 regulated genes (containing 874 alternative TSSs) have in total 151 promoter proximal ERα binding site, accounting for 23% of all such promoter proximal binding sites. Moreover, compared with downregulated genes, of which 5% having ERα binding sites close by, a significantly higher proportion (37%) of the upregulated genes contain promoter proximal ERα binding sites. Furthermore, to correlate the ERα binding with gene expression response dynamics, we compared the distance of the TSSs to the closest ERα binding sites among the six gene groups. As shown in **figure 18B**, again, in general, the upregulated genes have significantly closer ERα binding sites ($p < 2.2 \times 10^{-16}$ by Mann Whitney test), whereas the downregulated genes cannot be distinguished from the non-regulated genes in this regard. Interestingly, among the upregulated genes, 43% of early upregulated genes have at least one promoter proximal ERα binding sites, which is significant higher (Fisher's exact test, $p < 2.2 \times 10^{-16}$) than initial upregulated gene (35%), and late upregulated genes (30%) ($p < 2.2 \times 10^{-16}$ by Fisher's exact test) (**Figure 18C**). In addition, the upregulated genes show also higher enrichment of distal ERα binding sites (beyond 10Kb window around gene TSS), compared to the downregulated genes. Again, among the upregulated gene groups, the early upregulated genes show the highest enrichment.
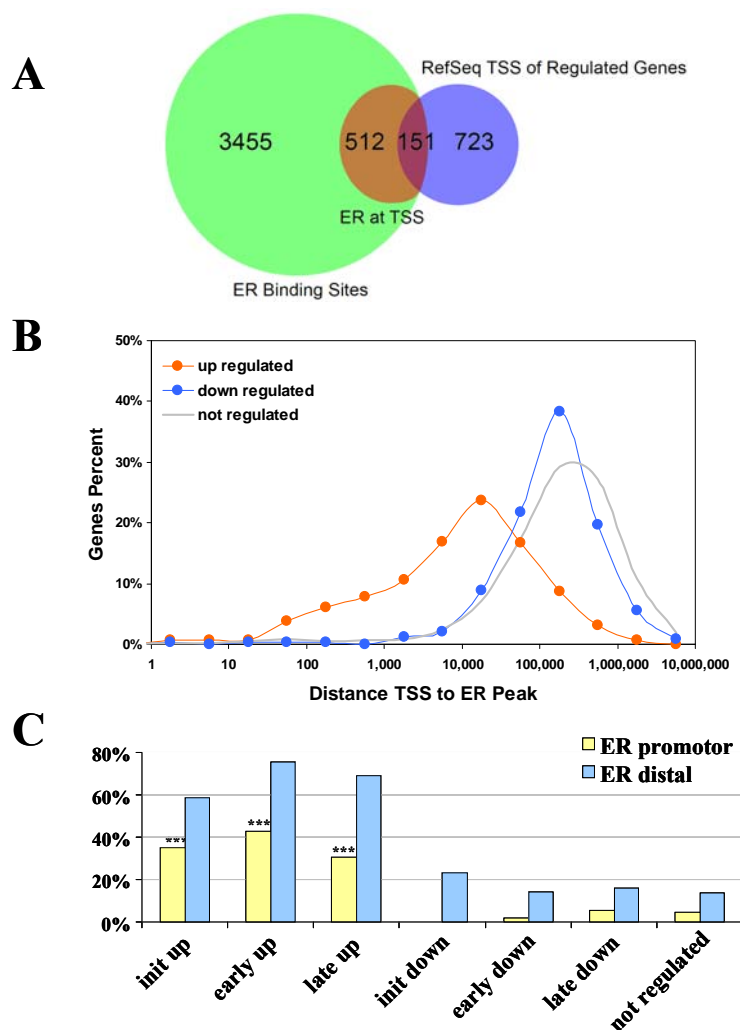
Figure 18. The relationship between ERα binding sites and E2 regulated gene expression changes. (A) The Venn diagram shows the overlap of the TSSs of E2 regulated genes with promoter proximal ERα binding sites (within 5Kb of TSSs). (B) The fractions of upregulated genes, downregulated genes and non-regulated genes with the distances of their TSSs to the closest ERα binding sites are plotted. The upregulated genes have significantly closer ERα binding sites ($p < 2.2 \times 10^{-16}$ by Mann Whitney test), whereas the downregulated genes can not be distinguished from the non-regulated genes in this regard. (C) The fraction of the six groups of E2 regulated genes with a promoter proximal or distal ERα binding site. Yellow bars marked with three black stars designate statistical significance by Fisher's exact test.

6.6. Potential Pathophysiological Relevance of Newly Detected ERα Regulated Genes

ERα target genes can induce the massive downstream pathway cascades, finally eliciting the cellular response to E2 stimulus in breast cancer cells. Hence, the knowledge on ERα direct target genes is a crucial step in understanding pathophysiological effect of E2 signaling. In

this study, we combined gene expression profiling and ERα binding sites, to identify a much more reliable list of ERα target genes. Here, I focused on three novel target genes with putative association with potential pathophysiological relevance.

We define Sin3B as an ERα target gene, because 1) two ERα binding sites were found within 10Kb of its TSS and 2) ~1.5 fold change of gene expression were detected based on both newly synthesized RNA and bulk RNA sequencing at 150min after E2 treatment (**Figure 19A**). Previous findings (Farias et al., 2010) have indicated that Sin3B could be integrated into a functional complex acting as histone deacetylase and was mainly involved in transcription repression. Interactions between the PAH functional domain of Sin3B and the SID site of other transcription factors, such as Mad1 (Silverstein et al., 2005), are an important cause of E2 and tamoxifen insensitivity in triple-negative MDA-MB-231 breast cancer cells. Moreover, Rampalli et al. reported that tumor suppressor SMAR1 interacts with histone deacetylation complex 1, Sin3 (containing both Sin3A and Sin3B), and pocket retinoblastomas to form a multiprotein repressor complex (Rampalli et al., 2005). This repressor complex can then induce the deacetylation of cyclin D1 promoter, resulting in the repression of cyclin D1, which is one of the frequently overexpressed proteins and one of the commonly amplified genes in breast cancer. Taken together, the intensive interactions of Sin3B with some transcription corepressors as well as transcription factors may imply its potential importance in modulating gene transcription mediated by ERα in response to E2 stimulus in breast cancer cells. The precise molecular mechanism uncovering the interplay between Sin3B and ERα would be an interesting topic in the follow-up studies.

Another novel ERα target gene is Sox3, which together with Sox2 belong to the SoxB1 subgroup of transcription factors. In our study, there is an ERα binding sites located within 5Kb of TSS; simultaneously, Sox3 had 23, 12 and 8 fold change of gene expression based on newly synthesized RNA-Seq after E2 treatment for 45min, 150min and 300min respectively whereas 3 fold changes of gene expression were detected based on bulk RNA-Seq at all the three time points (**Figure 19B**). As well known, Sox2 is one of the key transcription factors required in induced pluripotent stem cells (Zhao et al., 2008), while Sox3 is known to take part in primary neurogenesis as well as in spermatogonial differentiation (Archer et al., 2011; Laronda et al., 2011). Study in *Xenopus* has revealed that Sox3 might function as a tumor suppressor and play an important role in regulating Wnt signaling pathways (Zorn et al.,

1999). To date, no studies have revealed any functional roles of Sox3 in breast cancer cells. However, based on our study, SOX3 is very likely a directly target of E2 stimulation. Therefore, further dissection of the molecular mechanisms underlying the role that SOX3 plays in overall E2 response would be important for better understanding of the regulatory network mediated by E2 in breast cancer cells.
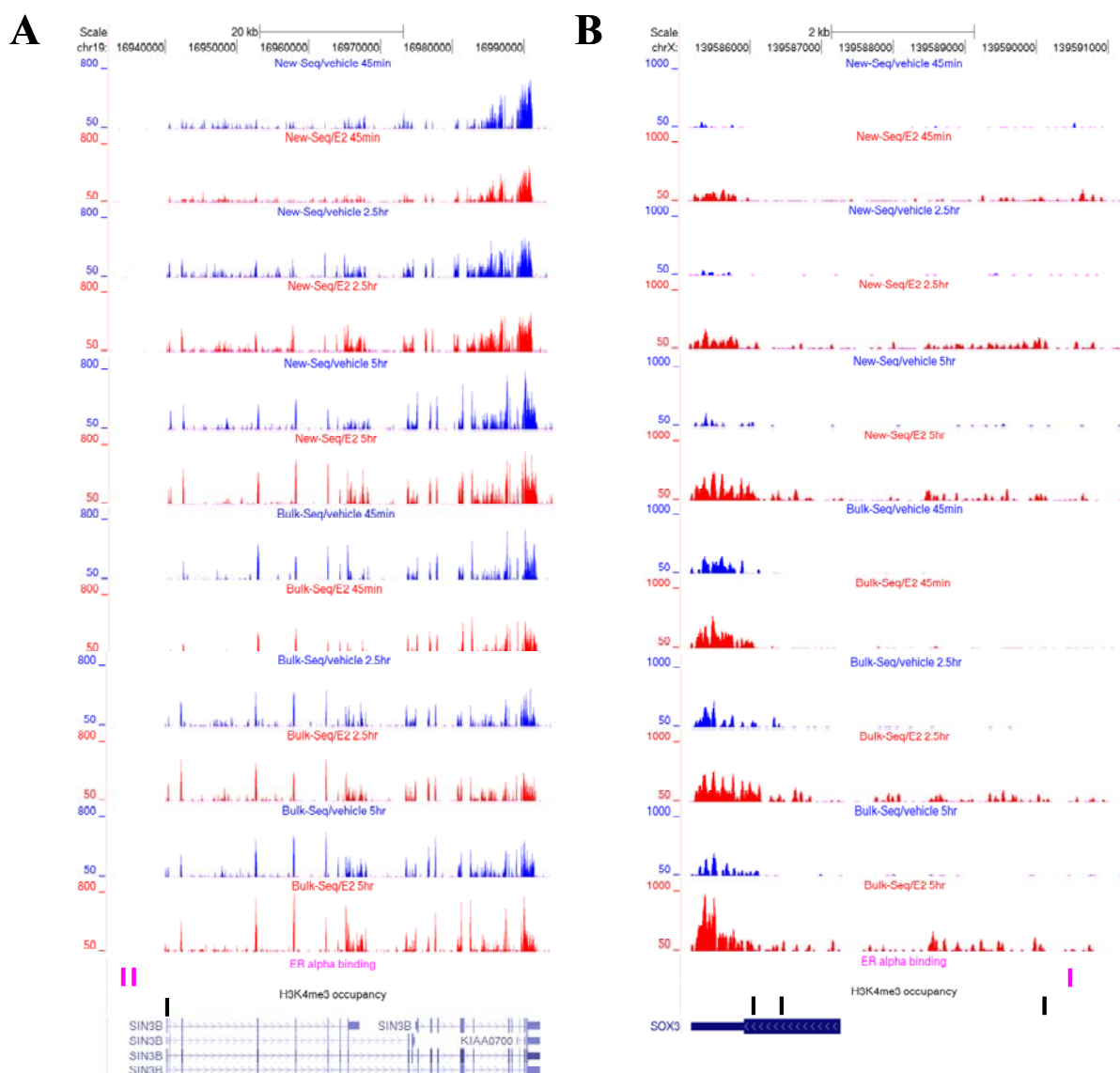
Figure 19. Overview of two novel ERα target genes. Screenshots show the transcription levels estimated based on newly synthesized RNA-Seq and bulk RNA-Seq at all the three time points after treatment on *SIN3B* (A) and *SOX3* (B) as well as ERα binding (**pink** bar) at 45min after E2 treatment and H3K4me3 occupancy (**black** bar) at 2.5hr after E2 treatment.

