MOLECULAR MECHANISMS UNDERLYING THE REGULATION OF GENE EXPRESSION AND GROWTH OF BREAST CANCER CELLS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Miao Sun

August 2013

MOLECULAR MECHANISMS UNDERLYING THE REGULATION OF GENE EXPRESSION AND GROWTH OF BREAST CANCER CELLS

Miao Sun, Ph.D.

Cornell University 2013

Breast cancer is a serious public health issue, and a full understanding of its etiology and pathophysiology is a primary focus in the field. Molecularly, the combined action of a plethora of factors in multiple pathways is involved in the regulation of the breast tumoriogenic process. Characterization of a more complete spectrum of the molecular factors will provide insights into the development of new and improved diagnostic, prognostic and therapeutic tools for treating breast cancer. To this end, my studies utilize a combination of molecular biology and bioinformatic methods, to uncover the mechanisms underlying the regulation of gene expression and growth in human breast cancer cells.

To investigate the molecular crosstalk of the estrogen and c-Jun N-terminal kinase 1 (JNK1) signaling pathways, I monitored the genomic localization of estrogen receptor α (ERα) and JNK1 in basal and estrogen-stimulated MCF-7 breast cancer cells. I found that JNK1 binds to the promoter of many genes. ERα is required for the binding of JNK1 to the estrogen-induced sites, and JNK1 in turn functions as a coregulator of ERα. The convergence of ERα and JNK1 at target promoters regulates estrogen-dependent gene expression, as well as downstream estrogen-dependent cell growth responses.

Furthermore, the implication of long noncoding RNAs (lncRNAs) in breast cancer is also coming to light. I developed a computational approach that integrates information from multiple

genomic datasets, and generated a comprehensive catalog of 1888 expressed lncRNA genes in MCF-7 cells. Almost half of them are first annotated in this study, and more than a quarter are estrogen-regulated. Close examination revealed many interesting features. Interestingly, cell type-specific expression of lncRNAs predicts the intrinsic molecular subtypes of breast cancer, suggesting its potential utility as prognostic marker. Lastly, by selecting lncRNAs with elevated expression in breast tumors, and whose differential expression across a wide spectrum of tissues and cell types correlates with important cell viability genes, we identified a number of lncRNAs that are required for the normal growth of human breast cancer cells.

Collectively, my studies expanded our understanding of the molecular mechanisms underlying breast cancer biology, and suggested new targets for therapeutic interventions.

BIOGRAPHICAL SKETCH

Miao Sun was born and raised in Wuxi, Jiangsu Province, China. At the age of 14, she was nominated and awarded a full scholarship from the Ministry of Education, Singapore, and left for the Nanyang Girls' High School in Singapore to continue her study. She then went on to the Raffles' Junior College and the National University of Singapore, where she gained a strong background in general biology, including topics in physiology, microbiology, biodiversity, biochemistry, molecular and cell biology, pharmacology, immunology, neural biology and computational biology. During the summer of 2006, she received a Summer Undergraduate Research Fellowship from the California Institute of Technology to participate in a summer research program in the lab of Dr. David Baltimore. In 2007, she completed her undergraduate studies and received a B.Sc. with 1st Class Honors, with a major in Biomedical Sciences and a minor in Computational Sciences. In the same year, she joined the Graduate Field of Biochemistry, Molecular and Cell Biology at Cornell University, and in 2008, she joined the lab of Dr. W. Lee Kraus, and relocated with the lab in 2010 to the University of Texas Southwestern Medical Center at Dallas. During she graduate studies, Miao integrated molecular biology, cell-based and computational approaches to study the functional roles played by cellular kinases and long noncoding RNAs in mediating transcriptional regulation and cellular outcomes in estrogen-responsive human breast cancer cells. She received multiple awards, including the American Heart Association Pre-doctoral Fellowship in 2009. In 2013, she defended her Ph.D. thesis.

ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Dr. W. Lee Kraus, for taking me as a student 6 years ago, and for guiding me along this scientific journey. I am deeply grateful for what he has contributed to my scientific development, providing continual support and lots of opportunities, so that I can evolve from a naïve undergraduate to an actual scientific researcher. I would also like to thank my committee members, Dr. John Lis, Dr. Andrew Clark and Dr. Jeffery Pleiss, for following my research, providing scientific insights and for all their advice and support throughout the years.

Thank you to all my lab-mates in the Kraus Lab, you are more than just my colleagues, but friends, allies, even a second family to me. I have enjoyed great conversations and scientific interactions with each one of you, and have received lots of help from you guys both inside and outside of the lab. On the Ithaca side, I wish to especially thank Dr. Matthew Gamble, a past lab member, and Dr. Charles Danko, for introducing me to the field of computational biology and patiently teaching me the basics of programming. I have learnt so much from both of you. Also thanks to past lab member Dr. Gary Isaacs, my first bay-mate in the Kraus lab, for a fruitful collaboration, and to Dr. Raga Krishnakumar for your generosity, both at a scientific level and a personal level. On the Dallas side, special thanks to my dear “comrades”, Xin, Ziying, Bryan, and Shrikanth, who have moved with me from Ithaca to Dallas. We have really bonded and depended on each other to “survive” those “difficult” times. Thank you Xin for being a wonderful friend, and a fellow n^{th}-year graduate student who has travelled the path with me. Despite of all the odds and difficulties, we have hanged on together to finally see the end. Thank you Ziying for being a great bay-mate and a great buddy. I have particularly enjoyed our “dates”

with shopping sprees and dinners, and those scientific conservations as well as “insightful” arguments regarding “important national and international issues”. Thanks to Dr. Shrikanth Gadad and Rui Li for collaborating with me and contributing to my project experimentally and intellectually, and to Dr. Hector Franco and Dr. Daeseok Kim for making the lab a “fun” place. To the rest of my lab, Minho, Anusha, Tulip, Keun, Shino, Rachel, Jane, Debbie, Pam and Connie, I was lucky to have worked and interacted with all of you.

Thank you to all my other friends, including Andrew Manford, Satyaki Prasad, Molly Shook, Liu Yang, Cheng Chen, and especially the newly “doctorated” Minxing Li, your friendship has been a true blessing, and while I sincerely wish you all the success as we move on to our separate ways, I have missed and will miss all you very much. Special thanks to Lu Huang, my boyfriend, for having travelled the journey with me from Ithaca to Dallas, literally, and for being the most caring person to me in the past 3 years, sharing with me happy times and sad times. It is great comfort to have you as my rock. Thank you!

Most importantly, words cannot even begin to express my gratitude towards my parents, who have showered me with unconditional love throughout my life. You have always put me ahead of yourselves and everything else, and are willing to give me everything you have, and still thinking that is not enough. Despite of being physically apart for about 15 year, our hearts are always together. Thank you for everything! I look forward to moving closer to you and I am determined to spend more time with you in the next phase of my life!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

5C	Chromosome Conformation Capture Carbon Copy
3P-seq	Poly(A) Position Profiling by Sequencing
ANRIL	Antisense noncoding RNA in the INK4 locus
agRNA	Anti-Gene RNA
asRNA	Antisense RNA
ATP	Adenosine Triphosphate
AP-1	Activating Protein-1
BP	Base Pairs
CAGE	Cap-Analysis Gene Expression
CBP	CREB Binding Protein
CEBPα	Ccaat-Enhancer-Binding Protein α
ceRNA	Competing Endogenous RNA
CHART	Capture Hybridization Analysis of RNA Targets
ChIP	Chromatin Immunoprecipitation
ChIRP	Chromatin Isolation by RNA Purification
CNC	Coding-Noncoding Coexpression Network
CPC	Coding Potential Calculator
CSF	Coding Substitution Frequency
E2	17β-Estradiol
EGF	Epidermal Growth Factor
EGFR	Epidermal Growth Factor Receptor
ENCODE	Encyclopedia of DNA Elements

ER	Estrogen Receptor
ERE	Estrogen Response Element
eRNA	Enhancer RNA
ERK	Extracellular Signal-Regulated Kinase
ESC	Embryonic Stem Cell
EZH2	Enhancer Of Zeste Homolog 2 (Drosophila)
FISH	Fluorescence In Situ Hybridization
FOS	Finkel-Biskis-Jinkins Osteosarcoma Virus Oncogene
GAPDH	Glyceraldehyde 3-Phosphate Dehydrogenase
GAS5	Growth Arrest-Specific 5
GFP	Green Fluorescent Protein
GR	Glucocorticoid Receptor
GO	Gene Ontology
GRO-seq	Global Nuclear Run-On Sequencing
H3K4me1	Histone 3 Lysine 4 Monomethylation
H3K4me3	Histone 3 Lysine 4 Trimethylation
H3K27ac	Histone 3 Lysine 27 Acetylation
H3K36me3	Histone 3 Lysine 36 Trimethylation
hnRNP	Heterogeneous Nuclear Ribonucleoproteins
HOTAIR	HOX antisense intergenic RNA
HOTTIP	HOXA transcript at the distal tip
HER2	Human Epidermal Growth Factor Receptor 2
H-InvDB	H-Invitational Databases

HuR	Human Antigen R
IGF	Insulin-Like Growth Factor
JNK	Jun N-Terminal Kinase
JUN	Avian sarcoma virus 17 oncogene
KB	Kilobase Pairs
LincRNA	Long Intergenic Noncoding RNA
LM-PCR	Ligantion-Mediated PCR
LncRNA	Long Noncoding RNA
LSD1	Lysine (K)-Specific Demethylase 1
MALAT1	Metastasis-Associated Lung Adenocarcinoma Transcript 1
MAPK	Mitogen-Activated Protein Kinase
MAST	Motif Alignment and Search Tool
MCF7	Michigan Cancer Foundation - 7
MEME	Multiple Em for Motif Elicitation
MLL	Mixed Lineage Leukemia
MPK-1	MAPK Phosphotase 1
mRNA	Messenger RNA
NAT	Natural Antisense Transcript
NET-seq	Nascent Elongating Transcript Sequencing
ncFANs	Non-coding RNA Function Annotation Server
ncRNA-a	ncRNA-Activating
NFκB	Nuclear Factor κ-Light-Chain-Enhancer of Activated B Cells
NR	Nuclear Receptor

NRED	Noncoding RNA Expression Database
NT	Nucleotides
PANDA	P21 Associated ncRNA DNA Damage Activated
PINC	Pregnancy-Induced Noncoding RNA
PR	Progesterone Receptor
PRC2	Polycomb Repressive Complex 2
Pol II	RNA Polymerase II
POU	PIT1, OCT1, Unc-86
PPARγ	Peroxisome Proliferator-Activated Receptor
qPCR	Quantitative Polymerase Chain Reaction
RACE	Rapid Amplification of Polymerase Ends
RIP	RNA Immunoprecipitation
rRNA	Ribosomal RNA
RT-qPCR	Reverse Transcription-qPCR
shRNA	Short Hairpin RNA
siRNA	Small Interfering RNA
snoRNA	Small Nucleolar RNA
snRNA	Small Nuclear RNA
SP1	Simian-Virus-40-Protein-1
SRA	Steroid Receptor RNA Activator
SRC	Steroid Receptor Coactivators
STAU1	Staufen, RNA Binding Protein, Homolog 1
SWI/SNF	SWItch/Sucrose Non-Fermentable

TESS	Transcription Element Search System
TINCR	Terminal Differentiation-Induced ncRNA
tRNA	Transfer RNA
TSS	Transcription Start Site
TTS	Transcription Termination Site
UTR	Un-Transcribed Region
XCI	X Chromosome Inactivation
XIST	X-Inactive-Specific Transcript

CHAPTER 1

An Introduction to the Mechanisms Underlying the Regulation of Gene Expression and Growth of Breast Cancer Cells

1.1. Summary

Extensive efforts have been undertaken to improve our understanding of breast cancer. We now know that the estrogen signaling pathway has established roles in the development of breast cancer, and its interplay with the growth factor signaling pathways has been associated with endocrine resistance in breast cancer. Moreover, the recent revelation of a large number of long noncoding RNAs (lncRNAs), has again introduced new players that are involved in the breast tumorigenic process. Clearly, breast cancer is a complex disease that involves the interplay of a wide variety of molecular factors. Therefore, my studies will aim to identify a more complete spectrum of such molecular factors, and to uncover molecular mechanisms underlying the regulation of gene expression and growth in breast cancer cells.