It is known that the growth of solid tumors including breast cancer is always accompanied by angiogenesis. Angiogenesis, defined as the process of new blood vessel formation, plays a pivotal role in both local tumor growth and distant metastasis in breast cancer (Schneider and

Miller, 2005). Angiotensin II from the renin-angiotensin system (RAS) has been demonstrated to promote angiogenesis by interacting with vascular endothelial growth factor (VEGF), one of the well-known angiogenic factors (Amaral et al., 2001). Both normal and cancerous breast epithelial cells are targets of angiotensin II that functions as growth factors and eventually promotes angiogenesis (Ladd et al., 2011). Angiotensin II is converted from angiotensin I by the angiotensin-converting enzyme (ACE). The roles of ACE and ACE inhibitors have been studied in breast cancer. For instance, *ACE* gene insertion/deletion polymorphism may function as susceptibility factor in breast cancer risk and disease-free survival in Caucasian postmenopausal women. ACE inhibitors have been reported to trigger breast cancer incidence (Ganz et al., 2011); however, the molecular mechanism underlying such correlation is poorly understood. Studies using tissue material from kidney cortex, kidney, medulla, lung, and aorta of ovariectomized Sprague-Dawley rats demonstrated that E2 treatment via ER indeed regulates tissue ACE activity by reducing ACE mRNA level (Gallagher PE et al., 1999). No consensus ERE, but a consensus AP-1 site was found close to the ACE promoter region. Therefore, it has been postulated that ERα together with AP-1 binds to an AP-1 site on ACE promoter to inhibit ACE gene expression (Gallagher et al., 1999). However, the precise mechanism by which the E2-ERα complex regulates ACE gene expression remains elusive.

In this study, ACE was identified as an ERα target gene due to 1) ERα binding sites residing around its gene locus and 2) more than 2 fold change of expression based on both Bulk RNA and newly synthesized RNA-Seq. Our ChIP-qPCR analysis suggested ACE-EN2 may be the putative enhancer for ACE gene regulation, which is supported by ChIA-PET results showing intensive interaction between ACE-EN2, ACE-EN1 and CYB561-EN (**Figure 13D**). It is therefore conceivable that ACE gene regulation by E2-ERα complex on a transcriptional level might be due to the combinatorial binding of ERα on the three sites and subsequent chromatin looping. ACE mRNA level regulated by ERα increased in MCF-7 breast cancer cells, whereas it was reduced in vascular cells. Therefore, distinct cellular context involving cell type-specific ERα coregulators and cooperating TFs may account for diverged ACE gene regulation mediated by ERα. However, these preliminary results previously uncharacterized have provided researchers a new perspective on ACE regulation by ERα in breast cancer cells apart from in vascular cells, and may even shed lights also on the precise mechanism in cardiovascular system.

7. References

Agius P, Arvey A, Chang W, et al. High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. *PLoS Computational Biology* 2010; 6: e1000916.

Amaral SL, Papanek PE, Greene AS. Angiotensin II and VEGF are involved in angiogenesis induced by short-term exercise training. *Heart and Circulatory Physiology* 2001; 281: 1163–1169.

American Cancer Society. *Breast Cancer Facts & Figures 2009-2010*. Atlanta: American Cancer Society, Inc. (Accessed March, 2011, at http://www.cancer.org/)

Anderson E. Progesterone receptors: (animal models and cell signaling in breast cancer) the role of estrogen and progesterone receptors in human mammary development and tumorigenesis. *Breast Cancer Research* 2002; 4: 197–201.

Archer TC, Jin J, Casey ES. Interactions of Sox1, Sox2, Sox3 and Oct4 during primary neurogenesis. *Developmental Biology* 2011; 350: 429–440.

Ariazi EA, Brailoiu E, Yerrum S, et al. The G protein-coupled receptor GPR30 inhibits proliferation of estrogen receptor-positive breast cancer cells. *Cancer Research* 2010; 70: 1184–1194.

Barkhem T, Haldosen LA, Gustafsson JA, et al. *pS2* Gene expression in HepG2 cells: complex regulation through crosstalk between the estrogen receptor α, an estrogen-responsive element, and the activator protein 1 response element. *Molecular Pharmacology* 2002; 61: 1273–1283.

Barone I, Brusco L, Fuqua SAW. Estrogen receptor mutations and changes in downstream gene expression and signaling. *Clinical Cancer Research* 2010; 16: 2702–2708.

Bernstein BE, Mikkelsen TS, Xie X, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 2006; 125: 315–326.

Bourdeau V, Deschênes J, Métivier R, et al. Genome-wide identification of high-affinity estrogen response elements in human and mouse. *Molecular Endocrinology* 2004; 18: 1411–1427.

Brauch H, Murdter TE, Eichelbaum M, et al. Pharmacogenomics of tamoxifen therapy. *Clinical Chemistry* 2009; 55: 1770–1782.

Bronson MW, Hillenmeyer S, Park RW, et al. Estrogen coordinates translation and transcription, revealing a role for NRSF in human breast cancer cells. *Molecular Endocrinology* 2010; 24: 1120–1135.

Brown AMC, Jeltsch JM, Roberts M, et al. Activation of *pS2* gene-transcription is a primary response to estrogen in the human breast cancer cell line MCF-7. *Proc Natl Acad Sci USA* 1984; 81: 6344–6348.

Bryne JC, Valen E, Tang MH, et al. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Research* 2008; 36: 102–106.

Brzozowski AM, Pike ACW, Dauter Z, et al. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* 1997; 389: 753–758.

Buckley MF, Sweeney KJ, Hamilton JA, et al. Expression and amplification of cyclin genes in human breast cancer. *Oncogene* 1993; 8: 2127–2133.

Cabili MN, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Gene & Development* 2011, doi: 10.1101/gad.17446611.

Calin GA, Liu CG, Ferracin M, et al. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* 2007; 12: 215–229.

Carroll JS, Brown M. Estrogen receptor target gene: an evolving concept. *Molecular Endocrinology* 2006; 20: 1707–1714.

Carroll JS, Liu XS, Brodsky AS, et al. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* 2005; 122: 33–43.

Carroll JS, Meyer CA, Song J, et al. Genome-wide analysis of estrogen receptor binding sites. *Nature Genetics* 2006; 38: 1289–1297.

Charn TH, Liu ET, Chang EC, et al. Genome-wide dynamics of chromatin binding of estrogen receptors alpha and beta: Mutual Restriction and Competitive Site Selection. *Molecular Endocrinology* 2010; 24: 47–59.

Cheung E, Kraus WL. Genomic analyses of hormone signaling and gene regulation. *Annual Review of Physiology* 2010; 72: 191–218.

Cheung E, Schwabish MA, Kraus WL. Chromatin exposes intrinsic differences in the transcriptional activities of estrogen receptors alpha and beta. *EMBO Journal* 2003; 22: 600–611.

Cicatiello L, Addeo R, Sasso A, et al. Estrogens and progesterone promote persistent CCND1 gene activation during G1 by inducing transcriptional derepression via c-Jun/c-Fos/estrogen receptor (progesterone receptor) complex assembly to a distal regulatory element and recruitment of cyclin D1 to its own gene promoter. *Molecular and Cellular Biology* 2004; 24: 7260–7274.

Cicatiello L, Mutarelli M, Grober OM, et al. Estrogen receptor alpha controls a gene network in luminal-like breast cancer cells comprising multiple transcription factors and microRNAs. *American Journal of Pathology* 2010; 176: 2113–2130.

Cirillo LA, Lin FR, Cuesta I, et al. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA1) and GATA-4. *Molecular Cell* 2002; 9: 279–289.

Cirillo LA, McPherson CE, Bossard P, et al. Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *The EMBO Journal* 1998; 17: 244–254.

Deblois G, Hall JA, Perry MC, et al. Genome-wide identification of direct target genes implicates estrogen-related receptor alpha as a determinant of breast cancer heterogeneity. *Cancer Research* 2009; 69: 6149–6157.

DeGraffenried LA, Hopp TA, Valente AJ, et al. Regulation of the estrogen receptor alpha minimal promoter by Sp1, USF-1 and ER alpha. *Breast Cancer Research and Treatment* 2004; 85: 111–120.

DeNardo DG, Kim HT, Hilsenbeck S, et al. Global gene expression analysis of estrogen receptor transcription factor cross talk in breast cancer: Identification of estrogen-induced activator protein-1-dependent genes. *Molecular Endocrinology* 2005; 19: 362–378.

Deroo BJ, Korach KS. Estrogen receptors and human disease. *Journal of Clinical Investigation* 2006; 116: 561–570.

Dong XY, Guo P, Sun XD, et al. Estrogen up-regulates ATBF1 transcription but causes its protein degradation in estrogen receptor-alpha-positive breast cancer cells. *Journal of Biological Chemistry* 2011; 286: 13879–13890.

Eeckhoute J, Keeton EK, Lupien M, et al. Positive cross-regulatory loop ties GATA-3 to estrogen receptor alpha expression in breast cancer. *Cancer Research* 2007; 67: 6477–6483.

Ehlers MRW, Riordan JF. Angiotensin-converting enzyme: New concepts concerning its biological role. *Biochemistry* 1989; 28: 5311–5318.

Ehlers MRW, Riordan JF. Angiotensin-converting enzyme: biochemistry and molecular biology, in "Hypertension: Pathophysiology, Diagnosis, and Management" (Laragh, J. H., and Brenner, B. M., Eds.). New York: Raven Press, 1990: 1217–1231.

Faddy HM, Robinson JA, Lee WJ, et al. Peroxisome proliferator-activated receptor alpha expression is regulated by estrogen receptor alpha and modulates the response of MCF-7 cells to sodium butyrate. *International Journal of Biochemistry and Cell Biology* 2006; 38: 255–266.

Farias EF, Petrie K, Leibovitch B, et al. Interference with Sin3 function induces epigenetic reprogramming and differentiation in breast cancer cells. *Proc Natl Acad Sci USA* 2010; 107: 11811–11816.

Fraga MF, Ballestar E, Villar-Garea A, et al. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nature Genetics* 2005; 37: 391–400.

Fullwood MJ, Liu MH, Pan YF, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 2009; 462: 58–64.

Gallagher PE, Li P, Lenhart JR, et al. Estrogen regulation of angiotensin-converting enzyme mRNA. *Hypertension* 1999; 33: 323–328.

Ganz PA, Habel LA, Weltzien EK, et al. Examining the influence of beta blockers and ACE inhibitors on the risk for breast cancer recurrence: results from the LACE cohort. *Epidemiology* 2011; 129: 549–556.

Gao H, Falt S, Sandelin A, et al. Genome-wide identification of estrogen receptor alpha-binding sites in mouse liver. *Molecular Endocrinology* 2008; 22: 10–22.

Gaub MP, Bellard M, Scheuer I, et al. Activation of the ovalbumin gene by the estrogen receptor involves the fos-jun complex. *Cell* 1990; 63: 1267–1276.

Gavras H. Angiotensin converting enzyme inhibition and its impact on cardiovascular disease. *Circulation* 1990; 81: 381–388.

Gery S, Tanosaki S, Bose S, et al. Down-regulation and growth inhibitory role of C/EBPα in breast cancer. *Clinical Cancer Research* 2005; 11: 3184-3190.

Glass CK, Rosenfeld MG. The coregulator exchange in transcriptional functions of nuclear receptors. *Genes Development* 2000; 14: 121–141.

Gojis O, Rudraraju B, Gudi M, et al. The role of SRC-3 in human breast cancer. *Nature Reviews Clinical Oncology* 2010; 7: 83–89.

Green KA, Carroll JS. Opinion: oestrogen-receptor-mediated transcription and the influence of co-factors and chromatin state. *Nature Reviews Cancer* 2007; 7: 713–722.

Green S, Walter P, Kumar V, et al. Human oestrogen receptor cDNA: sequence, expression and homology to v-erb-A. *Nature* 1986; 320: 134–139.

Gupta RA, Shah N, Wang KC, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 2010; 464: 1071–1076.

Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 2009; 458: 223–227.

Guttman M, Donaghey J, Carey BW, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 2011, doi: 10.1038/nature10398.

Hakim O, Sung MH, Hager GL. 3D shortcuts to gene regulation. *Current Opinion in Cell Biology* 2010; 22: 305–313.

Hao L, Rizzo P, Osipo C, et al. Notch-1 activates estrogen receptor-alpha-dependent transcription via IKK alpha in breast cancer cells. *Oncogene* 2010; 29: 201–213.

Heldring N, Pike A, Andersson S, et al. Estrogen receptors: How do they signal and what are their targets. *Physiological Reviews* 2007; 87: 905–931.

Holmes KA, Song JS, Liu XS, et al. Nkx3–1 and LEF-1 function as transcriptional inhibitors of estrogen receptor activity. *Cancer Research* 2008; 68: 7380–7385.

Hortobagyi GN. Progress in endocrine therapy for breast carcinoma. *Cancer* 1998; 83: 1–6.

Inoue A, Omoto Y, Yamaguchi Y, et al. Transcription factor EGR3 is involved in the estrogen-signaling pathway in breast cancer cells. *Journal of Molecular Endocrinology* 2004; 32: 649–666.

Hu M, Yu J, Taylor JMG, et al. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Research* 2010; 38: 2154–2167.

Ioannidis JP, Allison DB, Ball CA, et al. Repeatability of published microarray gene expression analyses. *Nature Genetics* 2009; 41: 149–155.

Irie T, Park SJ, Yamashita R, et al. Predicting promoter activities of primary human DNA sequences. *Nucleic Acids Research* 2011; 39: 11: e75

Jakowlew SB, Breathnach R, Jeltsch JM, et al. Sequence of the *pS2* messenger-RNA induced by estrogen in the human breast cancer cell line MCF-7. *Nucleic Acids Research* 1984; 12: 2861–2878.

Jensen EV, Jacobson HI. Basic guides to the mechanism of estrogen action. *Recent Progress in Hormone Research* 1962; 18: 387–414.

Joseph R, Orlov YL, Huss M, et al. Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. *Molecular Systems Biology* 2010; 6: 456.

Joshi SR, Ghattamaneni RB, Scovell WM. Expanding the paradigm for estrogen receptor binding and transcriptional activation. *Molecular Endocrinology* 2011; 25: 980–994.

Kamalakaran S, Radhakrishnan SK, Beck WT. Identification of estrogen-responsive genes using a genome-wide analysis of promoter elements for transcription factor binding sites. *Journal of Biological Chemistry* 2005; 280: 21491–21497.

Katzenellenbogen BS. Estrogen receptors: Bioactivities and interactions with cell signaling pathways. *Biology of Reproduction* 1996; 54: 287–293.

Katzenellenbogen BS, Choi I, Delage-Mourroux R, et al. Molecular mechanisms of estrogen action: selective ligands and receptor pharmacology. *Journal of Steroid Biochemistry and Molecular Biology* 2000; 74: 279–285.

Kim TH, Barrera LO, Zheng M, et al. A high-resolution map of active promoters in the human genome. *Nature* 2005; 436: 876–880.

Kleinhitpass L, Schorpp M, Wagner U, et al. An estrogen-responsive element derived from the 5' flanking region of the *xenopus* vitellogenin A2 gene functions in transfected human cells. *Cell* 1986; 46: 1053–1061.

Klinge CM. Estrogen receptor interaction with estrogen response elements. *Nucleic Acids Research* 2001; 29: 2905–2919.

Kocanova S, Kerr EA, Rafique S, et al. Activation of estrogen-responsive genes does not require their nuclear co-localization. *PLoS Genetics* 2010; 6: e10009.

Kwon YS, Garcia-Bassets I, Hutt KR, et al. Sensitive ChIP-DSL technology reveals an extensive estrogen receptor alpha-binding program on human gene promoters. *Proc Natl Acad Sci USA* 2007; 104: 4852–4857.

Ladd AMGZ, Vásquez AA, Sayed-Tabatabaei FA, et al. Angiotensin-converting enzyme gene insertion/deletion polymorphism and breast cancer risk. *Cancer Epidemiology, Biomarkers & Prevention* 2005; 14: 2143–2146.

Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 2009; 10: R25.

Laronda MM, Jameson JL. Sox3 functions in a cell-autonomous manner to regulate spermatogonial differentiation in mice. *Endocrinology* 2011; 152:1606–1615.

Lechner T, Carrozza MJ, Yu Y, et al. Sds3 (suppressor of defective silencing 3) is an integral component of the yeast Sin3-Rpd3 histone deacetylase complex and is required for histone deacetylase activity. *Journal of Biology Chemistry* 2000; 275: 40961–40966.

Lee TI, Jenner RG, Boyer LA, et al. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* 2006; 125: 301–313.

Le Romancer M, Treilleux I, Bouchekioua-Bouzaghou K, et al. Methylation, a key step for non-genomic estrogen signaling in breast tumors. *Steroids* 2010; 75: 560–564.

Lefrancois P, Zheng W, Snyder M. Guide to yeast genetics. *Methods in Enzymology* 2010; 470: 77–104.

Lin CY, Ström A, Vega VB, et al. Discovery of estrogen receptor α target genes and response elements in breast tumor cells. *Genome Biology* 2004; 5: R66.

Lin CY, Vega VB, Thomsen JS, et al. Whole genome cartography of estrogen receptor alpha binding sites. *PLoS Genetics* 2007; 3: 867–885.

Lin ZH, Reierstad S, Huang CC, et al. Novel estrogen receptor-alpha-binding sites and estradiol target genes identified by chromatin immunoprecipitation cloning in breast cancer. *Cancer Research* 2007; 67: 5017–5024.

Liu ZH, Chen SL. ER regulates an evolutionarily conserved apoptosis pathway. *Biochemical and Biophysical Research Communications* 2010; 400: 34–38.

Lo R, Burgoon L, MacPherson L, Ahmed S, et al. Estrogen receptor-dependent regulation of CYP2B6 in human breast cancer cells. *Biochimica et Biophysica Acta-Gene Regulatory Mechanisms* 2010; 1799: 469–479.

Lokuge S, Frey BN, Foster JA, et al. The rapid effects of estrogen: a mini-review. *Behavioral Pharmacology* 2010; 21: 465–472.

Lupien M, Eeckhoute J, Meyer CA, et al. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* 2008; 132: 958–970.

Lupien M, Meyer CA, Bailey ST, et al. Growth factor stimulation induces a distinct ER alpha cistrome underlying breast cancer endocrine resistance. *Genes Development* 2010; 24: 2219–2227.

Madaio RA, Spalletta G, Cravello L, et al. Overcoming endocrine resistance in breast cancer. *Current Cancer Drug Targets* 2010; 10: 519–528.

Madak-Erdogan Z, Kieser KJ, Kim SH, et al. Nuclear and extranuclear pathway inputs in the regulation of global gene expression by estrogen receptors. *Molecular Endocrinology* 2008; 22: 2116–2127.

Madak-Erdogan Z, Lupien M, Stossi F, et al. Genomic collaboration of estrogen receptor alpha and extracellular signal-regulated kinase 2 in regulating gene and proliferation programs. *Molecular and Cellular Biology* 2011; 31: 226–236.

Mandal S, Davie JR. Estrogen regulated expression of the p21 (Waf1/Cip1) gene in estrogen receptor positive human breast cancer cells. *Journal of Cellular Physiology* 2010; 224: 28–32.

Mangelsdorf DJ, Thummel C, Beato M, et al. The nuclear receptor superfamily: The second decade. *Cell* 1995; 83: 835–839.

Mason CE, Shu FJ, Wang C, et al. Location analysis for the estrogen receptor-alpha reveals binding to diverse ERE sequences and widespread binding within repetitive DNA elements. *Nucleic Acids Research* 2010; 38: 2355–2368.

Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics* 2006; 7: 29–59.

Mefford D, Mefford JA. Enumerating the gene sets in breast cancer, a "direct" alternative to hierarchical clustering. *BMC Genomics* 2010; 11: 482.

Moon SY, Zheng Y. Rho GTPase activating proteins in cell regulation. *Trends in Cell Biology* 2003; 13:13–22.

Morelli C, Lanzino M, Garofalo C, et al. Akt2 inhibition enables the forkhead transcription factor FoxO3a to have a repressive role in estrogen receptor alpha transcriptional activity in breast cancer cells. *Molecular and Cellular Biology* 2010; 30: 857–870.

Mosselman S, Polman J, Dijkema R. ER beta: identification and characterization of a novel human estrogen receptor. *FEBS Letters* 1996; 392: 49–53.

Nagashima T, Suzuki T, Kondo S, et al. Integrative genome-wide expression analysis bears evidence of estrogen receptor-independent transcription in heregulin-stimulated MCF-7 cells. *PLoS ONE* 2008; 3: 12: e1803.

Nilsson S, Ma S, Treuter E, et al. Mechanisms of estrogen action. *Physiology Review* 2001; 81: 1535–1565.

Noonan JP, McCallion. Genomics of long-range regulatory elements. *Annual Review of Genomics and Human Genetics* 2010; 11: 1–24.

O'Day E, Lai A. MicroRNAs and their target gene networks in breast cancer. *Breast Cancer Research* 2010; 12: 201.

O'Lone R, Frith MC, Karlsson EK, et al. Genomic targets of nuclear estrogen receptors. *Molecular Endocrinology* 2004; 18: 1859–1875.

Omalley BW, McGuire WL, Middleto P. Altered gene expression during differentiation-population changes in hybridizable RNA after stimulation of chick oviduct with estrogen. *Nature* 1968; 218: 1249–1251.

Paech K, Webb P, Kuiper GGJM, et al. Differential ligand activation of estrogen receptors ERα and ERβ at AP1 sites. *Science* 1997; 277: 1508–1510.

Pike JW. Genome-scale techniques highlight the epigenome and redefine fundamental principles of gene regulation. *Journal of Bone and Mineral Research* 2010; 26: 1155–1162.