1.2. Breast Cancer as a "Multi-Factorial" Disease

Breast cancer is the most common form of cancer and the second leading cause of cancer death in American women. The estimated annual incidence of breast cancer worldwide is about one million cases (Dumitrescu and Cotarla 2005). Clearly, this is a serious public health issue, and efforts to understand the etiology and pathophysiology of the disease are essential. To this end, extensive efforts in breast cancer research has uncovered important aspects of the molecular basis of the disease, and has been instrumental in pushing forward significant medical advances in both breast cancer detection and treatment. We now know that breast cancer is a complex disease that manifests in many different forms, which often lead to different prognosis and responses to treatment options. At the molecular level, the combined action of a plethora of protein factors in multiple signaling pathways is involved in the regulation and fine-tuning of the breast tumoriogenic process. Identification and characterization of a more complete spectrum of

molecular factors at play will further our knowledge of breast cancer biology and provide insights into the development of new and improved diagnostic, prognostic and therapeutic tools for treating breast cancer.

1.3. Estrogen Signaling Pathways in Breast Cancer

Breast cancer presents itself as a classical model of hormone-dependent malignancy. It is well accepted that estrogens, the primary female sex hormone that play a central role in normal mammary gland development, are also pivotally involved in mammary tumoriogenesis as a result of their potent mitogenic effects (Manavathi, Dey et al. 2013). A prolonged or increased exposure to estrogens has been associated with an increased breast cancer risk (Pike, Gerkins et al. 1979; Begg, Kuller et al. 1987), while reduced exposure to estrogens results in the opposite effects (Hulka 1997).

The molecular activities of estrogens are mediated through the estrogen receptor (ER) proteins, which consist of two isoforms, ERα (Greene, Gilna et al. 1986) and ERβ (Kuiper, Enmark et al. 1996), and belong to a conserved superfamily of nuclear receptor proteins that function as sequence-specific, DNA-binding transcription factors in the nucleus (Mangelsdorf, Thummel et al. 1995; Kininis, Chen et al. 2007). In the classical model (Fig. 1.1A), ERs dimerize upon ligand activation and bind directly to genomic DNA through estrogen response element (ERE) sequences (Kumar and Chambon 1988). In an alternative model (Fig. 1.1B), liganded ERs indirectly interact with genomic DNA through immediate transcription factors (e.g. NF-κB, Sp1 and AP-1), or tethering factors, via their recognition elements (Gaub, Bellard et al. 1990; Weisz and Rosales 1990; Umayahara, Kawamori et al. 1994; Kushner, Agard et al. 2000). In both cases, a variety of cofactors including (1) histone modifying enzyme complexes that

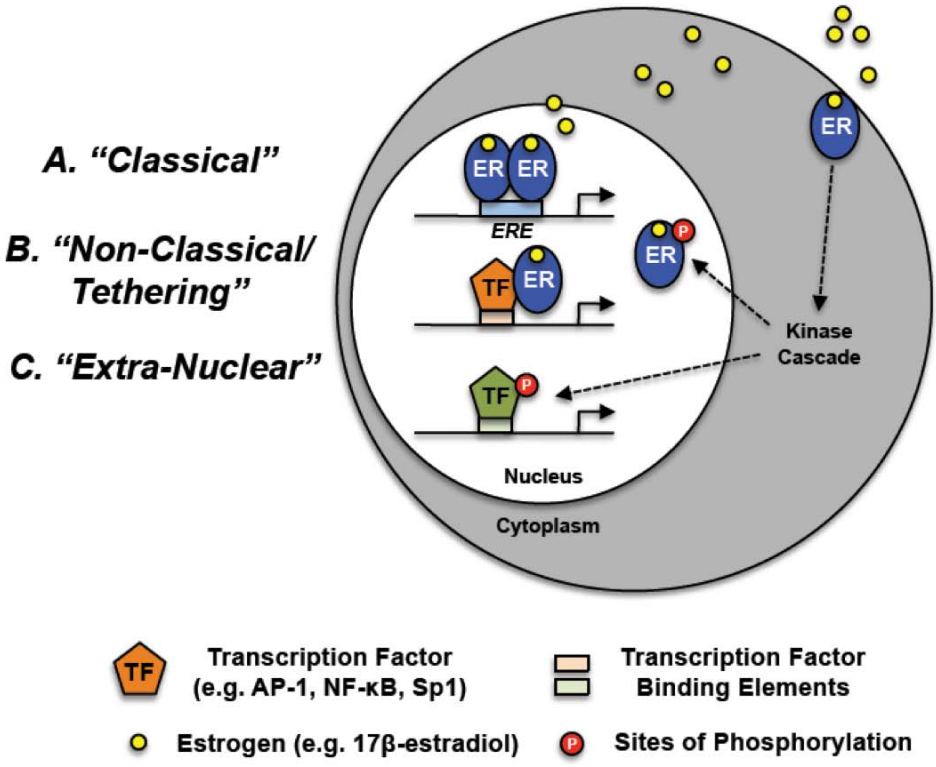


Figure 1.1 Estrogen-dependent signaling pathways.

Estrogen signaling pathways include: (A) "Classical", the ligand-dependent binding of ER directly to the ERE; (B) "Non-Classical/Tethering", the ligand-dependent binding of ER to DNA-bound transcription factors, the so-called tethering factors, and (C) "Extra-Nuclear", the activation of kinase cascades by membrane-associated ER.

contain members of the steroid receptor coactivator family of proteins as the receptor binding subunit (Leo and Chen 2000), (2) chromatin remodeling complexes such as SWI/SNF (Guyon, Narlikar et al. 1999; Robyr, Wolffe et al. 2000), and (3) Mediator complexes which contain Med220/TRAP220 as the primary receptor binding subunit (Malik and Roeder 2000; Rachez and Freedman 2001), act together with activated ERs to modify histones, alter chromatin structure, and regulate the recruitment and activity of RNA polymerase II (Pol II) transcriptional machinery (Fig. 1.2) (Wong, Lin et al. 2002). The interplay between ERs and these transcriptional cofactors leads to profound changes in the expression of estrogen-responsive genes that are associated with hormone-dependent physiological outcomes, such as in the case of promoting the development of breast cancer.

About 2/3 of human breast cancers are ER-positive and likely to be estrogen-responsive at the time of diagnosis. Therefore, aromatase inhibitors that suppress estrogen production, and anti-estrogens that target ERs to antagonize the effects of estrogens, collectively known as the endocrine therapy, are often used clinically as first- and second-line treatment options for these early-stage breast cancers. Nevertheless, about half of these patients are tolerant or acquire resistance to endocrine therapy, and breast cancer remains a devastating disease. Additional factors and pathways are involved in breast cancer biology and we are clearly far away from gaining a full understanding.

1.4. Interplay with the MAPK Signaling Pathways

In addition to the estrogen signaling pathway, components of the mitogen-activated protein kinases (MAPKs) pathways have also been implicated in hormone-dependent breast cancer(Smith 1998; Lange 2004). The MAPK pathways are mediated by the MAP kinases

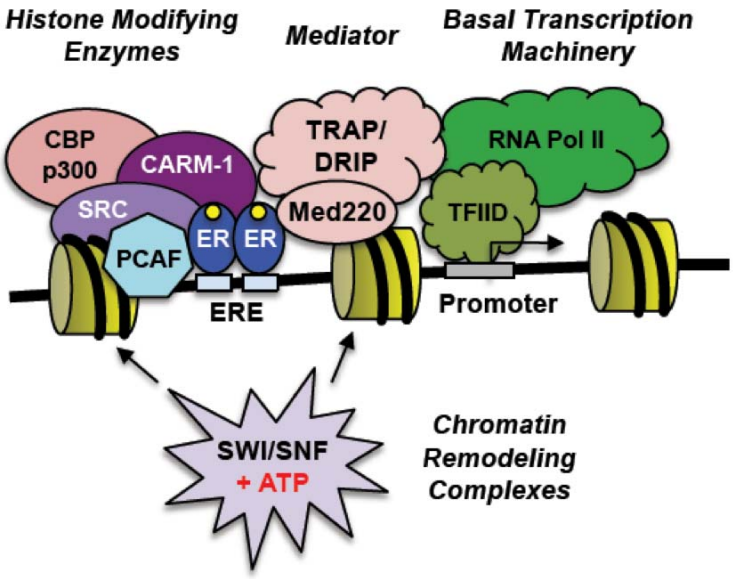


Figure 1.2 ER-mediated activation of gene expression through collaboration with additional transcriptional coregulators.

After binding estrogen, ER either dimerizes and binds to ERE (the classical pathway), or tethers to DNA-bound transcription factors (the non-classical, tethering pathway), where it then recruits a cohort of factors such as histone modifying proteins, chromatin remodeling proteins, and proteins associated with the basal transcriptional machinery. The classical pathway of ER-mediated transcriptional activation is illustrated here as an example.

comprising the extracellular signal-regulated kinases (ERKs), the c-Jun N-terminal kinases (JNKs) and p38. They are activated by upstream kinases as extracellular signals such as the growth factors act through membrane-associated receptors to initiate a signaling cascade, and function to phosphorylate downstream effectors to control a variety of cellular processes such as cell proliferation and cell survival programs.

There is a functional interplay between the estrogen and the MAPK signaling pathways. ERα is significantly phosphorylated at Serine-118, in response to either estrogen binding or activation of the MAPK pathway, and this phosphorylation influences the transactivation activity of ERα (Fig. 1.1C) (Lannigan 2003). The JNK family of MAPKs, as they are named, phosphorylate c-Jun at its N-terminus to modulate the activity of the ERα-associated transcription factor AP-1 (Fig. 1.1C) (Hibi, Lin et al. 1993; Dai, Rubie et al. 1995; Ip and Davis 1998). Indeed, increased activity of the MAPK pathway is one of the hallmarks of more aggressive cancers and is often associated with the aforementioned hormone-refractory breast cancer. In these breast cancer cells, it is believed that the cellular physiology switches from ERα nuclear-initiated pathways to increased involvement of extra-nuclear-activated MAPK pathways (Sivaraman, Wang et al. 1997; Santen, Song et al. 2002; Hutcheson, Knowlden et al. 2003; Britton, Hutcheson et al. 2006; Jordan and O'Malley 2007; McGlynn, Kirkegaard et al. 2009).

At the molecular level, ERα seems to be a possible point of convergence between the estrogen and MAPK signaling pathways in ERα-positive human breast cancer cells. In the absence of estrogens, ERα can be activated by growth factors such as the epidermal growth factor (EGF) and the insulin-like growth factors (IGFs) (Bunone, Briand et al. 1996; Ignar-Trowbridge, Pimentel et al. 1996). EGF is known to stimulate signaling via the classical MAPK cascade, and hormone-refractory breast tumors are typically dependent on the overexpression of

EGF receptor (EGFR) and HER2, the membrane-associated receptors for EGF (Benz, Scott et al. 1992; Pietras, Arboleda et al. 1995; Kurokawa, Lenferink et al. 2000; Nicholson, Hutcheson et al. 2001). Lupien and colleagues have demonstrated in their study that in MCF-7 cells, an ERα-positive human breast cancer cell line, EGF signaling resulted in a unique set of ERα genomic targets, which is distinct from the estrogen-activated ERα cistrome but consistent with the molecular profiles of HER2-positive human breast cancer cells (Lupien, Meyer et al. 2010). Their findings suggested a molecular explanation for endocrine resistance as observed in a subset of ERα-positive breast cancers, and demonstrated an important role of growth factor/MAPK signaling in estrogen-responsive breast cancer cells.

In addition, Madak-Erdogan and colleagues have shown that upon activation by the estrogens, ERK2 physically interacts with ERα and colocalizes with ERα at chromatin binding sites across the genome in MCF-7 breast cancer cells (Fig. 1.3) (Madak-Erdogan, Lupien et al. 2011). This genomic colocalization leads to regulation of estrogen-dependent gene expression and cell proliferation programs (Fig. 1.3) (Madak-Erdogan, Lupien et al. 2011). Their results established a role of ERK2, a MAP kinase, as an ERα coregulator, which functions at gene promoters and distal enhancer to modulate the genomic actions of ERα, resulting in the regulation of transcriptional outcomes and cell growth responses.

As mentioned earlier, JNK kinases are the other MAP kinases that are potentially involved in a functional crosstalk with the estrogen signaling pathway, due to their association with the ERα-associated tethering factor AP-1. Nevertheless, to what extent this functional crosstalk occurs and affects the physiology of estrogen-responsive breast cancer cells, and whether JNK kinases, such as JNK1, behave in a similar way as ERK2, is still unknown and will be addressed in Chapter 2 of my thesis.