Planas-Silva MD, Shang YF, Donaher JL, et al. AIB1 enhances estrogen-dependent induction of cyclin D1 expression. *Cancer Research* 2001; 61: 3858–3862.

Prest SJ, May FEB, Westley BR. The estrogen-regulated protein, TFF1, stimulates migration of human breast cancer cells. *The FASEB Journal* 2002; 16: 592–594.

Rampalli S, Pavithra L., Bhatt A, et al. Tumor Suppressor SMAR1 Mediates Cyclin D1 Repression by Recruitment of the SIN3/Histone Deacetylase 1 Complex. *Molecular and Cellular Biology* 2005; 25: 8415–8429.

Renoir JM, Radanyi C, Faber LE, et al. The non-DNA-binding hetero-oligomeric form of mammalian steroid hormone receptors contains an hsp90-bound 59-kilodalton protein. *Journal of Biological Chemistry* 1990; 265: 10740–10745.

Roh TY, Cuddapah S, Cui K, et al. The genomic landscape of histone modifications in human T cells. *Proc Natl Acad Sci USA* 2006; 103: 15782–15787.

Roh TY, Wei G, Farrell CM, et al. Genome-wide prediction of conserved and non-conserved enhancers by histone acetylation patterns. *Genome Research* 2007; 17: 74–81.

Russo J, Russo IH. Development of the human breast. *Maturitas* 2004; 49: 2–15.

Sakabe NJ, Nobrega MA. Genome-wide maps of transcription regulatory elements. *Wiley Interdisciplinary Reviews-Systems Biology and Medicine* 2010; 2: 422–437.

Sanchez R, Nguyen D, Rocha W, et al. Diversity in the mechanisms of gene regulation by estrogen receptors. *Bioessays* 2002; 24: 244–254.

Schmidt D, Schwalie P, Ross-Innes CS, et al. A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Research* 2010; 20: 578–588.

Schnerder BP and Miller KD. Angiogenesis of breast cancer. *Journal of Clinical Oncology* 2005; 23: 1178–1190.

Schunkert H, Danser AHJ, Hense HW, et al. Effects of estrogen replacement therapy on the renin-angiotensin system in postmenopausal women. *Circulation* 1997; 95: 39–45.

Seligson DB, Horvath S, Shi T, et al. Global histone modification patterns predict risk of prostate cancer recurrence. *Nature* 2005; 435: 1262–1266.

Shang YF, Hu X, DiRenzo J, et al. Cofactor dynamics and sufficiency in estrogen receptor-regulated transcription. *Cell* 2000; 103: 843–852.

Shibata H, Spencer TE, Onate SA, et al. Role of co-activators and co-repressors in the mechanism of steroid/thyroid receptor action. *Recent Progress in Hormone Research* 1997; 52: 141–156.

Shu FJ, Sidell N, Yang DZ, et al. The tri-nucleotide spacer sequence between estrogen response element half-sites is conserved and modulates ER alpha-mediated transcriptional responses. *Journal of Steroid Biochemistry and Molecular Biology* 2010; 120: 172–179.

Silverstein RA, Ekwall K. Sin3: a flexible regulator of global gene expression and genome stability. *Current Genetics* 2005; 47:1–17.

Simpson NE, Lambert WM, Watkins R, et al. High levels of Hsp90 co-chaperone p23 promote tumor progression and poor prognosis in breast cancer by increasing lymph node metastases and drug resistance. *Cancer Research* 2010; 70: 8446–8456.

Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* 2003; 100: 8418–8423.

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 2001; 98: 10869–10874.

Soule HD, Vazquez J, Long A, et al. A human cell line from a pleural effusion derived from a breast carcinoma. *The Journal of the National Cancer Institute* 1973; 51: 1409–1416.

Spink BC, Bennett JA, Pentecost BT, et al. Long-term estrogen exposure promotes carcinogen bioactivation, induces persistent changes in gene expression, and enhances the tumorigenicity of MCF-7 human breast cancer cells. *Toxicology and Applied Pharmacology* 2009; 240: 355–366.

Stender JD, Kim K, Charn TH, et al. Genome-wide analysis of estrogen receptor alpha DNA binding and tethering mechanisms identifies Runx1 as a novel tethering factor in receptor-mediated transcriptional activation. *Molecular and Cellular Biology* 2010; 30: 3943–3955.

Stossi F, Madak-Erdogan Z, Katzenellenbogen BS. Estrogen receptor alpha represses transcription of early target genes via p300 and CtBP1. *Molecular and Cellular Bi*ology 2009; 29: 1749–1759.

Sun G, Porter W, Safe S. Estrogen-induced retinoic acid receptor α1 gene expression: role of estrogen receptor-Sp1 complex. *Molecular Endocrinology* 1998; 12: 882–890.

Sun J, Nawaz Z, Slingerland JM. Long-range activation of GREB1 by estrogen receptor via three distal consensus estrogen responsive elements in breast cancer cells. *Molecular Endocrinology* 2007; 21: 2651–2662.

Sutherland RL, Musgrove EA. Cyclin D1 and mammary carcinoma: new insights from transgenic mouse models. *Breast Cancer Research* 2002; 4: 14–17.

Tan SK, Lin ZH, Chang CW, et al. AP-2γ regulates estrogen receptor-mediated long-range chromatin interaction and gene transcription *The EMBO Journal* 2011; 30: 2569 –2581.

Taslim C, Wu J, Yan P, et al. Comparative study on ChIP-Seq data: normalization and binding pattern characterization. *Bioinformatics* 2009; 25: 2334–2340.

Theodorou V, Carroll JS. Estrogen receptor action in three dimensions: looping the loop. *Breast Cancer Research* 2010; 12: 2470.

Thomas C, Gustafsson J. The different roles of ER subtypes in cancer biology and therapy. *Nature Reviews Cancer* 2011; 11: 597–608.

Tsai MC, Spitale RC, Chang HY. Long intergenic noncoding RNAs: new links in cancer progression. *Cancer Resarch* 2011; 71: 3–7.

Tsuchiya Y, Nakajima M, Yokoi T. Cytochrome P450-mediated metabolism of estrogens and its regulation in human. *Cancer Letters* 2005; 227: 115–124.

Tuteja G, White P, Schug J, et al. Extracting transcription factor targets from ChIP-Seq data. *Nucleic Acids Research* 2009; 37: 10: e11310.

Van de Vijver MJ, He YD, van't Veer LJ, et al. A gene expression signature as a predictor of survival in breast cancer. *NEJM* 2002; 347: 1999–2009.

van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; 415: 530–536.

Waldminghaus T, Skarstad K. ChIP on Chip: surprising results are often artifacts. *BMC Genomics* 2010; 11: 414.

Wang C, Yu J, Kallen CB. Two estrogen response element sequences near the PCNA gene are not responsible for its estrogen-enhanced expression in MCF7 Cells. *PLoS ONE* 2008; 10: e352310.

Webb P, Nguyen P, Valentine C, et al. Estrogen receptor enhances AP-1 activity by two distinct mechanisms with different requirements for receptor transactivation functions. *Molecular Endocrinology* 1999; 13: 1672–1685.

Wei G, Wei L, Zhu J, et al. Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4[+] T cells. *Immunity* 2009; 30: 155–167.

Welboren WJ, Sweep F, Span PN, et al. Genomic actions of estrogen receptor alpha: what are the targets and how are they regulated? *Endocrine-Related Cancer* 2009; 16: 1073–1089.

Yang J, Jubb AM, Pike L, et al. The histone demethylase JMJD2B is regulated by estrogen receptor alpha and hypoxia, and is a key mediator of estrogen induced growth. *Cancer Research* 2010; 70: 6456–6466.

Zeisig BB, Kwok C, Zelent A, et al. Recruitment of RXR by homotetrameric RARalpha fusion proteins is essential for transformation. *Cancer Cell* 2007; 12: 36–51.

Zhang Y, Liang J, Li Y, et al. (2010). CCCTC-binding factor acts upstream of FOXA1 and demarcates the genomic response to estrogen. *Journal of Biological Chemistry* 2010: 285: 28604–28613.

Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biology* 2008; 9: R137.

Zhao C, Dahlman-Wright K, Gustafsson JA. Estrogen signaling via estrogen receptor beta. *Journal of Biological Chemistry* 2010; 285: 39575–39579.

Zhao C, Gao H, Liu Y, et al. Genome-wide mapping of estrogen receptor-beta-binding regions reveals extensive cross-talk with transcription factor activator protein-1. *Cancer Research* 2010; 70: 5174–5183.

Zhao R, Daley GQ. From fibroblasts to iPS cells: Induced pluripotency by defined factors. *Journal of Cellular Biochemistry* 2008; 105: 949–955.

Zorn AM, Barish GD, Williams BO, et al. Regulation of Wnt signaling by Sox proteins: XSox17α/β and XSox3 physically interact with β-catenin. *Molecular Cell* 1999; 4: 487–98.

8. German Summary (Zusammenfassung)

Das Steroidhormon Östrogen spielt eine zentrale Rolle bei der Genregulation in hormonresponsiblen Brustkrebszellen, und zwar vermittels Estrogenrezeptoren (ERs), von denen ERα eines der wichtigsten Isoforme darstellt. Nach der Aktivierung durch E2 bindet ERα in kurzer Zeit an Tausenden von genomischen Positionen, rekrutiert Coregulatoren, Chromatinmodifikatoren und den grundlegenden Transkriptionsapparat und beeinflusst letztlich auf diese Weise die Expression von Zielgene.

Um die Mechanismen aufzudecken, durch die ERα die Genexpression verändert, wurden genomweit sowohl die ERα Bindestellen als auch die durch E2-Stimulation verursachten Änderungen der Genexpression untersucht. Aus unseren ChIP-Seq Ergebnissen, ergänzt durch bereits zuvor veröffentlichte Daten, konnten 4543 zuverlässige ERα Bindestellen für den Zeitpunkt 45 Minuten nach der Behandlung mit E2 bestimmt werden, zum großen Teil außerhalb von Promotorregionen gelegen. Interessanterweise konnten wir entgegen den meisten bisheriger Studien eine erhebliche Anzahl von ERα Bindungsvorkommen in Abwesenheit von E2 beobachten. Zur Bestimmung der Transkriptionsänderungen durch Stimulation mit E2 kamen moderne RNA-Seq-Verfahren zur Anwendung. Um zwischen einer Regulation der Transkription und post-transkriptionalen Vorgängen unterscheiden zu können, wurde die Genexpression sowohl für die gesamte RNA als auch für neu-synthetisierte RNA zu verschiedenen Zeitpunkten nach der Behandlung mit E2 gemessen. Eine differentielle Regulation wurde bei insgesamt 722 Gene festgestellt, welche nach dem Muster ihrer Response-Dynamik in sechs Gruppen klassifiziert werden konnten. Die weitere Untersuchung der Beziehung zwischen ERα-Bindungstellen und differentiell exprimierten Genen zeigte, dass hochgeregelte Gene zu einem signifikant höheren Anteil ERα-Bindestellen in Promotornähe aufweisen als heruntergeregelte Gene. Unter den hochgeregelten Genen besaßen diejenigen, die ihre Expression bereits zu einem frühen Zeitpunkt geändert hatten, die höchste Anreicherung an ERα-Bindestellen in promotornahen oder -fernen Regionen. Insgesamt scheint die unmittelbare Auswirkung auf die Transkription durch ERα größtenteils in schneller Transkriptionsinduktion zu bestehen.

9. Publications

Sun W, Hu Y, Gogol-doering A, Chen W. Genome-wide analysis of the transcriptional regulatory networks mediated by estrogen receptor alpha in MCF-7 human breast cancer cells. 2012 (**Manuscript in preparation**)

Former publications (on antiviral research)

1. Yang J, **Sun W**, Liu S. Research progress on HIV entry inhibitors from natural source. *Chinese Traditional and Herbal Drugs,* 2009 (11).

2. **Sun W**, Wang H, Xia C, Wu S, Jiang S, Jiang Z, Liu, S. 1,2,6-Tri-O-galloyl-β-D-glucopy-ranose inhibits HIV induced cell fusion by targeting HIV gp41. *Journal of Southern Medical University*, 2008 (07).

3. Yang J, **Sun W**, Yang X, Chen W, Lv L, Duan W. Screening of HIV-1 fusion inhibitors from extracts of Chinese medicinal herb Fructus Canarii. *China Pharmacy,* 2008 (21).

10. Erklärung

Hiermit versichere ich, dass ich die Dissertation selbstständig verfasst habe und dass ich keine anderen als die in der Arbeit genannten Hilfsmitteln und Quellen benutzt habe.

Datum: Berlin, March, 2012   Unterschrift: Wei Sun

11. Appendix

Table 5. The E2-regulated genes identified by both the newly synthesized and bulk RNA-Seq. [#] In cluster, +, - and 0 represent upregulation, down regulation and no significant change, respectively. Each cluster, from left to right, represents gene changes at 45min, 2.5hr and 5hr by newly synthesized RNA-Seq, and 45min, 2.5hr and 5hr by bulk RNA-Seq.* Distance designates the distance of E2-regulated gene TSSs to ERα binding sites.

Chr	Gene Symbol	Cluster[#]	Group	Distance*
chr1	RGS16	C+00+00	init up	42059
chr1	KLHL21	C+00+00	init up	8981
chr1	PHF13	C+00+00	init up	976
chr1	ELF3	C+00+00	init up	282
chr1	ITLN2	C+00+00	init up	138156
chr1	EDN2	C+00+00	init up	24179
chr1	WNT9A	C+00+00	init up	1148353
chr1	FAM46C	C+00000	init up	423417
chr1	MMACHC, PRDX1	C+00000	init up	4577
chr1	ARL8A	C+00000	init up	11838
chr1	CITED4	C000+00	init up	57239
chr1	PPFIA4	C++++++	early up	8344
chr1	IL20	C++++++	early up	558
chr1	FAIM3,IL24	C++++++	early up	142
chr1	HPDL	C++++++	early up	22044
chr1	FABP3	C++++++	early up	4048
chr1	NBPF1	C++++++	early up	1071
chr1	ADORA1	C+++0++	early up	25684
chr1	MANEAL	C+++0++	early up	640
chr1	ANXA9	C+++0++	early up	68
chr1	ETNK2	C+++0++	early up	856
chr1	FLJ32224, SLC2A1	C+++0++	early up	14324
chr1	EFHD2	C+++0++	early up	4538
chr1	PPM1J	C+++0++	early up	673505
chr1	FAM63A	C+++0++	early up	160
chr1	C1orf15-NBL1, C1orf151, NBL1	C+++0++	early up	140
chr1	PGBD5	C0++0++	late up	122996
chr1	PDZK1	C0++0++	late up	953
chr1	NR5A2	C0++0++	late up	10709
chr1	C1orf126, KAZN	C0++0++	late up	2063
chr1	C1orf226	C0++0++	late up	210359
chr1	MIR556, NOS1AP	C0++0++	late up	82106
chr1	NKAIN1	C0++0++	late up	18989
chr1	CELSR2	C0++0++	late up	3371
chr1	ZBTB40	C0++0++	late up	20510
chr1	ATP13A2	C0++00+	late up	17741

Chr	Gene Symbol	Cluster#	Group	Distance*
chr1	PPP1R12B	C0++00+	late up	7789
chr1	UBE2T	C0++00+	late up	6262
chr1	BAI2, MIR4254	C0++0++	late up	13370
chr1	SLC25A24	C0++0++	late up	700
chr1	C1orf130	C0++0++	late up	11966
chr1	DYRK3	C0++0++	late up	55961
chr1	PADI2	C0++00+	late up	62954
chr1	JAK1, RAVER2	C0++00+	late up	16804
chr1	C1orf163	C00+0++	late up	713605
chr1	SLC19A2	C00+00+	late up	43188
chr1	SLC30A1	C0++0++	late up	56825
chr1	GRIK3	C00+00+	late up	4545
chr1	DPH2	C00+00+	late up	97428
chr1	HIVEP3	C00+00+	late up	111
chr1	SPAG17, WDR3	C00+00+	late up	747894
chr1	TXNIP	C-00-00	init down	287778
chr1	C1orf133, SERTAD4	C-00-00	init down	686272
chr1	JUN	C-00-00	init down	2431732
chr1	MCL1	C-00-00	init down	12144
chr1	C1orf51	C-00000	init down	204540
chr1	ZC3H12A	C------	early down	138631
chr1	ID3	C------	early down	441599
chr1	EFNA1	C---0--	early down	36312
chr1	MIR554, TUFT1	C0-0--	late down	26026
chr1	PIK3R3	C0-0--	late down	615720
chr1	IGSF3	C0--0--	late down	514163
chr1	ARHGEF2	C0--0--	late down	237978
chr1	PVRL4	C0--0--	late down	29410
chr1	LOC388692	C0--00-	late down	69810
chr1	SELENBP1	C0--00-	late down	140909
chr1	C1orf115	C0--00-	late down	2458433
chr1	IRF6	C0--00-	late down	259415
chr1	FLJ39609	C0--00-	late down	154396
chr1	RHBG	C0--00-	late down	86302
chr1	SLC9A1	C0--00-	late down	207815
chr1	TRIM45	C00-0--	late down	60066
chr1	RFX5	C00-00-	late down	166985
chr1	CNKSR1	C00-00-	late down	420188
chr1	PLEKHA6	C00-00-	late down	20764
chr1	RAP1GAP	C00-00-	late down	12199
chr1	EFNA4	Crest	rest	27489
chr2	DDX11L2, WASH2P	C+00+00	init up	406755
chr2	INHBB	C+00+00	init up	5693
chr2	SDC1	C+00+00	init up	31865
chr2	CYP1B1	C++++++	early up	61
chr2	FZD7	C++++++	early up	48695
chr2	KCNF1	C++++++	early up	134115
chr2	FHL2	C+++0++	early up	7114
chr2	MYEOV2	C+++0++	early up	65360

Chr	Gene Symbol	Cluster#	Group	Distance*
chr2	C2orf18	C+++0++	early up	6111
chr2	TUBA3D	C0+++++	early up	2945
chr2	SLC1A4	C0++0++	late up	12931
chr2	GPR39, LYPD1	C0++0++	late up	70672
chr2	TGFA	C0++0++	late up	62467
chr2	GREB1	C0++0++	late up	264
chr2	HAAO	C0++0++	late up	616
chr2	CYP26B1	C0++0++	late up	421274
chr2	TNS1	C0++0++	late up	89
chr2	MARS2	C0++0++	late up	137473
chr2	SOS1	C0++0++	late up	56325
chr2	REEP1	C0++00+	late up	71624
chr2	LONRF2	C0++00+	late up	254525
chr2	MREG	C0++00+	late up	30343
chr2	SPEG	C0++00+	late up	12804
chr2	RAMP1	C00+0++	late up	81455
chr2	DUSP2	C00+00+	late up	184410
chr2	PRKCE	C00+00+	late up	100329
chr2	IRS1	C00+00+	late up	647414
chr2	SLC1A4	C0++0++	late up	12931
chr2	FZD5	C-00000	init down	417793
chr2	ID2	C------	early down	2176
chr2	WNT6	C---0--	early down	48993
chr2	FAM110C	C---0--	early down	343021
chr2	RPRM	C0--0--	late down	924755
chr2	REL	C0--0--	late down	1189539
chr2	GAL3ST2	C0--0--	late down	816262
chr2	SEMA4C	C0--0--	late down	70775
chr2	EFHD1	C0--00-	late down	254326
chr2	NEU4	C0--00-	late down	850182
chr2	MGAT4A	C0--00-	late down	429353
chr2	EPB41L5	C0--00-	late down	253319
chr2	NCRNA00152	C0--00-	late down	3
chr2	EFEMP1	C0--00-	late down	305782
chr2	TFPI	C00-00-	late down	1731164
chr2	MYO1B	C00-00-	late down	458330
chr2	PSD4	C00-00-	late down	2915
chr2	HS6ST1	Crest	rest	2522
chr3	FAM43A	C+00+00	init up	37355
chr3	BCL6, LOC100131635, RTP2	C+00+00	init up	183130
chr3	PPM1M	C+00000	init up	186715
chr3	TWF2	C+00000	init up	193514
chr3	SEMA3G	C++++++	early up	12104
chr3	SIAH2	C++++++	early up	4530
chr3	HES1	C++++++	early up	18110
chr3	KCTD6	C++++++	early up	24721
chr3	EPHB3	C++++++	early up	32518
chr3	LOC100287227, TIPARP	C++++++	early up	4

Chr	Gene Symbol	Cluster[#]	Group	Distance*
chr3	MAP6D1	C++++++	early up	20260
chr3	ALS2CL	C++++++	early up	10605
chr3	IL17RB	C+++0++	early up	158
chr3	LRIG1, SLC25A26	C0++0++	late up	24665
chr3	ATRIP	C0++0++	late up	4575
chr3	SLCO2A1	C0++0++	late up	10110
chr3	PXK	C0++0++	late up	134485
chr3	HDAC11	C0++0++	late up	301
chr3	SH3BP5	C0++0++	late up	39139
chr3	GRIP2, SLC6A6	C0++0++	late up	3053
chr3	SKIL	C0++0++	late up	7716
chr3	SEMA3B	C0++00+	late up	1099
chr3	FLNB	C0++00+	late up	31577
chr3	RRP9	C0++00+	late up	491017
chr3	SLC25A36	C0++00+	late up	201
chr3	PDHB	C0++00+	late up	33522
chr3	ATP13A3	C0++00+	late up	26937
chr3	PLXND1	C00+0++	late up	38447
chr3	TRIM59	C-00-00	init down	247597
chr3	CSRNP1	C0--0--	late down	1242994
chr3	TM4SF1	C0--00-	late down	130091
chr3	IL17RE	C0--00-	late down	405322
chr3	PRRT3	C0--00-	late down	368764
chr3	LOC344595	C0--00-	late down	559729
chr3	SIDT1	C0--00-	late down	4570939
chr3	CHRD	C0--00-	late down	34536
chr3	IL17RC	C0--00-	late down	387848
chr3	ABCC5	C00-0--	late down	171928
chr3	CMTM8	C00-00-	late down	128220
chr3	ABTB1	C00-00-	late down	94169
chr3	OSBPL10, ZNF860	C00-00-	late down	128608
chr4	FGFR3	C+00+00	init up	18753
chr4	WFS1	C++++++	early up	8526
chr4	FGFBP2	C+++0++	early up	121
chr4	PDLIM3	C+++0++	early up	21352
chr4	NPY1R	C0++0++	late up	52950
chr4	DOK7	C0++0++	late up	61
chr4	PPM1K	C0++0++	late up	50383
chr4	PPP2R2C	C0++00+	late up	2251
chr4	TSPAN5	C0++00+	late up	3266
chr4	NPY5R	C0++00+	late up	41242
chr4	LOC348926	C00+0++	late up	25
chr4	MAML3, MGST2	C00+00+	late up	11073
chr4	NAF1	C00+00+	late up	218824
chr4	CCNG2	C------	early down	389003
chr4	ANKRD56	C---0--	early down	129649
chr4	KIAA0922	C0--0--	late down	219548
chr4	TMEM150C	C0--00-	late down	472964