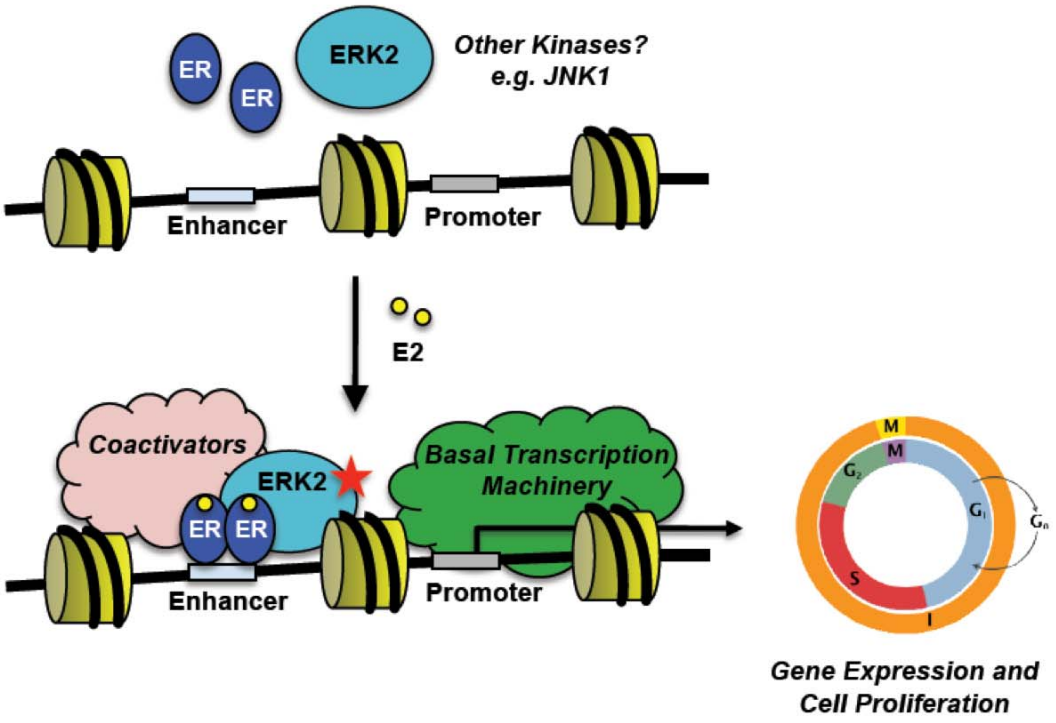


Figure 1.3 ERK2 as an ER coregulator in regulating gene and proliferation programs.

Model depicting the relationship between ERK2 and ERα in the hormonal regulation of gene expression and cell proliferation. Estrogen signaling leads to rapid activation of ERK2 (red star) and its colocalization with ERα at enhancer binding sites (and/or gene promoters). ERK2 collaborate with ERα and a cohort of additional coactivators to regulate hormone stimulation of prolideration and cell cycle-related genes. The findings indicate that ERK2, and possibly other cellular kinases, not only possess a signaling function but also a nuclear role at chromatin in facilitating the estrogen/ER-dependent transcriptional regulation. Modified from Madak-Erdogan, Lupien *et al.* 2011.

1.5. LncRNAs : A New Class of Regulators in Breast Cancer

Traditionally, identification and characterization of important players in the regulation of gene expression and cellular outcomes in human breast cancer has been focused on proteins. Nevertheless, recent advances in high-throughput sequencing technologies have revealed that the genome is extensively transcribed, yielding a large repertoire of noncoding RNAs. This includes long noncoding RNAs, mRNA-like molecules that do not code for proteins, which are emerging as a new class of RNAs that have significant impact on almost all aspects of life. While the study of lncRNA function is still in its infancy, a role for a number of these transcripts has recently been established in cancer in general, and more specifically, in breast cancer.

HOTAIR, for example, is one of the best-characterized lncRNAs to date. It is highly induced in metastatic breast cancer samples, and has been shown to be an independent predictor of breast cancer survival and metastasis-free survival (Gupta, Shah et al. 2010). Over-expression of *HOTAIR* in breast cancer cell lines results in increased cell invasion in vitro and metastasis in vivo (Gupta, Shah et al. 2010). In contrast to *HOTAIR*, which shows elevated expression in more aggressive breast cancers, the lncRNA *GAS5* is expressed at reduced levels in tumors compared to unaffected breast epithelial tissues (Mourtada-Maarabouni, Pickard et al. 2009). Consistently, *GAS5* expression induces growth arrest and apoptosis, and reduced expression of *GAS5* in breast cancers is associated with a poorer prognosis (Mourtada-Maarabouni, Pickard et al. 2009).

SRA is another example of an lncRNA that has been implicated in mammary carcinogenesis. It serves as a molecular scaffold that coordinates the functions of various transcription factors and coregulators, including ERα and its well-known coactivator SRC (Lanz, McKenna et al. 1999). Possibly through interacting and modulating the activity of ERα and its cofactors, *SRA* serves as a regulator of breast tumoriogenesis. In animal models, crossing *SRA*

transgenic mice with MMTV-ras mice, a model highly susceptible to the development of mammary tumors, reduces the incidence of mammary neoplasia (Lanz, Chua et al. 2003). In contrast, knockdown of *SRA* in MDA-MB-231 human mammary cancer cells, a model of invasive breast cancer with elevated expression of *SRA*, leads to reduced invasiveness (Foulds, Tsimelzon et al. 2010). Why reduced expression of *SRA* in one system and overexpression in another would both reduce aspects of cancer development and progression remains to be determined in future studies.

The abovementioned examples represent individual lncRNAs that have been implicated in the initiation and progression of breast tumorigeneisis. Furthermore, transcriptome profiling using global nuclear run-on sequencing (GROseq, a method for mapping transcriptionally active RNA polymerases across the genome), has identified a large number of lncRNA-like transcripts in MCF-7 human breast cancer cells (Fig. 1.4) (Hah, Danko et al. 2011). About a quarter of these transcripts are regulated by 17β-estradiol (E2), and many of the E2-upregulated lncRNAs have ERα binding sites with their proximal promoter regions, suggesting a direct involvement of ERα in the regulation of their expression (Hah, Danko et al. 2011). It is reasonable to suggest that some of these transcripts are by definition lncRNAs, and could in turn regulate the estrogenic transcription programs, and possibly play a previously under-appreciated role in breast cancer biology. To this end, in the second part of my thesis, I will examine the extent of contribution of lncRNAs to the transcription outcome and cell proliferation programs in estrogen-responsive human breast caner cells, so as to ultimately explore the utility of functionally important lncRNAs in clinical applications to facilitate the treatment of breast cancer.

1.6. Conclusions

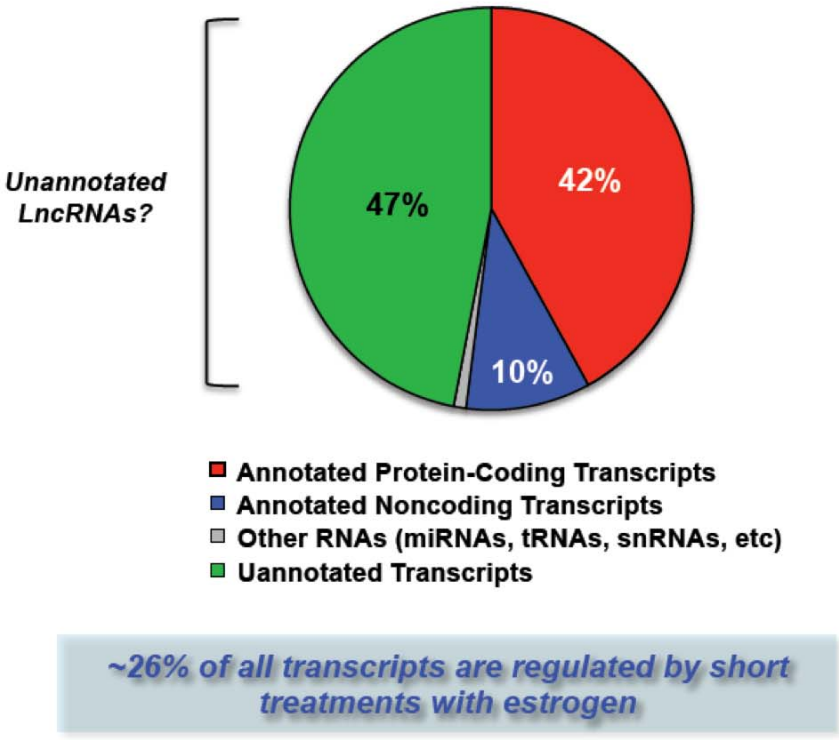


Figure 1.4 Transcriptome profiling in MCF-7 cells using GROseq identified a large number of lncRNA-like transcripts.

GROseq in MCF-7 cells across an E2 treatment time course of 0, 10, 40 and 160 min. captured the location and orientation of all actively transcribing RNA polymerases genome-wide. Transcription units were called *de novo* using a Hidden Markov Model-based statistical model. When the called transcripts are mapped to existing annotations, 10% of them correspond to loci of annotated lncRNA genes; while 47% of them are previously unannotated, some of them are likely unannotated lncRNA genes expressed in basal and E2-stimulated MCF-7 cells. About a quarter of all transcripts are E2-regulated, suggesting a previously underappreciated population of E2-regulated lncRNA genes in MCF-7 cells. Modified from Hah, Danko *et al.* 2011.

In conclusion, our molecular understanding of breast cancer is constantly expanding. It is accepted by now that breast cancer is a complex disease that integrates the activities of a plethora of molecular factors. Therefore, an improved understanding of a more complete spectrum of molecular factors and pathways that are involved in the breast tumorigenic process requires continual investigation, and is the focus of my studies in this dissertation thesis.

REFERENCES

Begg, L., L. H. Kuller, et al. (1987). "Endogenous sex hormone levels and breast cancer risk." Genet Epidemiol 4(4): 233-247.

Benz, C. C., G. K. Scott, et al. (1992). "Estrogen-dependent, tamoxifen-resistant tumorigenic growth of MCF-7 cells transfected with HER2/neu." Breast Cancer Res Treat 24(2): 85-95.

Britton, D. J., I. R. Hutcheson, et al. (2006). "Bidirectional cross talk between ERalpha and EGFR signalling pathways regulates tamoxifen-resistant growth." Breast Cancer Res Treat 96(2): 131-146.

Bunone, G., P. A. Briand, et al. (1996). "Activation of the unliganded estrogen receptor by EGF involves the MAP kinase pathway and direct phosphorylation." EMBO J 15(9): 2174-2183.

Dai, T., E. Rubie, et al. (1995). "Stress-activated protein kinases bind directly to the delta domain of c-Jun in resting cells: implications for repression of c-Jun function." Oncogene 10(5): 849-855.

Dumitrescu, R. G. and I. Cotarla (2005). "Understanding breast cancer risk -- where do we stand in 2005?" J Cell Mol Med 9(1): 208-221.

Foulds, C. E., A. Tsimelzon, et al. (2010). "Research resource: expression profiling reveals unexpected targets and functions of the human steroid receptor RNA activator (SRA) gene." Mol Endocrinol 24(5): 1090-1105.

Gaub, M. P., M. Bellard, et al. (1990). "Activation of the ovalbumin gene by the estrogen receptor involves the fos-jun complex." Cell 63(6): 1267-1276.

Greene, G. L., P. Gilna, et al. (1986). "Sequence and expression of human estrogen receptor complementary DNA." Science 231(4742): 1150-1154.

Gupta, R. A., N. Shah, et al. (2010). "Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis." Nature 464(7291): 1071-1076.

Guyon, J. R., G. J. Narlikar, et al. (1999). "Stable remodeling of tailless nucleosomes by the human SWI-SNF complex." Mol Cell Biol 19(3): 2088-2097.

Hah, N., C. G. Danko, et al. (2011). "A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells." Cell 145(4): 622-634.

Hibi, M., A. Lin, et al. (1993). "Identification of an oncoprotein- and UV-responsive protein kinase that binds and potentiates the c-Jun activation domain." Genes Dev 7(11): 2135-2148.

Hulka, B. S. (1997). "Epidemiologic analysis of breast and gynecologic cancers." Prog Clin Biol Res 396: 17-29.

Hutcheson, I. R., J. M. Knowlden, et al. (2003). "Oestrogen receptor-mediated modulation of the EGFR/MAPK pathway in tamoxifen-resistant MCF-7 cells." Breast Cancer Res Treat 81(1): 81-93.

Ignar-Trowbridge, D. M., M. Pimentel, et al. (1996). "Peptide growth factor cross-talk with the estrogen receptor requires the A/B domain and occurs independently of protein kinase C or estradiol." Endocrinology 137(5): 1735-1744.

Ip, Y. T. and R. J. Davis (1998). "Signal transduction by the c-Jun N-terminal kinase (JNK)--from inflammation to development." Curr Opin Cell Biol 10(2): 205-219.