Chr	Gene Symbol	Cluster[#]	Group	Distance*
chr4	MXD4	C0--00-	late down	156698
chr4	NPNT	C00-00-	late down	166201
chr4	LETM1	Crest	rest	2409
chr5	N4BP3	C+00+00	init up	1691
chr5	LOC553103, MIR3936, SLC22A4 SLC22A5	C++++++	early up	802
chr5	CXXC5	C++++++	early up	4366
chr5	PITX1	C++++++	early up	12635
chr5	MSX2	C++++++	early up	237849
chr5	SNX18	C+++0++	early up	138563
chr4	LETM1	Crest	rest	2409
chr5	WWC1	C0++0++	late up	4890
chr5	STC2	C0++0++	late up	123893
chr5	SLC26A2	C0++0++	late up	56012
chr5	SNX24	C0++0++	late up	3636
chr5	FAM134B	C0++0++	late up	9818
chr5	P4HA2	C0++0++	late up	26663
chr5	ZSWIM6	C0++0++	late up	814
chr5	PPARGC1B	C0++0++	late up	38594
chr5	GEMIN5	C0++00+	late up	107881
chr5	NOP16	C00+0++	late up	484
chr5	RHOBTB3	C00+00+	late up	2350
chr5	RGNEF	C00+00+	late up	3945
chr5	DUSP1	C-00-00	init down	173407
chr5	ARRDC3, LOC100129716	C-00-00	init down	555608
chr5	IRX2	C-00000	init down	51055
chr5	FBN2	C---0--	early down	560722
chr5	PLK2	C0-0--	late down	364381
chr5	ARHGEF37	C0-0--	late down	98798
chr5	SLC12A2	C0--00-	late down	1014175
chr5	FLJ38109, GALNT10	C00-00-	late down	64642
chr5	ADAMTS19	C00-00-	late down	217879
chr5	CMYA5	C00-00-	late down	495924
chr6	LSM2	C+00000	init up	3642
chr6	HLA-DRB6	C++++++	early up	162
chr6	HLA-DRB1	C++++++	early up	17097
chr6	KCNK5	C++++++	early up	37491
chr6	MICB	C++++++	early up	11397
chr6	IER3	C++++++	early up	7197
chr6	HSPA1A	C++++++	early up	4905
chr6	HLA-DRB5	C+++0++	early up	24208
chr6	RBM24	C+++0++	early up	106761
chr6	GRM4	C0+++++	early up	5158
chr6	ELOVL2	C0++0++	late up	2257
chr6	TPD52L1	C0++0++	late up	555
chr6	PTP4A1	C0++0++	late up	80793
chr6	TPBG	C0++0++	late up	412
chr6	MBOAT1	C0++0++	late up	1563

Chr	Gene Symbol	Cluster[#]	Group	Distance*
chr6	C6orf141	C0++00+	late up	2751309
chr6	PKIB	C0++00+	late up	251
chr6	VEGFA	C0++00+	late up	34599
chr6	MAN1A1	C0++00+	late up	45
chr6	MIR1913, RPS6KA2	C00+00+	late up	88338
chr6	SLC29A1	C00+00+	late up	56
chr6	COL12A1	C00+00+	late up	236177
chr6	HIST1H2BC	C-00000	init down	164390
chr6	HIST1H3J	C000-00	init down	190146
chr6	PNRC1	C------	early down	427837
chr6	BAK1	C---0--	early down	182015
chr6	FAM83B	C0--0--	late down	17757
chr6	C6orf211	C0--0--	late down	181481
chr6	C6orf155	C0--0--	late down	1200902
chr6	RMND1	C0--0--	late down	181586
chr6	NEDD9	C0--0--	late down	164943
chr6	CDKN1A	C0--0--	late down	59400
chr6	SPDEF	C0--00-	late down	4021
chr6	PACSIN1	C0--00-	late down	36205
chr6	RNF144B	C00-0--	late down	69820
chr6	TRERF1	C00-00-	late down	18413
chr6	SLC17A5	C00-00-	late down	624731
chr6	SASH1	C00-00-	late down	290830
chr6	TFEB	C00-00-	late down	339273
chr6	ESR1	C00-00-	late down	32229
chr7	PDGFA	C+00+00	init up	101
chr7	AZGP1P1	C+00+00	init up	452230
chr7	TNRC18	C000+00	init up	3
chr7	AMZ1	C++++++	early up	582
chr7	EN2	C++++++	early up	18935
chr7	LFNG	C++++++	early up	959
chr7	PODXL	C+++0++	early up	69305
chr7	LOC401321	C+++0++	early up	25251
chr7	LOC100134229	C0++0++	late up	132960
chr7	FOXK1	C0++0++	late up	58
chr7	DPY19L2P3, LOC100271874, LOC646762	C0++0++	late up	228916
chr7	ZNF107	C0++0++	late up	7611
chr7	UFSP1	C0++0++	late up	438924
chr7	JHDM1D	C0++0++	late up	133567
chr7	BMPER	C0++00+	late up	58741
chr7	PLXNA4	C0++00+	late up	1758
chr7	LOC441208	C00+00+	late up	4702
chr7	INSIG1	C-00000	init down	6888
chr7	SEPT7P2	C000-00	init down	107564
chr7	C7orf68	C------	early down	11656
chr7	ZNF467	C0--0--	late down	230207
chr7	GLCCI1	C0--0--	late down	183450
chr7	PRR15	C0--00-	late down	107588

Chr	Gene Symbol	Cluster[#]	Group	Distance*
chr7	FAM20C	C0--00-	late down	31
chr7	LRCH4	C00-00-	late down	135075
chr7	GTF2IRD1	C00-00-	late down	41035
chr7	LAMB1	C00-00-	late down	104
chr7	KIAA1147	C00-00-	late down	630317
chr7	ARL4A	Crest	rest	91445
chr8	ARC	C+00+00	init up	25715
chr8	KLF10	C+00+00	init up	16078
chr8	PPP1R3B	C+00+00	init up	30577
chr8	EGR3	C++++++	early up	41252
chr8	ZNF703	C++++++	early up	484
chr8	PDLIM2	C++++++	early up	4509
chr8	TSPYL5	C++++++	early up	11408
chr8	NEIL2	C+++0++	early up	19967
chr8	GEM	C+++0++	early up	17918
chr8	MYC	C0+++++	early up	61710
chr8	SGK223	C0++0++	late up	66722
chr8	MYBL1	C0++0++	late up	90034
chr8	SLC7A2	C0++0++	late up	10726
chr8	RRS1	C0++0++	late up	83733
chr8	SDC2	C0++0++	late up	133716
chr8	GRHL2	C0++0++	late up	11799
chr8	SYBU	C0++0++	late up	67692
chr8	TMEM64	C0++00+	late up	241724
chr8	FLJ10661	C0++00+	late up	8501
chr8	DEPTOR	C00+00+	late up	199372
chr8	TRIB1	C-00-00	init down	938839
chr8	OSGIN2	C-00000	init down	81275
chr8	EIF3E	C000-00	init down	1291397
chr8	PLEKHF2	C---0--	early down	74063
chr8	ZBTB10	C---0--	early down	5812683
chr8	TP53INP1	C0--0--	late down	258643
chr8	TACC1	C0--0--	late down	225266
chr8	MAFA	C0--0--	late down	30966
chr8	CLU	C0--00-	late down	56823
chr8	PLEKHA2	C00-00-	late down	398315
chr8	PTK2B	C00-00-	late down	346070
chr8	NDRG1	C00-00-	late down	3591
chr9	GADD45G	C+00+00	init up	43105
chr9	KLF4	C+00+00	init up	33958
chr9	TPRN	C+00+00	init up	24590
chr9	NACC2	C000+00	init up	15122
chr9	PTGES	C++++++	early up	2782
chr9	NXNL2	C++++++	early up	4005
chr9	SLC25A25	C++++++	early up	9962
chr9	IER5L	C++++++	early up	9159
chr9	BAG1	C++++++	early up	29580
chr9	AQP3	C++++++	early up	2085

Chr	Gene Symbol	Cluster[#]	Group	Distance*
chr9	B4GALT1	C+++0++	early up	32648
chr9	FAM102A	C+++0++	early up	5165
chr9	KIF12	C+++0++	early up	13845
chr9	C9orf95	C+++0++	early up	13047
chr9	CLIC3	C+++0++	early up	4110
chr9	OLFM1	C+++0++	early up	51279
chr9	DPM2	C+++0++	early up	9149
chr9	UGCG	C+++0++	early up	9446
chr9	OSTF1	C0++0++	late up	12783
chr9	TMEM2	C0++0++	late up	3065
chr9	SUSD3	C0++0++	late up	5261
chr9	JAK2	C0++0++	late up	54550
chr9	MIR101-2, RCL1	C0++0++	late up	2506
chr9	LMX1B	C0++0++	late up	28570
chr9	NOL6	C0++00+	late up	28213
chr9	AK3	C0++00+	late up	14215
chr9	SIGMAR1	C00+00+	late up	3145
chr9	FGD3	C00+00+	late up	53
chr9	COL27A1, MIR455	C00+00+	late up	70047
chr9	PSAT1	C00+00+	late up	700109
chr9	C9orf47, S1PR3	C0--0--	late down	269034
chr9	CORO2A	C00-00-	late down	1467672
chr10	MAP3K8	C+00+00	init up	59750
chr10	BAMBI	C+00+00	init up	16979
chr10	NCOA4	C+00000	init up	3066
chr10	MIR3155, PFKFB3	C++++++	early up	1266
chr10	RET	C+++0++	early up	5097
chr10	CXCL12	C+++0++	early up	125240
chr10	ASB13	C+++0++	early up	6150
chr10	MSMB	C+++0++	early up	1990
chr10	CASP7	C+++0++	early up	301
chr10	PPIF	C+++0++	early up	10778
chr10	RHOBTB1	C0++0++	late up	400
chr10	SFXN2	C0++0++	late up	3675
chr10	DPYSL4	C0++0++	late up	8360
chr10	ZMIZ1	C0++0++	late up	928
chr10	C10orf2	C0++0++	late up	104088
chr10	C10orf137	C0++0++	late up	17817
chr10	LOC100128098, ST8SIA6	C0++00+	late up	8058
chr10	MYOF	C0++00+	late up	47058
chr10	INPP5F	C0++00+	late up	36870
chr10	TDRD1	C0++00+	late up	60075
chr10	PPRC1	C0++00+	late up	301246
chr10	MIR604, MIR938, SVIL	C00+00+	late up	4429
chr10	RRP12	C00+00+	late up	170446
chr10	KLF6	C-00-00	init down	36618
chr10	FAM160B1	C000-00	init down	314202

Chr	Gene Symbol	Cluster[#]	Group	Distance*
chr10	GATA3	C---0--	early down	72692
chr10	DUSP5	C0--0--	late down	291670
chr10	BLNK	C0--00-	late down	1138408
chr10	FAM21B	C0--00-	late down	275844
chr10	KIAA1217, MIR603, PRINS	C00-00-	late down	546354
chr10	NEURL	C00-00-	late down	16632
chr10	DKK1	Crest	rest	2511999
chr10	CALHM2	Crest	rest	90911
chr11	FUT4	C+00+00	init up	301149
chr11	KCTD21	C+00000	init up	31978
chr11	TSKU	C++++++	early up	5024
chr11	P2RY2	C++++++	early up	338
chr11	ASCL2	C++++++	early up	63389
chr11	SLC25A45	C++++++	early up	2250
chr11	EHD1	C++++++	early up	195304
chr11	ALDH3B1	C+++0++	early up	8193
chr11	CTSD,DUSP8, FAM99A, FAM99B, KRTAP LOC338651, LOC402778, MOB2	C+++0++	early up	606
chr11	PRSS23	C+++0++	early up	23921
chr11	RHOD	C+++0++	early up	5330
chr11	ZNF259	C+++0++	early up	21010
chr11	FRMD8	C+++0++	early up	6518
chr11	SIDT2	C+++0++	early up	4573
chr11	CCND1	C+++0++	early up	3
chr11	H19, MIR675	C+++0++	early up	13525
chr11	STARD10	C+++0++	early up	601
chr11	USP35	C+++0++	early up	32276
chr11	CTR9	C+++0++	early up	8267
chr11	TH	C0++0++	late up	2519
chr11	ENDOD1	C0++0++	late up	1870
chr11	PGR	C0++0++	late up	48908
chr11	PLA2G16	C0++0++	late up	5688
chr11	SYT12	C0++0++	late up	70
chr11	BARX2	C0++0++	late up	51194
chr11	BACE1, BACE1-AS	C0++0++	late up	2575
chr11	GAB2	C0++0++	late up	29761
chr11	CCDC86, GPR44	C0++0++	late up	25215
chr11	DDX10	C0++0++	late up	2068
chr11	DGAT2	C0++00+	late up	23427
chr11	RAB30, SNORA70E	C0++00+	late up	37210
chr11	FJX1	C00+00+	late up	177811
chr11	GRIK4	C00+00+	late up	22973
chr11	SLC3A2	C00+00+	late up	202631
chr11	WEE1	C-00-00	init down	322410
chr11	PPFIBP2	C0-0--	late down	716260
chr11	BCL9L, CXCR5	C0--0--	late down	447483
chr11	PTPRCAP	C0--00-	late down	188860

Chr	Gene Symbol	Cluster[#]	Group	Distance*
chr11	TM7SF2	C0--00-	late down	15747
chr11	VEGFB	C0--00-	late down	78200
chr11	SYTL2	C00-00-	late down	688828
chr11	EFCAB4A	C00-00-	late down	179701
chr11	CREB3L1	C00-00-	late down	68268
chr11	MUC5B	C00-00-	late down	144427
chr11	GUCY1A2	C00-00-	late down	557143
chr12	HOXC9	C+00+00	init up	9618
chr12	HOXC8	C+00+00	init up	18631
chr12	TNFRSF1A	C+00000	init up	21950
chr12	CD9	C+00000	init up	36807
chr12	HSPB8	C++++++	early up	213
chr12	HOXC4, HOXC5, HOXC6, MIR615	C++++++	early up	26383
chr12	FKBP4	C+++0++	early up	883
chr12	MIR4304, PITPNM2	C0++0++	late up	40318
chr12	B4GALNT3	C0++0++	late up	63894
chr12	NCKAP5L	C0++0++	late up	24060
chr12	NCOR2	C0++0++	late up	38
chr12	MED13L, MIR620	C0++00+	late up	108813
chr12	RHOF, TMEM120B	C0++00+	late up	36093
chr12	BRI3BP	C0++00+	late up	396
chr12	CHPT1, SYCP3	C0++00+	late up	112
chr12	METTL1	C00+00+	late up	146166
chr12	RND1	C---0--	early down	91689
chr12	GPR81	C---0--	early down	22208
chr12	NR4A1	C---0--	early down	96033
chr12	AMIGO2	C0-0--	late down	570120
chr12	GPRC5A	C0-0--	late down	303852
chr12	RARG	C0-0--	late down	62196
chr12	DNAJC22	C0-0--	late down	159193
chr12	SOCS2	C0--0--	late down	126509
chr12	KRT80	C0--00-	late down	8774
chr12	PKP2	C0--00-	late down	1122898
chr12	TSPAN9	C0--00-	late down	281264
chr12	KRT81	C0--00-	late down	41929
chr12	GALNT6	C0--00-	late down	137009
chr12	PMEL	C0--00-	late down	929507
chr12	LIMA1, MIR1293	C00-00-	late down	297183
chr12	LIN7A, MIR617, MIR618	C00-00-	late down	1183708
chr12	DRAM1	C00-00-	late down	179576
chr12	CDKN1B	Crest	rest	130198
chr12	BCL7A	Crest	rest	16614
chr13	KBTBD6	C++++++	early up	30737
chr13	LOC100289410, MCF2L	C+++0++	early up	306
chr13	ATP11A	C0++0++	late up	24585
chr13	ELF1	C0++0++	late up	2062
chr13	GAS6	C0++00+	late up	86
chr13	FARP1, MIR3170, RNF113B	C00+00+	late up	84831

Chr	Gene Symbol	Cluster[#]	Group	Distance*
chr13	C13orf27	C-00000	init down	956793
chr13	SGCG	C00-00-	late down	180887
chr14	ZFP36L1	C+00+00	init up	29380
chr14	PAX9	C+00+00	init up	41
chr14	BATF	C+00+00	init up	891
chr14	BMP4	C+00+00	init up	10237
chr14	GLRX5	C+00000	init up	4376
chr14	DHRS2	C+00000	init up	4313
chr14	FOS	C++++++	early up	18514
chr14	DEGS2	C++++++	early up	73
chr14	TMED8	C+++0++	early up	2433
chr14	ZFYVE21	C+++0++	early up	22086
chr14	ACOT4	C+++0++	early up	10186
chr14	CALM1	C+++0++	early up	94730
chr14	GSC	C0+++++	early up	125919
chr14	C14orf49	C0++0++	late up	9663
chr14	CCDC88C	C0++0++	late up	37
chr14	TMEM229B	C0++0++	late up	18215
chr14	GPR68	C0++0++	late up	22732
chr14	ITPK1, ITPK1-AS1	C0++0++	late up	235
chr14	RPS6KL1	C0++0++	late up	115
chr14	SAMD15	C0++0++	late up	2624
chr14	PIGH, PLEKHH1	C0++00+	late up	401
chr14	SIX4	C-00-00	init down	95213
chr14	FOXA1	C------	early down	54869
chr14	TGFB3	C0--0--	late down	419459
chr14	KHNYN, SDR39U1	C0--0--	late down	240971
chr14	NFATC4	C0--00-	late down	177975
chr14	NYNRIN	C0--00-	late down	209822
chr14	SLC7A8	C0--00-	late down	82073
chr14	CLMN	C00-00-	late down	45100
chr15	FAM174B	C+00000	init up	86
chr15	GCOM1, GRINL1A	C000+00	init up	184
chr15	C15orf59	C++++++	early up	8295
chr15	CD276	C+++0++	early up	11013
chr15	CT62	C+++0++	early up	10599
chr15	PEX11A	C+++0++	early up	3083
chr15	CHSY1	C+++0++	early up	2501
chr15	CA12	C0++0++	late up	889
chr15	RASGRP1	C0++0++	late up	28777
chr15	FMN1	C0++0++	late up	152
chr15	THBS1	C0++0++	late up	8113
chr15	DAPK2	C0++0++	late up	100267
chr15	AEN	C0++0++	late up	487354
chr15	SCARNA14, TIPIN	C0++00+	late up	26
chr15	TPM1	C0++00+	late up	47559
chr15	THSD4	C00+00+	late up	36547
chr15	LOC100507217	C-00-00	init down	124032

Chr	Gene Symbol	Cluster[#]	Group	Distance*
chr15	ARRDC4	C-00-00	init down	41516
chr15	ISL2	C-00000	init down	25306
chr15	BMF	C0--0--	late down	50746
chr15	SMAD6	C0--0--	late down	315377
chr15	SEMA7A	C0--0--	late down	135565
chr15	SQRDL	C0--00-	late down	1440095
chr15	FURIN	C0--00-	late down	870151
chr15	LEO1	C00-0--	late down	192652
chr15	SMAD3	C00-00-	late down	12133
chr15	CYP1A1	C00-00-	late down	45629
chr15	MIR629, TLE3	C00-00-	late down	42829
chr16	SLC7A5P2	C+00+00	init up	5408
chr16	SOCS1	C+00+00	init up	2583
chr16	RPUSD1	C+00+00	init up	30502
chr16	CORO7	C++++++	early up	200
chr16	LCAT, SLC12A4	C++++++	early up	3694
chr16	OSGIN1	C+++0++	early up	1
chr16	ADCY9	C+++0++	early up	1901
chr16	SLC7A5	C+++0++	early up	850
chr16	ABCA3	C0+++++	early up	10022
chr16	USP31	C0++0++	late up	50
chr16	CRISPLD2	C0++0++	late up	16999
chr16	CBFA2T3	C0++0++	late up	5
chr16	PDPR	C0++0++	late up	23177
chr16	AMFR	C0++00+	late up	51887
chr16	ABAT	C0++00+	late up	4685
chr16	MT1F	C00+0++	late up	27928
chr16	MIR662, MSLNL	C00+00+	late up	16493
chr16	TFAP4	C00+00+	late up	61177
chr16	TNFRSF12A	C-00000	init down	52814
chr16	CRNDE	C-00000	init down	499769
chr16	ITPRIPL2	C0--0--	late down	873289
chr16	IRX3	C0--0--	late down	137821
chr16	LITAF	C0--00-	late down	113095
chr16	YPEL3	C0--00-	late down	1870
chr16	CLN3	C0--00-	late down	593440
chr16	KREMEN2	C0--00-	late down	109933
chr16	CLDN9	C0--00-	late down	61693
chr16	TUBB3	C0--00-	late down	154001
chr16	FA2H	C0--00-	late down	254925
chr16	ROGDI	C0--00-	late down	112040
chr16	ATP2A1, RABEP2	C0--00-	late down	276929
chr16	KIFC3	C0--00-	late down	152978
chr16	PPL	C0--00-	late down	246045
chr16	CCDC64B	C0--00-	late down	37632
chr16	RAB26	C0--00-	late down	179921
chr16	PRRT2	C00-0--	late down	274966
chr16	AKTIP, RBL2	C00-00-	late down	229374