Jordan, V. C. and B. W. O'Malley (2007). "Selective estrogen-receptor modulators and antihormonal resistance in breast cancer." J Clin Oncol 25(36): 5815-5824.

Kininis, M., B. S. Chen, et al. (2007). "Genomic analyses of transcription factor binding, histone acetylation, and gene expression reveal mechanistically distinct classes of estrogen-regulated promoters." Mol Cell Biol 27(14): 5090-5104.

Kuiper, G. G., E. Enmark, et al. (1996). "Cloning of a novel receptor expressed in rat prostate and ovary." Proc Natl Acad Sci U S A 93(12): 5925-5930.

Kumar, V. and P. Chambon (1988). "The estrogen receptor binds tightly to its responsive element as a ligand-induced homodimer." Cell 55(1): 145-156.

Kurokawa, H., A. E. Lenferink, et al. (2000). "Inhibition of HER2/neu (erbB-2) and mitogen-activated protein kinases enhances tamoxifen action against HER2-overexpressing, tamoxifen-resistant breast cancer cells." Cancer Res 60(20): 5887-5894.

Kushner, P. J., D. A. Agard, et al. (2000). "Estrogen receptor pathways to AP-1." J Steroid Biochem Mol Biol 74(5): 311-317.

Lange, C. A. (2004). "Making sense of cross-talk between steroid hormone receptors and intracellular signaling pathways: who will have the last word?" Mol Endocrinol 18(2): 269-278.

Lannigan, D. A. (2003). "Estrogen receptor phosphorylation." Steroids 68(1): 1-9.

Lanz, R. B., S. S. Chua, et al. (2003). "Steroid receptor RNA activator stimulates proliferation as well as apoptosis in vivo." Mol Cell Biol 23(20): 7163-7176.

Lanz, R. B., N. J. McKenna, et al. (1999). "A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex." Cell 97(1): 17-27.

Leo, C. and J. D. Chen (2000). "The SRC family of nuclear receptor coactivators." Gene 245(1): 1-11.

Lupien, M., C. A. Meyer, et al. (2010). "Growth factor stimulation induces a distinct ER(alpha) cistrome underlying breast cancer endocrine resistance." Genes Dev 24(19): 2219-2227.

Madak-Erdogan, Z., M. Lupien, et al. (2011). "Genomic collaboration of estrogen receptor alpha and extracellular signal-regulated kinase 2 in regulating gene and proliferation programs." Mol Cell Biol 31(1): 226-236.

Malik, S. and R. G. Roeder (2000). "Transcriptional regulation through Mediator-like coactivators in yeast and metazoan cells." Trends Biochem Sci 25(6): 277-283.

Manavathi, B., O. Dey, et al. (2013). "Derailed estrogen signaling and breast cancer: an authentic couple." Endocr Rev 34(1): 1-32.

Mangelsdorf, D. J., C. Thummel, et al. (1995). "The nuclear receptor superfamily: the second decade." Cell 83(6): 835-839.

McGlynn, L. M., T. Kirkegaard, et al. (2009). "Ras/Raf-1/MAPK pathway mediates response to tamoxifen but not chemotherapy in breast cancer patients." Clin Cancer Res 15(4): 1487-1495.

Mourtada-Maarabouni, M., M. R. Pickard, et al. (2009). "GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer." Oncogene 28(2): 195-208.

Nicholson, R. I., I. R. Hutcheson, et al. (2001). "Modulation of epidermal growth factor receptor in endocrine-resistant, oestrogen receptor-positive breast cancer." Endocr Relat Cancer 8(3): 175-182.

Pietras, R. J., J. Arboleda, et al. (1995). "HER-2 tyrosine kinase pathway targets estrogen receptor and promotes hormone-independent growth in human breast cancer cells." Oncogene 10(12): 2435-2446.

Pike, M. C., V. R. Gerkins, et al. (1979). "The hormonal basis of breast cancer." Natl Cancer Inst Monogr(53): 187-193.

Rachez, C. and L. P. Freedman (2001). "Mediator complexes and transcription." Curr Opin Cell Biol 13(3): 274-280.

Robyr, D., A. P. Wolffe, et al. (2000). "Nuclear hormone receptor coregulators in action: diversity for shared tasks." Mol Endocrinol 14(3): 329-347.

Santen, R. J., R. X. Song, et al. (2002). "The role of mitogen-activated protein (MAP) kinase in breast cancer." J Steroid Biochem Mol Biol 80(2): 239-256.

Sivaraman, V. S., H. Wang, et al. (1997). "Hyperexpression of mitogen-activated protein kinase in human breast cancer." J Clin Invest 99(7): 1478-1483.

Smith, C. L. (1998). "Cross-talk between peptide growth factor and estrogen receptor signaling pathways." Biol Reprod 58(3): 627-632.

Umayahara, Y., R. Kawamori, et al. (1994). "Estrogen regulation of the insulin-like growth factor I gene transcription involves an AP-1 enhancer." J Biol Chem 269(23): 16433-16442.

Weisz, A. and R. Rosales (1990). "Identification of an estrogen response element upstream of the human c-fos gene that binds the estrogen receptor and the AP-1 transcription factor." Nucleic Acids Res 18(17): 5097-5106.

Wong, R. J., D. T. Lin, et al. (2002). "Diagnostic and prognostic value of [(18)F]fluorodeoxyglucose positron emission tomography for recurrent head and neck squamous cell carcinoma." J Clin Oncol 20(20): 4199-4208.

CHAPTER 2

Estrogen Regulates JNK1 Genomic Localization to Control Gene Expression and Cell Growth in Breast Cancer Cells*

* This work was published as Sun M, Isaacs GD, Hah N, Heldring N, Fogarty EA, Kraus WL. (2012). "Estrogen regulates JNK1 genomic localization to control gene expression and cell growth in breast cancer cells." Mol Endocrinol. **26**(5):736-47 © the Endocrine Society. Minor changes have been made. Isaacs GD contributed equally to this work (Fig. 2.1A-C, 2.6C, 2.7, 2.8, 2.9B, 2.10-2.12); Heldring N assisted Isaacs GD in some of the experiments; Fogarty EA helped with ChIP-qPCR in Fig. 2.5B; and Hah N performed the proliferation assay in Fig. 2.9A.

2.1. Summary

Steroid hormone and mitogen-activated protein kinase (MAPK) signaling pathways functionally intersect, but the molecular mechanisms of this crosstalk are unclear. Here we demonstrate a functional convergence of the estrogen and c-Jun N-terminal kinase 1 (JNK1) signaling pathways at the genomic level in breast cancer cells. We find that JNK1 binds to many promoters across the genome. Although most of the JNK1 binding sites are constitutive, a subset of are estrogen-regulated (either induced on inhibited). At the estrogen-induced sites, estrogen receptor alpha (ERα) is required for the binding of JNK1 by promoting its recruitment to estrogen response elements (EREs) or other classes of DNA elements through a tethering mechanism, which in some cases involves AP-1. At estrogen-regulated promoters, JNK1 functions as a transcriptional coregulator of ERα in a manner that is dependent on its kinase activity. The convergence of ERα and JNK1 at target gene promoters regulates estrogen-dependent gene expression outcomes, as well as downstream estrogen-dependent cell growth responses. Analysis of existing gene expression profiles from breast cancer biopsies suggests a role for functional interplay between ERα and JNK1 in the progression and clinical outcome of breast cancers.

2.2. Introduction

Diverse signaling pathways regulate a wide variety of cellular processes in mammalian cells, including global transcription programs, to control both physiological and disease states (Kininis and Kraus 2008; Cheung and Kraus 2010). The signaling pathways controlled by estrogens, such as the predominant natural form 17β-estradiol (E2), are good examples of the signal-dependent transcriptional control of cellular outcomes. Estrogens bind to cognate nuclear

estrogen receptor proteins, ERα and ERβ, which function as sequence-specific, DNA-binding transcription factors in the nucleus to directly regulate the transcription of estrogen-responsive genes (Mangelsdorf, Thummel et al. 1995; Warner, Nilsson et al. 1999; Heldring, Pike et al. 2007). ERs bind directly to genomic DNA through ERE sequences (Kumar and Chambon 1988) or indirectly through other transcription factors (e.g., activating protein-1; AP-1) using a tethering mechanism (Gaub, Bellard et al. 1990; Weisz and Rosales 1990; Umayahara, Kawamori et al. 1994; Kushner, Agard et al. 2000; Kushner, Agard et al. 2000), where they recruit a variety of coregulator proteins that mediate transcriptional outcomes (Glass, Rose et al. 1997; Acevedo and Kraus 2004). The genes regulated by estrogens play key roles in the sexual development and fertility of both males and females (Hess 2003; Findlay, Liew et al. 2010), as well as the regulation of metabolic processes in fat, liver and bone tissues (Couse and Korach 1999; Hewitt, Harrell et al. 2005; Li and Shen 2005; Murphy and Korach 2006; Pallottini, Bulzomi et al. 2008). They also play important roles in the aberrant mitogenic and proliferative processes that underlie breast and uterine cancer (Prall, Rogan et al. 1998; Foster, Henley et al. 2001; Sommer and Fuqua 2001). In this regard, the expression of ERα in cells is a well-known prognostic indicator for breast cancers, and a variety of synthetic estrogen antagonists that target ERs are used as therapeutic agents for breast cancers to reverse the mitogenic actions of estrogens (Kuiper, van den Bemd et al. 1999; Johnston 2001; McDonnell, Chang et al. 2001).

In contrast to the nuclear actions of estrogens, growth factors act through cytoplasmic membrane receptors to stimulate intracellular signaling pathways, including mitogen activated protein kinase cascades, which indirectly regulate gene expression through a variety of target transcription factors (Turjanski, Vaque et al. 2007). The MAPK family comprises a conserved set of proteins that are activated by a series of upstream kinases that form a phosphorylation

relay. Activated MAPKs, including the JNKs and the extracellular signal-regulated kinases (ERKs), phosphorylate downstream effectors at serine or threonine residues to control a variety of cellular processes (Davis 2000; Chang and Karin 2001; Johnson and Lapadat 2002; Vlahopoulos and Zoumpourlis 2004). In addition to the direct stimulation of proliferation and cellular survival programs, growth factor signaling pathways functionally interact with estrogen signaling pathways to promote endocrine therapy-resistant growth of cancer cells. Indeed, functional crosstalk between steroid hormone and growth factor/MAPK signaling pathways was demonstrated nearly two decades ago in steroid hormone-dependent cancers (Smith 1998; Lange 2004), but our understanding of how these pathways converge at the genomic level to regulate gene expression remains incomplete.

AP-1, which is a heterodimer of c-Jun and c-Fos or related transcription factors, functions as a terminal downstream target of MAPK pathways (Karin 1995; Hess, Angel et al. 2004). The JNK family of MAPKs was first identified by its ability to specifically phosphorylate c-Jun to modulate the transcriptional activity of AP-1 (Hibi, Lin et al. 1993; Dai, Rubie et al. 1995; Ip and Davis 1998). Subsequent studies have shown that JNK also phosphorylates and regulates the activity of other transcription factors, and non-transcription factors, in response to a variety of extracellular stimuli (Davis 2000; Vlahopoulos and Zoumpourlis 2004). A number of previous studies have described considerable functional interplay between AP-1 and ERs (Webb, Lopez et al. 1995; Kushner, Agard et al. 2000; Teyssier, Belguise et al. 2001; Webb, Nguyen et al. 2003; Qi, Borowicz et al. 2004), including interactions at the level of chromatin through the aforementioned ER tethering pathway. The extent to which JNK family members, such as JNK1, play a role in estrogen signaling pathways and where such potential functional interactions might occur in the cell has not been examined in detail.

The prevailing view in the literature has been that the kinase-mediated phosphorylation events regulating transcriptional outcomes do not occur at the genes that they ultimately regulate. Nonetheless, the terminal kinases of various signaling pathways are found in the nucleus under activating conditions (Edmunds and Mahadevan 2004; Turjanski, Vaque et al. 2007). In addition, gene-specific and genomic analyses in yeast (Pascual-Ahuir, Struhl et al. 2006; Pokholok, Zeitlinger et al. 2006), *Drosophila* (Suganuma, Mushegian et al. 2010), and mammalian cells (Bruna, Nicolas et al. 2003; Edmunds and Mahadevan 2004; Narayanan, Adigun et al. 2005; Vicent, Ballare et al. 2006; Dawson, Bannister et al. 2009; Hu, Xie et al. 2009; Bungard, Fuerth et al. 2010; Madak-Erdogan, Lupien et al. 2011) have shown that a number of signaling kinases bind to the promoters of genes whose expression they regulate. For example, AMPK activates transcription in response to cellular stress through direct association with chromatin and phosphorylation of histone H2B at serine 36 (Bungard, Fuerth et al.). Likewise, cyclin A/cdk2 is recruited to gene promoters where is functions as a progesterone receptor coactivator (Narayanan, Adigun et al. 2005). In addition, ERK2 is recruited to ERα binding sites across the genome where it supports E2-induced gene expression (Madak-Erdogan, Lupien et al.). The extent to which other transcription factors and other kinase families collaborate in the nucleus in a similar manner remains to be determined.

In this study, we characterized the genomic relationships between ERα and JNK1 with respect to their binding to chromatin and subsequent transcriptional outcomes. Our results support a model for the estrogen- and ERα-dependent recruitment of pre-activated JNK1 to the promoters of estrogen target genes. JNK1, in turn, serves as a coregulator of ERα required for efficient estrogen-dependent transcription of these genes, and for downstream cell growth responses. Our study has identified a genomic nexus between the estrogen and JNK1 signaling

pathways, and similar genomic systems are likely to integrate the signaling pathways for other steroid hormones and signal-regulated nuclear kinases in broader cellular processes.

2.3. Results

Activated/phosphorylated JNK1 localizes to the nuclei of MCF-7 cells

To explore the nuclear actions of JNK1 and its potential role in the estrogen signaling pathway, we used the ERα-positive MCF-7 human breast cancer cell line. We first examined the extent to which JNK1 localizes to the nucleus in MCF-7 cells, and whether the natural ERα ligand E2 affects the activation (i.e., phosphorylation) and localization of JNK1. Like other MAPKs, JNK1 is regulated by the phosphorylation of a Thr-Pro-Tyr motif by upstream MAPK kinases (Davis 2000; Turjanski, Vaque et al. 2007). Phosphorylation of JNK1 promotes its translocation into the nucleus and activation of its enzymatic activity (Davis 2000; Turjanski, Vaque et al. 2007).

Fractionation of MCF-7 cells followed by Western blotting revealed that although JNK1 is present in both the cytoplasm and nucleus, only the phosphorylated form of JNK1 is detected in the nucleus (Fig. 2.1A). Treatment of the cells with E2 did not alter the localization of JNK1 or the fraction of phosphorylated JNK1 (Fig. 2.1A). Immunofluorescent staining of the MCF-7 cells confirmed that JNK1 is located in the cytoplasm and nucleus, and that E2 does not alter the localization of JNK1 (Fig. 2.1B). The constitutive JNK1 phosphorylation may be the result of HER-2-dependent MAPK hyperactivation (Oh, Lorant et al. 2001), or it may be related to the elevated kinase activity associated with breast cancer cell lines (Santen, Song et al. 2002). In either case, our results show that activated JNK1 is located in the nuclei of MCF-7 cells.

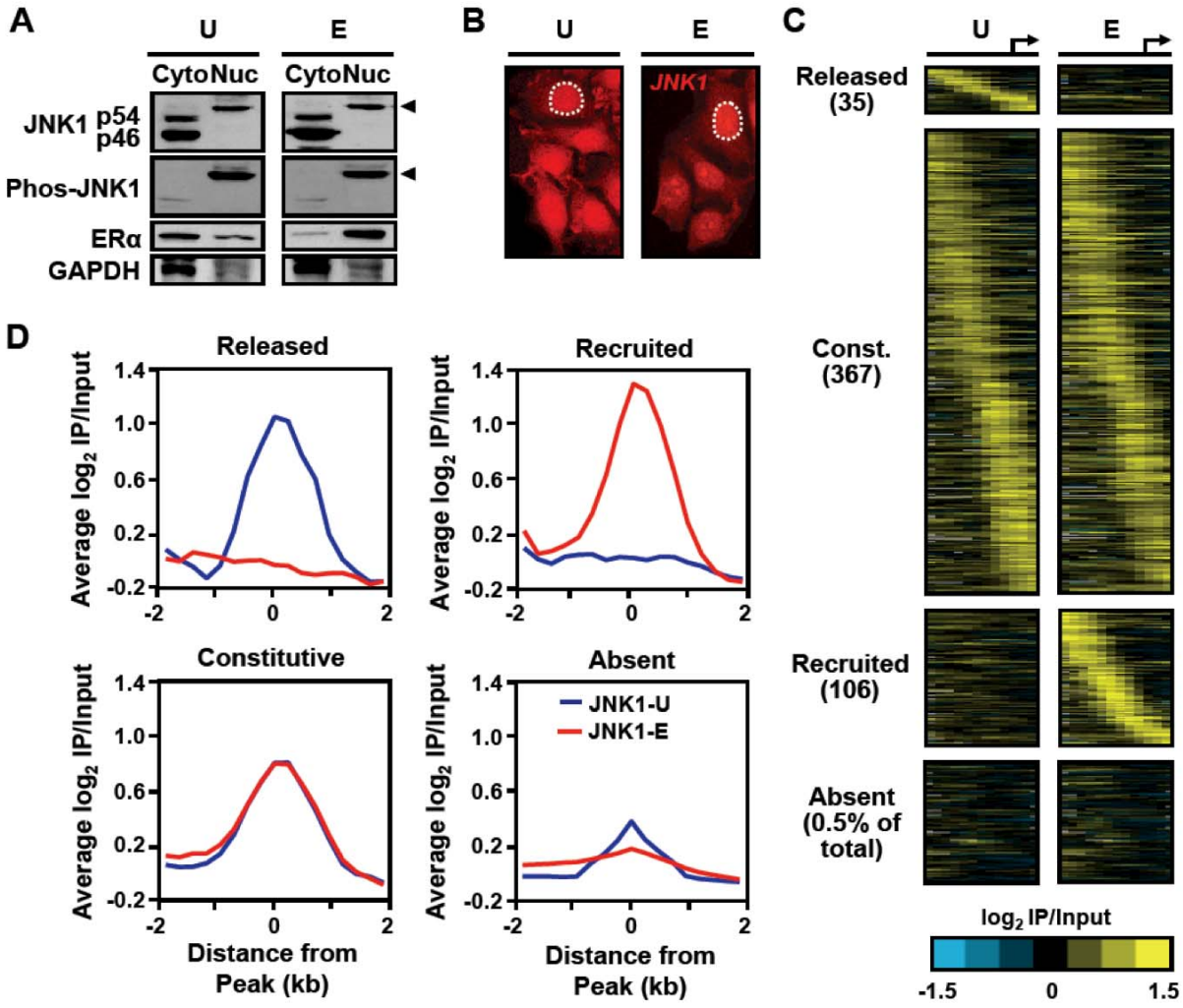
Figure 2.1 Estrogen signaling regulates JNK1 genomic localization program in MCF-7 cells.

(A) MCF-7 cells were treated with vehicle (U) or E2 (E) for 45 min. Cytoplasmic (Cyto) and nuclear (Nuc) extracts were made from both conditions and analyzed by Western blotting for JNK1 and phospho-JNK1. Arrows indicate the migration of phospho-JNK1. ERα was used as a control for cytoplasmic and nuclear fractionation, and GAPDH was used as a loading control.

(B) Immunofluorescent staining of JNK1 (red signal) in MCF-7 cells before and after E2 treatment. One of the nuclei in each panel is denoted by a white dotted line.

(C) Analysis of promoter-proximal JNK1-binding sites in MCF-7 cells before and after E2 treatment by ChIP-chip using Nimblegen promoter arrays containing approximately 19,000 unique promoters tiled from 2200 bp upstream to 500 bp downstream of the TSS (arrow). The data are shown as heatmaps of the JNK1 ChIP-chip $\log_2$ enrichment ratios in both treatment conditions for all promoters with significant binding in either condition, and for 0.5% of JNK1-absent promoters. They are shown in categories of released, constitutive, recruited and absent based on the fold changes of ChIP-chip signals between the E and U conditions. The promoters in each category are aligned by their positions relative to the TSS and ordered from those with the 5'-most JNK1 peak to those with the 3'-most JNK1 peak.

(D) Peak-centered averaging graphs (metagene analyses) of the $\log_2$ enrichment ratios from regions in the categories shown in (C). The probe signals are centered on JNK1 peaks and averaged for all promoters across the region from -2 kb to + 2 kb relative to the JNK1 peak.


A
U
E
CytoNuc
CytoNuc
JNK1 p54
p46
Phos-JNK1
ERα
GAPDH
B
U
E
JNK1
C
U
E
Released
(35)
Const.
(367)
Recruited
(106)
Absent
(0.5% of
total)
log2 IP/Input
-1.5
0
1.5
D
Released
Recruited
Constitutive
Absent
Average log2 IP/Input
1.4
1.0
0.6
0.2
-0.2
-2
0
2
JNK1-U
JNK1-E
Distance from
Peak (kb)

Estrogen signaling regulates the JNK1 genomic localization program in MCF-7 cells

Based on previous studies showing that a number of signaling kinases associate with chromatin, we considered the possibility that nuclear JNK1 may also associate with chromatin in MCF-7 cells. Furthermore, even though E2 treatment did not alter the subcellular distribution of JNK1, we also considered the possibility that E2 might alter the genomic localization of JNK1. To address these questions, we determined the genomic localization of JNK1 at all annotated promoters across the MCF-7 cell genome using chromatin immunoprecipitation coupled with hybridization to RefSeq promoter microarrays (i.e., ChIP-chip) with arrays spanning approximately -2 kb to +0.5 kb relative to the transcription start site. The ChIP DNA was prepared from MCF-7 cells treated with ("E") or without ("U") 100 nM E2.

Using stringent peak definition criteria, we identified more than 500 promoters with a significant peak of JNK1 in either treatment condition (~2% of all promoters on the array) (Fig. 2.1C). Gene-specific ChIP-qPCR of peak and non-peak regions confirmed the ChIP-chip results (Fig. 2.2). Interestingly, E2 treatment caused a redistribution of the JNK1 localization pattern, with 141 peaks changing upon E2 treatment. Thirty-five promoters showed a release of JNK1 upon E2 treatment and 106 promoters showed a recruitment of JNK1 upon E2 treatment, while the majority of JNK1-bound promoters (367) were unaffected by E2 treatment (i.e., constitutively bound by JNK1) (Fig. 2.1C). Averaging of JNK1 peak centered ChIP-chip data across these classes illustrates the distinct patterns of JNK1 promoter binding in response to E2 (Fig. 2.1D). These data show that JNK1 localizes to discrete genomic binding sites and that E2 regulates the JNK1 genomic localization program. Together with the data in Figs. 2.1A and 2.1B, we conclude that E2 treatment alters the occupancy of activated JNK1 on gene promoters in MCF-7 cells without altering the overall nuclear pool of JNK1.

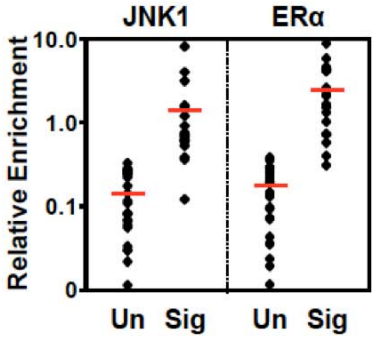


Figure 2.2 Confirmation of JNK1 and ERα peaks from the ChIP-chip analysis by ChIP-qPCR.

ChIP-qPCR was performed to determine JNK1 (left) and ERα (right) occupancy at significantly bound (Significant; Sig) and unbound (Un) regions as defined by the ChIP-chip analyses shown in Figs. 1 and 2. The relative enrichment of each qPCR amplicon tested is shown in the cluster plot, grouped by whether or not the region of the amplicon is bound by JNK1 or ERα based on ChIP-chip peak definition. Red bars represent the average signal in each group. Significant JNK1- and ERα-bound regions from the ChIP-chip analyses show enrichment of factor binding compared to the unbound regions in the ChIP-qPCR experiments, confirming the ChIP-chip analysis.

Gene ontology (GO) analyses showed that JNK1 binding is enriched in the promoters of genes that code for proteins involved in responses to stimuli, signal transduction, and RNA splicing (Table 2.1). These GO categories are driven largely by the group of genes with JNK1 constitutively bound at the promoter. A separate analysis of the genes associated with the JNK1-released and JNK1-recruited promoters also showed an enrichment of genes that code for proteins involved in transcriptional regulation and metabolism of steroids (Table 2.1). Together, these results point to a role for JNK1 in cell growth responses, much like those observed in response to the mitogenic actions of estrogens.