Chr	Gene Symbol	Cluster[#]	Group	Distance*
chr16	KIAA0513	C00-00-	late down	50302
chr16	GDPD3	C00-00-	late down	23197
chr16	CDYL2	C00-00-	late down	4831
chr17	TBX2	C+00+00	init up	33441
chr17	RASD1	C+00+00	init up	18209
chr17	CBX4	C+00+00	init up	25160
chr17	SLC16A3	C+00000	init up	28001
chr17	GGT6	C+00000	init up	15426
chr17	TRIM16	C+00000	init up	8812
chr17	KCTD11	C+00000	init up	128370
chr17	SMTNL2	C++++++	early up	7973
chr17	RARA	C++++++	early up	888
chr17	MIR3615, SLC9A3R1	C++++++	early up	904
chr17	ACE	C++++++	early up	2376
chr17	SOCS3	C++++++	early up	36242
chr17	TOB1	C++++++	early up	17482
chr17	FDXR	C++++++	early up	3954
chr17	METRNL	C++++++	early up	3289
chr17	CANT1	C++++++	early up	7929
chr17	IGFBP4	C+++0++	early up	4840
chr17	RAPGEFL1	C+++0++	early up	2931
chr17	UNC119	C+++0++	early up	15628
chr17	KRT13	C+++0++	early up	1290
chr17	ALDOC	C+++0++	early up	39933
chr17	PLCD3	C+++0++	early up	1462
chr17	TMEM104	C0++0++	late up	6520
chr17	TEX14	C0++0++	late up	19696
chr17	KRT15	C0++0++	late up	12115
chr17	ERN1	C0++0++	late up	1863
chr17	FOXK2	C0++0++	late up	2067
chr17	WDR45L	C0++00+	late up	11620
chr17	MYBBP1A, SPNS2	C0++00+	late up	20555
chr17	IMP5, LOC100128977, LOC100130148 MAPT, STH	C0++00+	late up	5486
chr17	COL1A1	C0++00+	late up	35544
chr17	CACNG4	C0++00+	late up	169
chr17	CUEDC1	C0++00+	late up	3388
chr17	TRIM47	C00+0++	late up	119879
chr17	PITPNC1	C00+00+	late up	89984
chr17	GRB7	C---0--	early down	276530
chr17	C17orf58	C---0--	early down	298229
chr17	RNF43	C0--0--	late down	222256
chr17	CDC42EP4	C0--0--	late down	300346
chr17	ERBB2	C0--00-	late down	226736
chr17	ST6GALNAC2	C0--00-	late down	53648
chr17	MMD	C0--00-	late down	452837
chr17	PMP22	C0--00-	late down	82501
chr17	PRR15L	C0--00-	late down	263959

Chr	Gene Symbol	Cluster#	Group	Distance*
chr17	HDAC5	C0--00-	late down	95988
chr17	SLC16A5	C00-00-	late down	219350
chr17	C17orf28	C00-00-	late down	103155
chr17	ATP2A3	C00-00-	late down	11908
chr18	SMAD7	C++++++	early up	728
chr18	RAB31	C0++0++	late up	2769
chr18	HAUS1	C0++00+	late up	9385
chr18	TGIF1	C-00-00	init down	26978
chr18	RAB27B	C0--0--	late down	802676
chr18	ACAA2	C0--00-	late down	321350
chr18	NEDD4L	C00-00-	late down	249863
chr19	ZNF552	C+00+00	init up	525
chr19	ZFP36	C+00+00	init up	255325
chr19	ZNF823	C+00+00	init up	810432
chr19	MIDN	C+00+00	init up	3108
chr19	LOC284454, MIR24-2	C+00+00	init up	410
chr19	C19orf33, YIF1B	C+00000	init up	4764
chr19	NR2F6	C+00000	init up	3542
chr19	OCEL1	C+00000	init up	13978
chr19	KCNK6	C++++++	early up	1565
chr19	ADAMTSL5	C++++++	early up	131
chr19	PLIN5	C++++++	early up	1742
chr19	GPR77	C++++++	early up	132
chr19	GADD45B	C++++++	early up	47235
chr19	CNTD2	C++++++	early up	276
chr19	CYP4F22	C++++++	early up	19755
chr19	S1PR5	C++++++	early up	15363
chr19	BLVRB	C+++0++	early up	446
chr19	STK11	C+++0++	early up	1123
chr19	PAK4	C+++0++	early up	4401
chr19	REEP6	C+++0++	early up	21870
chr19	PIN1	C+++0++	early up	10980
chr19	CHST8	C0++0++	late up	4725
chr19	CYP4F11	C0++0++	late up	286
chr19	SLC25A42	C0++0++	late up	546
chr19	ARHGEF18	C0++0++	late up	3409
chr19	SBNO2	C0++0++	late up	812
chr19	KDM4B	C0++0++	late up	82
chr19	HDGFRP2, PLIN4	C0++00+	late up	19234
chr19	ZNF91	C0++00+	late up	5010
chr19	JUNB	C-00-00	init down	242117
chr19	SERTAD3	C-00-00	init down	21668
chr19	IER2	C-00-00	init down	87353
chr19	SIX5	C-00000	init down	12753
chr19	LYPD3	C---0--	early down	129655
chr19	MYADM	C---0--	early down	552653
chr19	BBC3, MIR3190, MIR3191	C---0--	early down	92988
chr19	SUV420H2	C0--0--	late down	69134

Chr	Gene Symbol	Cluster[#]	Group	Distance*
chr19	PLEKHF1	C0--0--	late down	244914
chr19	MKNK2	C0--0--	late down	133921
chr19	GRAMD1A	C0--0--	late down	89
chr19	NOTCH3	C0--00-	late down	278593
chr19	PPAP2C	C0--00-	late down	163582
chr19	C19orf21	C0--00-	late down	23798
chr19	EPOR	C0--00-	late down	878765
chr19	LASS4	C0--00-	late down	114731
chr19	TLE2	C0--00-	late down	196807
chr19	DYRK1B	C00-00-	late down	408032
chr19	TJP3	Crest	rest	8917
chr20	MAFB	C+00+00	init up	1004324
chr20	TBC1D20	C+00000	init up	16645
chr20	SDC4	C+00000	init up	5396
chr20	C20orf96	C+00000	init up	144045
chr20	MIR663	C000+00	init up	55
chr20	SMOX	C++++++	early up	10188
chr20	SNAI1	C++++++	early up	65827
chr20	TFAP2C	C++++++	early up	25
chr20	C20orf160	C+++0++	early up	22573
chr20	C20orf24, SLA2, TGIF2 TGIF2-C20ORF24	C+++0++	early up	12568
chr20	KCNK15	C0++0++	late up	18911
chr20	RIMS4	C0++0++	late up	44519
chr20	PARD6B	C0++0++	late up	4015
chr20	RRBP1	C0++0++	late up	35734
chr20	EPB41L1	C0++00+	late up	28967
chr20	C20orf177, PPP1R3D	C-00-00	init down	47661
chr20	TRIB3	C-00000	init down	52564
chr20	ZNF217	C------	early down	60693
chr20	SALL4	C0--0--	late down	65439
chr20	BMP7	C0--00-	late down	70239
chr21	SIK1	C000+00	init up	6
chr21	MIR3648	C000+00	init up	1266
chr21	TMPRSS3	C+++0++	early up	120
chr21	TMPRSS2	C+++0++	early up	98416
chr21	TFF1	C+++0++	early up	219
chr21	NRIP1	C0++0++	late up	59279
chr21	C21orf96, RUNX1	C0++0++	late up	152173
chr21	TFF3	C0++00+	late up	15204
chr21	TIAM1	C00+00+	late up	29075
chr21	DSCAM	C00+00+	late up	533
chr22	SUSD2	C+00000	init up	14479
chr22	BCR, FBXW4P1	C000+00	init up	11376
chr22	C1QTNF6	C++++++	early up	1023
chr22	CDC42EP1	C++++++	early up	2713
chr22	GATSL3	C++++++	early up	1601
chr22	CRKL	C++++++	early up	463

Chr	Gene Symbol	Cluster[#]	Group	Distance*
chr22	A4GALT	C+++0++	early up	8491
chr22	SEC14L2	C+++0++	early up	11779
chr22	AP1B1, MIR3653, SNORD125	C0++0++	late up	2006
chr22	BIK	C0--0--	late down	88123
chr22	PIK3IP1	C0--0--	late down	61993
chr22	INPP5J	C0--00-	late down	106201
chr22	PRODH	C0--00-	late down	236713
chr22	MB	C00-00-	late down	1397974
chr22	CERK	C00-00-	late down	245993
chrX	SLC10A3	C+00000	init up	50976
chrX	SOX3	C++++++	early up	386
chrX	MGC16121, MIR424, MIR503	C+++0++	early up	21189
chrX	ARMCX6	C+++0++	early up	871861
chrX	ARMCX3	C0++0++	late up	877106
chrX	ZNF185	C0++00+	late up	1164
chrX	TMEM164	C0++00+	late up	50472
chrX	DYNLT3	C0++00+	late up	57418
chrX	SYTL5	C00+00+	late up	94144
chrX	OPHN1	C00+00+	late up	29032
chrX	TSPYL2	C---0--	early down	171601
chrX	RIBC1	C00-00-	late down	131287

Table 6. E2 regulated novel intergenic transcripts are summarized.* Joined qips regions are generated by using paired-end RNA-Seq data.

Chr	Begin	End	Name	Cluster	Group
chr1	192215873	192217022	qips2982	C00+00+	late up
chr2	238392574	238393438	qips48269	C+00+00	init up
chr6	80146681	80146791	qips75521	C0--00-	late down
chr7	42422769	42425877	qips78688	C00-00-	late down
chr7	56358787	56359104	qips79035	C000-00	init down
chr8	115126711	115126894	qips80554	C0--00-	late down
chr9	12972569	12973458	qips84844	C00+00+	late up
chr10	32046551	32048990	qips10586	C+00+00	init up
chr10	125710888	125714063	qips9622	C00+00+	late up
chr10	8910071	9016796	Joined qips 1*	C---0--	early down
chr12	52741798	52748478	qips19870	C++++++	early up
chr12	115026165	115026589	qips17618	C0++0++	late up
chr15	56363620	56366027	qips28543	C0--00-	late down
chr15	63729165	63730439	qips28964	C0++0++	late up
chr16	4233319	4235259	qips31787	C++++++	early up
chr17	30438667	30439383	qips34660	C0++00+	late up
chr18	20463035	20465840	qips38384	C0++00+	late up
chr19	15679880	15696024	Joined qips 2*	C0+++++	early up
chr19	16013228	16014034	qips40412	C000+00	init up
chr20	24911985	24913601	qips51646	C0--00-	late down
chr20	46639363	46640113	qips52805	C000+00	init up
chr20	46749673	46755514	qips52810	C00+00+	late up
chr20	52463272	52475949	qips53108	C00-00-	late down
chr20	52480126	52482710	qips53110	C---0--	early down
chr20	55310082	55338977	qips53165	C+++0++	early up
chrX	38338683	38339016	qips89157	C0++00+	late up
chrX	48634254	48635296	qips89448	C0--00-	late down
chrX	94754594	94754890	qips90289	C0++0++	late up
chrX	119118484	119125834	qips87854	C0--00-	late down