E2-recruited JNK1 colocalizes with ERα at many target promoters

Given the estrogen-dependent alterations in the JNK1 genomic localization program, we tested the possibility that some JNK1 peaks might correspond to sites of ERα binding. To do so, we performed an ERα ChIP-chip analysis using the same array platform that we used for the JNK1 ChIP-chip in Fig. 2.1. Of the 508 significant peaks of JNK1 binding that we identified in the three groups (i.e., released, constitutive, and recruited), ~15% overlapped with an ERα binding site. This increased dramatically to ~40% of the significant peaks of JNK1 when the analysis was limited to the 106 promoters with E2-induced binding of JNK1, suggesting the existence of two populations of E2-induced JNK1 binding sites: ERα-positive (~40%) and ERα-negative (~60%; based on significant peaks).

More broadly, correlation analysis of these 106 promoters showed a striking positive correlation between JNK1 and ERα binding (Fig. 2.3, A and B; Pearson correlation coefficient of 0.68). This analysis, which is more inclusive than the direct peak-to-peak comparison described above because it considers all detectable ChIP-chip signals regardless of significance, suggests

Table 2.1 Gene ontology analysis of JNK1-bound promoters.

Gene set	Ontology[a]	p-value[b]
All JNK1-bound promoters	• Response to stimulus	1.99×10^{-12}
	• Signal transduction	3.02×10^{-10}
	• GPCR signaling pathway	1.64×10^{-8}
	• RNA splicing	3.22×10^{-8}
JNK1-released promoters	• Regulation of transcription	7.67×10^{-4}
JNK1-constitutive promoters	• Response to stimulus	5.44×10^{-11}
	• GPCR signaling pathway	1.07×10^{-8}
	• Signal transduction	1.51×10^{-6}
	• RNA splicing	8.30×10^{-6}
JNK1-recruited promoters	• Regulation of transcription, DNA-dependent	3.65×10^{-7}
	• Response to drugs	1.22×10^{-6}
	• Signal transduction	1.71×10^{-5}
	• Metabolism of androgens and estrogens	6.0×10^{-6}
Five random gene sets[c]	• None	<0.001

[a] Ontologies were obtained using Genecodis for the (1) All JNK1-bound, (2) JNK1-released, (3) JNK1-constitutive, and (4) JNK1-recruited genes. The entire gene list represented on the ChIP-chip array was used as the background reference. GO terms representing less than 5 genes were not considered.

[b] p-values were determined by Genecodis using Chi-square tests. Randomized gene lists of equal size to each gene set analyzed were generated from the genes present on the ChIP-chip array to determine a significance threshold and demonstrate the specificity of ontology assignments.

[c] Five random gene sets were generated using the programming language R from the total number of genes present on the ChIP-chip array. No GO terms were enriched (i.e., all p-values were >0.001) in the random lists using the criteria described above.

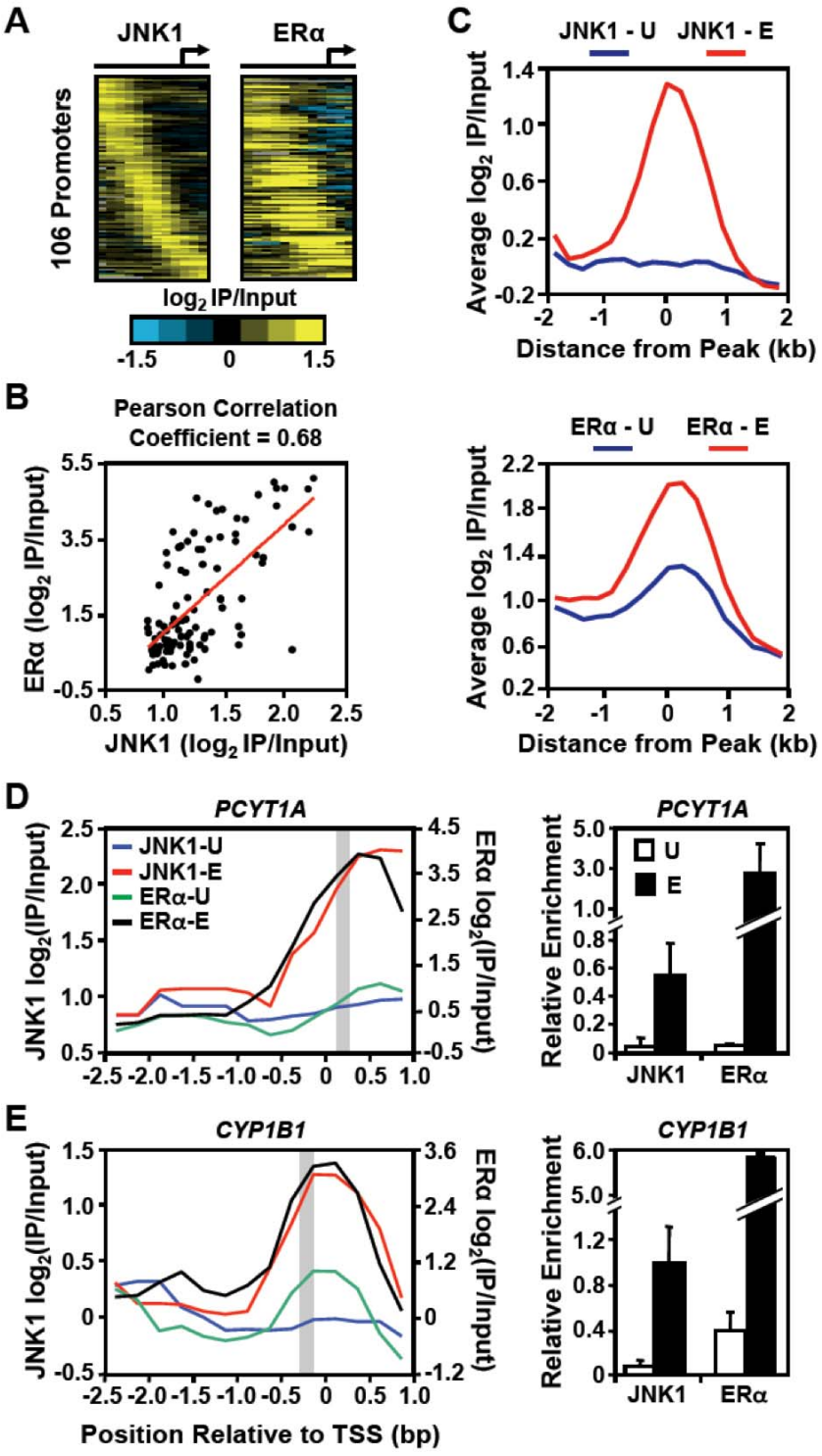
Figure 2.3 JNK1 recruitment correlates with ERα occupancy at target promoters in MCF-7 cells.

(A) Analysis of promoter-proximal JNK1 and ERα binding in MCF-7 cells. JNK1 and ERα ChIP-chip data from E2-treated MCF-7 cells for the 106 "JNK1-recruited" promoters from Fig. 1C are shown as heatmaps of $\log_2$ recruitment ratios aligned and ordered as in Fig. 1C.

(B) Pearson correlation analysis of the JNK1 and ERα ChIP-chip $\log_2$ recruitment ratios.

(C) Metagene analyses of the $\log_2$ enrichment ratios of JNK1 *(top)* and ERα *(bottom)* ChIP-chip data for the "JNK1-recruited" promoters from Fig. 1C before (U, in blue) and after (E, in red) E2 treatment. The probe signals are centered on the JNK1 peaks in both the JNK1 and ERα graphs.

(D and E) ChIP-chip *(left)* and ChIP-qPCR *(right)* analyses of JNK1 and ERα at two "JNK1-recruited" gene promoter regions (*PCYT1A* in panel D and *CYP1B1* in panel E) in MCF-7 cells treated with vehicle (U) or E2 (E). The average JNK1 ChIP-qPCR signals *(right)* of peak regions defined by ChIP-chip (gray boxes in *left* panels) are consistent with the array profiles. For the ChIP-qPCR analyses, each bar represents the mean + SEM, n = 3.


A
JNK1
ERα
106 Promoters
log2 IP/Input
-1.5
0
1.5
C
JNK1 - U
JNK1 - E
Average log2 IP/Input
Distance from Peak (kb)
B
Pearson Correlation
Coefficient = 0.68
ERα (log2 IP/Input)
JNK1 (log2 IP/Input)
ERα - U
ERα - E
Average log2 IP/Input
Distance from Peak (kb)
D
PCYT1A
JNK1-U
JNK1-E
ERα-U
ERα-E
JNK1 log2(IP/Input)
ERα log2(IP/Input)
PCYT1A
U
E
Relative Enrichment
JNK1
ERα
E
CYP1B1
JNK1 log2(IP/Input)
ERα log2(IP/Input)
Position Relative to TSS (bp)
CYP1B1
Relative Enrichment
JNK1
ERα

that there may be even greater colocalization of JNK1 and ERα binding. This result is further illustrated by (1) averaging ERα ChIP-chip signals centered on the E2-recruited JNK1 peaks (Fig. 2.3C) and (2) examining JNK1 and ERα ChIP signals from ChIP-chip and ChIP-qPCR data for the promoters of two target genes (*PCYT1A* and *CYP1B1*; Figs. 2.3D and 2.3E, respectively). These analyses show a pattern of ERα promoter binding that is induced in response to E2, much like JNK1 binding, suggesting that JNK1 and ERα are co-recruited to these genomic binding sites.

To confirm that JNK1 and ERα are present simultaneously at these binding sites, we performed ChIP-reChIP experiments for a selected set of JNK1-bound promoters (JNK1-recruited and constitutive). Whether we first immunoprecipitated ERα (Fig. 2.4A, *top*) or JNK1 (Fig. 2.4B, *top*), we were able to detect both proteins in the reChIP for three JNK1-recruited promoters (*GREB1*, *CYP1B1*, and *PCYT1A*; Fig. 2.4, *bottom*). In contrast, ERα did not show strong binding to the JNK1 constitutively-bound promoter, *ACO2*, and we were thus unable to reChIP JNK1 at this promoter, as expected (Fig. 2.4A, *bottom*). These results demonstrate that, for these promoters, E2-recruited JNK1 co-occupies its binding sites with ERα, indicating that E2 signaling causes the convergence of ERα and JNK1 pathways.

E2-dependent binding of JNK1 to many target promoters is dependent on ERα

Co-recruitment of JNK1 and ERα to specific sites in the genome following E2 treatment suggests a role for ERα in mediating the E2-induced genomic localization of JNK1 at these sites. To explore the dependency of E2-dependent JNK1 recruitment to genomic binding sites on ERα, we used small interfering RNAs (siRNAs) to knock down ERα in MCF-7 cells. A pool of ERα-targeting siRNAs, but not a control pool, effectively depleted ERα in the cells (Fig. 2.5A, *left*,

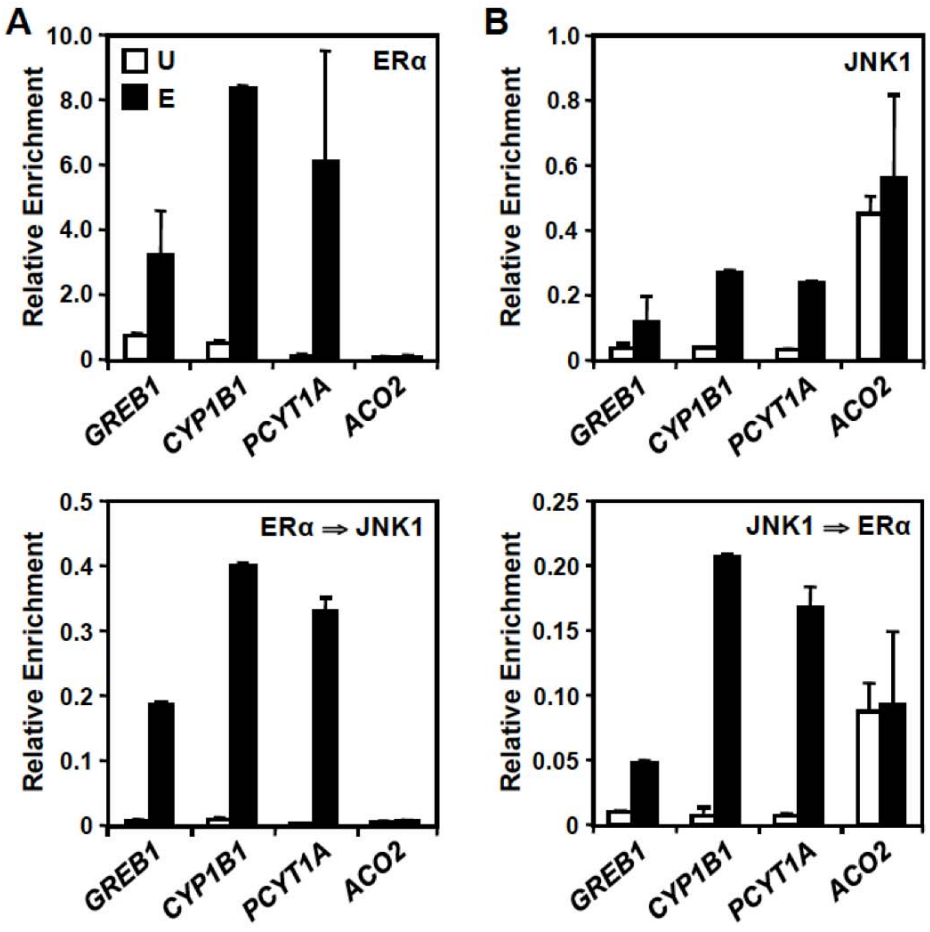


Figure 2.4 JNK1 and ERα colocalize at promoters of "JNK1-recruited" genes.

(A and B) ChIP-qPCR *(top)* and reChIP-qPCR *(bottom)* was performed reciprocally for ERα and JNK1 for three "JNK1-recruited" gene promoters (*GREB1*, *CYP1B1*, and *PCYT1A*) and one "JNK1-consititutive" gene promoter (*ACO2*). The initial ChIP *(top)* was performed using antibodies to ERα (A) or JNK1 (B). The recovered ChIP DNA was then immunopreciptated using antibodies to JNK1 (A) or ERα (B) in a reChIP experiment *(bottom)*. The ChIP and reChIP DNAs were analyzed by RT-qPCR. Each bar represents the mean + SEM, n = 3.

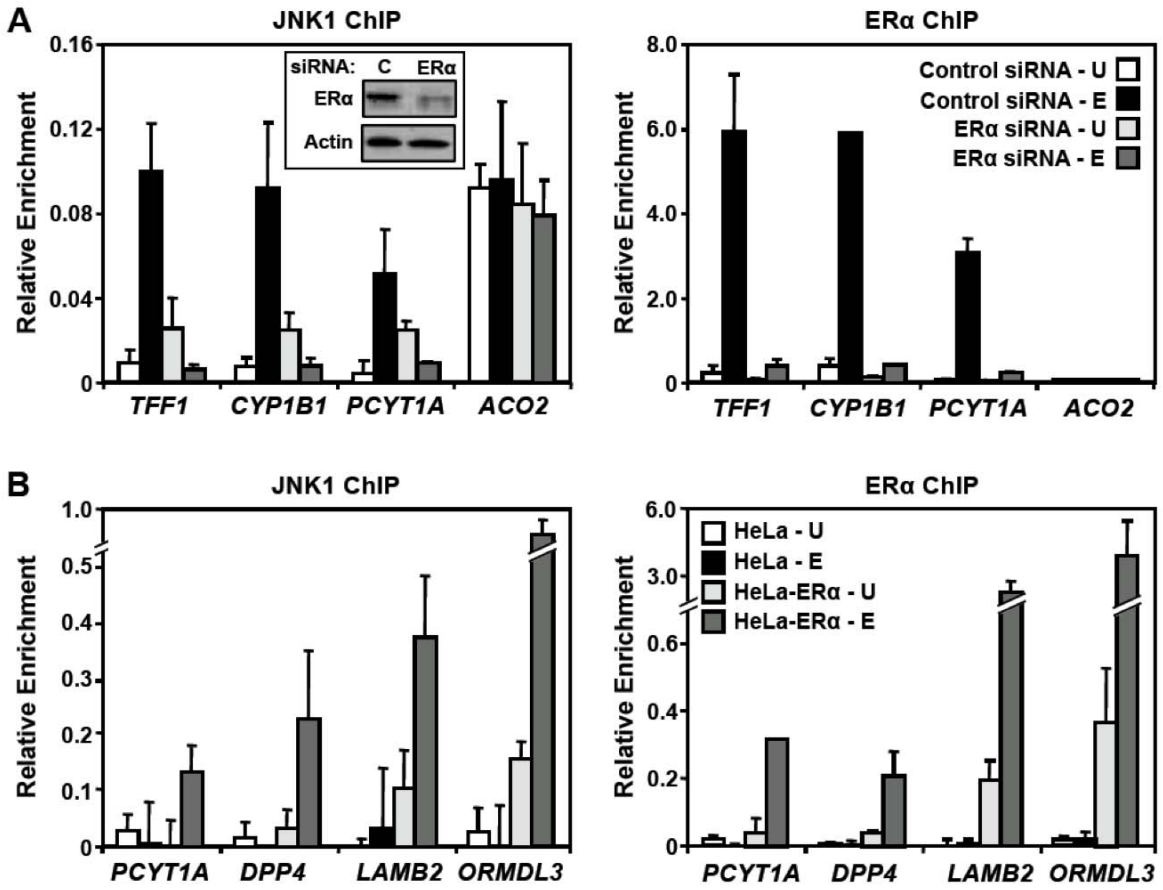


Figure 2.5 ERα binding at target promoters is required for JNK1 recruitment.

(A) MCF-7 cells were transfected with control or ERα siRNAs. Sixty hours after transfection, the cells were treated with vehicle (U) or E2 (E) for 45 min. and collected for Western blotting or ChIP-qPCR analyses. ChIP-qPCR analyses of the E2-dependent JNK1 *(left)* and ERα *(right)* binding to three "JNK1-recruited" gene promoters (*TFF1*, *CYP1B1*, and *PCYT1A*) and one "JNK1-constitutive" gene promoter (*ACO2*) are shown. Each bar represents the mean + SEM, n = 3. *(Inset in left panel)* Western blot analysis of siRNA-mediated ERα knockdown versus control ("C"). β-actin was used as a loading control.

(B) ChIP-qPCR analyses of E2-dependent JNK1 *(left)* and ERα *(right)* binding to four gene promoters (*PCYT2*, *DPP4*, *LAMB2*, and *ORMDL3*) in parental HeLa cells and HeLa cells stably expressing ERα (HeLa-ERα). The cells were treated with vehicle (U) or E2 (E) for 45 min and analyzed by ChIP-qPCR. Each bar represents the mean + SEM, n = 3.

inset). As shown in Fig. 2.5A, it also blocked E2-dependent recruitment of ERα *(left)* and JNK1 *(right)* to target gene promoters (e.g., *TFF1*, *CYP1B1*, *PCYT1A*). In contrast, depletion of ERα did not affect the localization of JNK1 to the constitutive JNK1-bound promoter of *ACO2* (Fig. 2.5A).

To further explore the dependency of E2-dependent JNK1 recruitment on ERα, we used HeLa cells lacking (i.e., HeLa) or stably expressing (i.e., HeLa-ERα) ERα. Using these two cell lines, we examined the promoter localization of ERα and JNK1 in response to E2 treatment. While no promoter localization of ERα or JNK was observed in the ERα-negative HeLa cells, E2-induced ERα and JNK1 recruitment was observed at specific gene promoters in the HeLa-ERα cells (Fig. 2.5B), consistent with the results from the MCF-7 cells. Together, these experiments using two different strategies show definitively that functionally active ERα is required for the E2-dependent recruitment of JNK1 to target promoters where they co-localize.

Transcription factor binding sites are found under JNK1 peaks

As noted above, ERα exhibits two distinct modes of genomic binding: (1) direct binding to genomic DNA through EREs (Heldring, Pike et al. 2007) or indirect binding through other transcription factors (e.g., AP-1) using a tethering mechanism (Kushner, Agard et al. 2000). To determine which mode of binding might direct the co-recruitment of ERα and JNK1, we employed a series of bioinformatic analyses. First, we used MEME and MAST in an unbiased search for DNA sequence motifs enriched under the JNK1 peaks in E2-treated condition. These results yielded a number of high confidence motifs (Fig. 2.6A and Table 2.2). Using TESS to predict the transcription factors that might bind to these sequences, we identified AP-1, as well as other transcription factors not previously associated with ERα-dependent gene regulation

Figure 2.6 Motif analysis of JNK1 peaks.

(A) Unbiased search for DNA sequence motifs enriched under the JNK1-bound regions in the E2-treated condition using MEME. A selection of some of the most significantly enriched motifs are shown as web logos of the position weight matrices. Motif predictions were examined by TESS, as well visual inspection, to determine transcription factors that are most likely to bind to the indicated sequence, as indicated.

(B) Targeted search for AP-1 and ERα (i.e., ERE) binding motifs (1) under all JNK1 peaks in both vehicle- and E2-treated conditions, (2) JNK1 peaks in the E2-treated condition, and (3) under all JNK1-recruited peaks. *(Top)* The position weight matrices used in the targeted search for AP-1 and ERα binding motifs are shown as web logos. The AP-1 position weight matrix is from TRANSFAC., whereas the ERα position weight matrix is based on information from O'Lone et al. (O'Lone, Frith et al. 2004). *(Bottom)* The position weight matrices for the AP-1 and ERα motifs were used with MAST to map the location of the motifs under JNK1 peaks in a directed search. The percent of JNK1 binding sites in each of the indicated groups with an AP-1 or ERα motifs motif is shown.

(C) c-Fos localizes with JNK1 and ERα at target promoters containing an AP-1 binding motif. ChIP-qPCR analyses of JNK1, ERα, and c-Fos binding at JNK1- and ERα-recruited regions before (U) and after estrogen (E) treatment. The *UGT2B15*, *SPTBN4*, *TFF1*, and *GREB1* gene promoters contain at least one predicted AP-1 motif under the JNK1 peak. The *TFF1*, *GREB1*, and *PLAC1* promoters, as well as the *BLK44* distal enhancer (Carroll, Liu et al. 2005), contain at least one ERE under the JNK1 peak. Each bar represents the mean + SEM, n = 3.

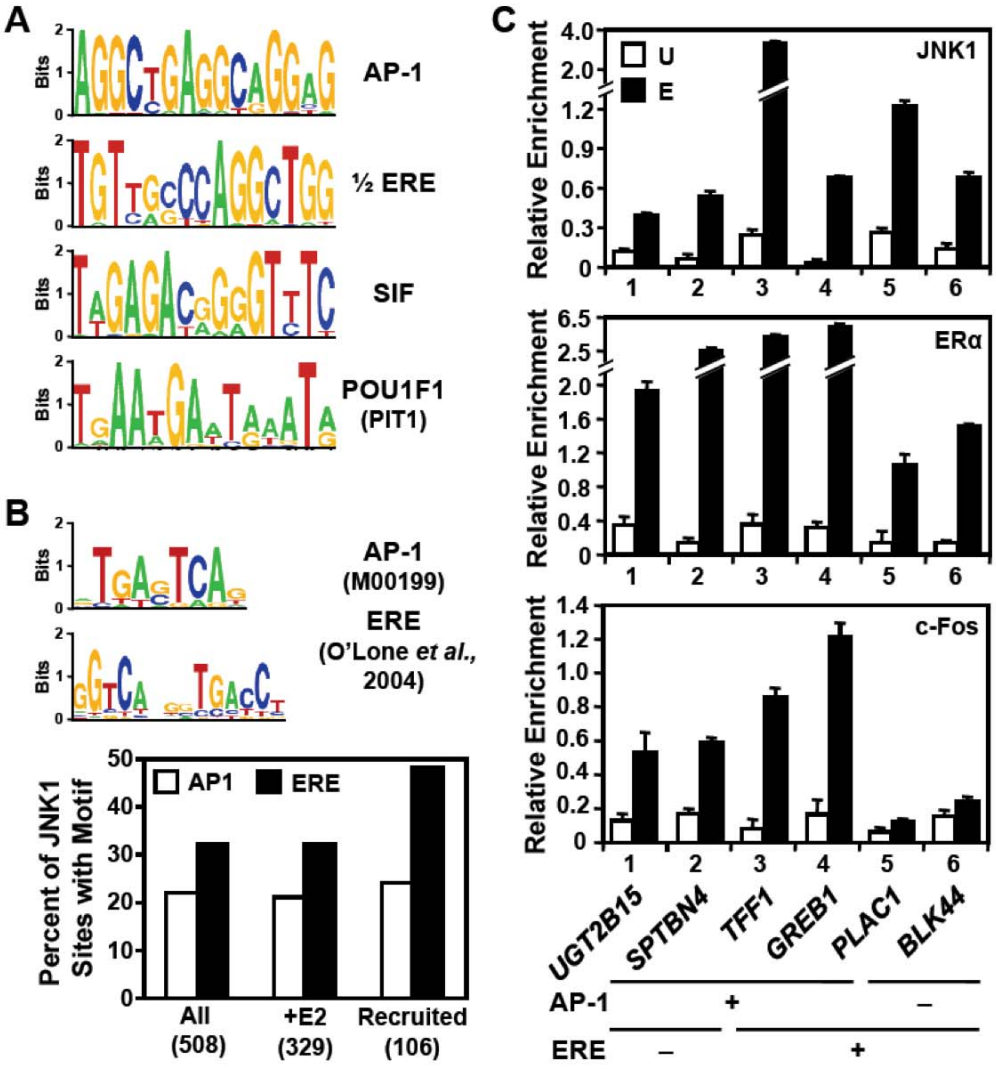

A
Bits
AP-1
½ ERE
SIF
POU1F1
(PIT1)
B
AP-1
(M00199)
ERE
(O'Lone et al.,
2004)
Percent of JNK1
Sites with Motif
AP1
ERE
All
(508)
+E2
(329)
Recruited
(106)
C
Relative Enrichment
U
E
JNK1
ERα
c-Fos
UGT2B15
SPTBN4
TFF1
GREB1
PLAC1
BLK44
AP-1
ERE

Table 2.2 Unbiased motif analysis of JNK1 peaks.

Sequence[a, b]	No. of sites[c] (No. of windows)[d]	p-value[e]	TESS call[f]	ID[g]
JNK1 - E peaks				
GCCTGTAAGTCCCAGC	50 (38)	3.24E-06	Bcd	Q00016
AGGCTGAGGCAGGAG	50 (36)	1.09E-07	AP-1	-
TGTTGCCCAGGCTGG	50 (40)	1.50E-12	½ ERE	R04883
GAGGTTGCAGTGAGC	47 (43)	1.16E-08	–	-
GCCACCACGCCCGGC	50 (36)	1.27E-03	Sp1	R01702
TAGAGACGGGGTTTC	50 (38)	8.02E-20	SIF	R02244
CTTGAGCCCAGGAGT	50 (37)	1.13E-10	LBP-1	I00191
GATCGTGCCACTGCA	38 (35)	7.24E-04	NF-1/L	R01322
CACCTCAGCCTCCCA	19 (18)	8.10E-07	Sp1	R02245
TGAATGAATAAATA	44 (38)	7.70E-82	POU1F1a	R00623
GCACAGCTTCCCTGC	19 (10)	3.96E-28	Sp1	R08166
CTCGAACTCCTGACC	22 (22)	3.71E-15	AP-1	R00368
			½ ERE	I00276
JNK1-recruited				
GCCTGTAATCCCAGC	43 (27)	1.74E-14	Bcd	Q00016
CTGCCTCAGCCTCCC	50 (23)	9.99E-16	Sp1	R02245
TGTTGCCCAGGCTGG	43 (25)	3.06E-10	½ ERE	R04883
GCCACCACGCCCGGC	36 (22)	1.24E-04	Sp1	R01702
GAAACCCCGTCTCTA	32 (23)	3.06E-10	SIF	R02244
GAGGATCACTTGAGC	47 (31)	7.36E-11	IRF-2	R00917
CAGTGAGCTGAGATC	22 (20)	6.12E-10	Zeste	R04948
AGTGCAGTGGC	18 (17)	3.88E-16	Sp1	R01021

[a] Unbiased search for DNA sequence motifs enriched under the JNK1-bound regions using MEME. De novo motif predictions were performed for: (1) all promoters showing JNK1 binding at in the E2-treated condition ("JNK1 - E peaks") or (2) promoters showing JNK1 recruitment upon E2 treatment ("JNK1-recruited"). The gene lists were formulated using the tools on the Galaxy browser (Elnitski, King et al. 2006) so genomic locations from JNK1-bound regions would not be present in the background regions. De novo motif detection was carried out using MEME (Bailey, Williams et al. 2006) on repeat-masked sequences. MAST (Motif Alignment and Search Tool) (Bailey, Williams et al. 2006) was used to scan for the locations of all motif instances within both bound and unbound sequences, using a p-value threshold of 1.5 x 10^{-4} (Bailey, Williams et al. 2006). Fisher's exact tests were used to determine enrichments relative to background, with p-values corrected for multiple testing using the Holm method in R.

[b] The sequences are listed 5' to 3'.

[c] Number of times each motif listed was identified in the promoter regions of the genes in each category (i.e., "JNK1 - E peaks" or "JNK1-recruited").

(continued on the next page)

Table 2.2. ***(continued)***

[d] The number of peak-containing windows in which the sequence is found. Note that two or more sites may be in one peak-containing window.

[e] p-value generated by a Fisher exact test.

[f] The transcription factor binding site associated with the sequence, as called by TESS (Schug 2008). TESS was used to predict the transcription factors that might bind to the enriched sequences from MEME. Position weight matrices for the predicted transcription factors were obtained from the TRANSFAC database (Wingender, Chen et al. 2001). Adjusted matrices for the predicted transcription factors were mapped to the JNK1-bound and JNK1-negative regions with MAST using a 6th order Markov model. Fisher's exact tests were used to determine the enrichments for each motif.

[g] The ID from TESS, IMD (information matrix database) ID for each transcription factor.

(e.g., the POU homeodomain transcription factor POU1F1/PIT1; Fig. 2.6A). Of note, the motifs did not include a canonical ERE, but did include an ERE half-site.

Next, in a directed search, we mapped probability weight matrices (obtained from TRANSFAC) for AP-1 and ERα binding motifs to the JNK1 peaks. Although full EREs were not identified in the unbiased search, we included the ERE probability weight matrix in this directed search based on our results showing a role for ERα in the E2-dependent recruitment of JNK1 to promoters. This analysis yielded high confidence sites for the AP-1 and ERα binding motifs (Fig. 2.6B, *top*). For all groups tested, we observed a greater enrichment of ERα binding motifs (i.e., EREs) than AP-1 binding motifs (Fig. 2.6B, *bottom*). These results of these bioinformatic analyses, together with the ChIP-chip results described above, suggest that the E2-dependent recruitment of JNK1 occurs through ERα using both (1) direct binding to EREs and (2) a tethering mechanism mediated by AP-1 and possibly other DNA-binding transcription factors.

We tested the validity of our bioinformatics analyses using gene-specific ChIP-qPCR assays. For this analysis, we focused on JNK1-recruited promoters (and one ERα enhancer, *BLK44*; (Carroll, Liu et al. 2005)) containing high confidence AP-1 motifs or EREs. Although these JNK1-recruited genomic regions showed E2-induced binding of JNK1 and ERα, as expected, only those with a high confidence AP-1 motif showed binding of c-Fos, a component of the AP-1 heterodimer (Fig. 2.6C). Interestingly, the binding of c-Fos was also stimulated by E2 treatment (Fig. 2.6C, *bottom*), as we have reported previously (Heldring, Isaacs et al. 2011), suggesting E2-induced binding of a complex containing ERα, AP-1, and JNK1 at target promoters. Together, these results support the validity of our bioinformatic analyses by demonstrating the recruitment of JNK1 and c-Fos to regions containing predicted AP-1 sites.

JNK1 acts as an ERα coregulator at E2-responsive genes

Next, we determined the role of JNK1 in the E2-dependent expression of target genes in MCF-7 cells using reverse transcription-qPCR (RT-qPCR) coupled with knockdown of JNK1 or chemical inhibition of JNK1 kinase activity. As shown in Fig. 2.7A, stable expression of a short hairpin RNA (shRNA) into the cells resulted in efficient knockdown of both JNK1 mRNA *(top)* and protein *(bottom)*. As expected, this also resulted in reduced occupancy of JNK1 at the promoters of estrogen-regulated genes (Fig. 2.8). Knockdown of JNK1 (Fig. 2.7B) or chemical inhibition of JNK catalytic activity using SP600125 (Fig. 2.7C) inhibited the E2-stimulated expression of some, but not all, estrogen target genes. Furthermore, the results with JNK1 knockdown were consistent with those from the JNK inhibitor (Fig. 2.7, B and C). Thus, JNK1 protein and its kinase activity are required for full E2-dependent regulation of estrogen target genes in MCF-7 cells, implicating JNK1 as a hormone-dependent transcriptional coregulator of ERα.

JNK1 is required for E2-dependent growth of MCF-7 cells

E2 regulates the transcription of estrogen-responsive genes, including a set of genes involved in cell growth control (Kininis and Kraus 2008). This transcriptional program underlies the potent mitogenic effects of E2 on estrogen-responsive cells, such as MCF-7 cells. To determine the role JNK1 in E2-dependent mitogenic responses, we determined the proliferation of MCF-7 cells in response to E2 treatment with or without JNK1 knockdown. Two different shRNAs targeting JNK1, expressed individually in the cells, reduced the E2-dependent proliferation of MCF-7 cells by about half (Fig. 2.9A). These results suggest that the impaired E2-dependent transcriptional responses that we observed upon JNK1 knockdown (Fig. 2.7B) are

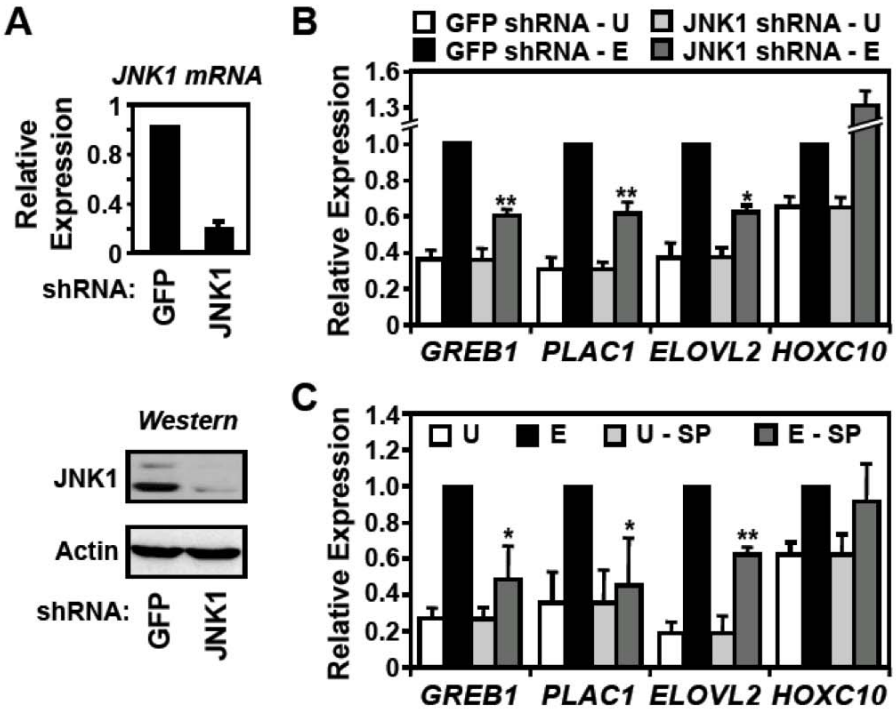


Figure 2.7 JNK1 activity is required for full estrogen-dependent transcriptional responses at estrogen target promoters.

(A) JNK1 was stably knocked down in MCF-7 cells by retroviral-mediated delivery of an shRNA construct followed by drug selection. An shRNA construct targeting GFP was used as a control. *(Top)* Analysis of JNK1 mRNA expression by RT-qPCR. β-actin mRNA was used as an internal control. Each bar represents the mean + SEM, n = 3. *(Bottom)* Analysis of JNK1 protein levels by Western blotting. β-actin was used as a loading control.

(B) Effect of JNK1 knockdown on estrogen-dependent gene expression. The E2-regulated expression of four JNK1-recruited genes (*GREB1*, *PLAC1*, *ELOVL2*, and *HOXC10*) in control (GFP) and JNK1 knockdown MCF-7 cells was monitored by RT-qPCR before (U) and after (E) a 3-hour treatment with E2. Each bar represents the mean + SEM, n = 3. Asterisks represent p-values: <0.05 (∗) or <0.01 (∗∗) (Student's t-test versus corresponding E2 control).

(C) Effect of inhibiting JNK catalytic activity on estrogen-dependent gene expression. The E2-regulated expression of the four JNK1-recruited genes shown in (B) were examined by RT-qPCR in the absence or presence of the JNK inhibitor SP600125 (SP) for 2 hours, before a 3-hour treatment with vehicle (U) or E2 (E). Each bar represents the mean + SEM, n = 3. Asterisks represent p-values: <0.05 (∗) or <0.01 (∗∗) (Student's t-test versus corresponding E2 control).

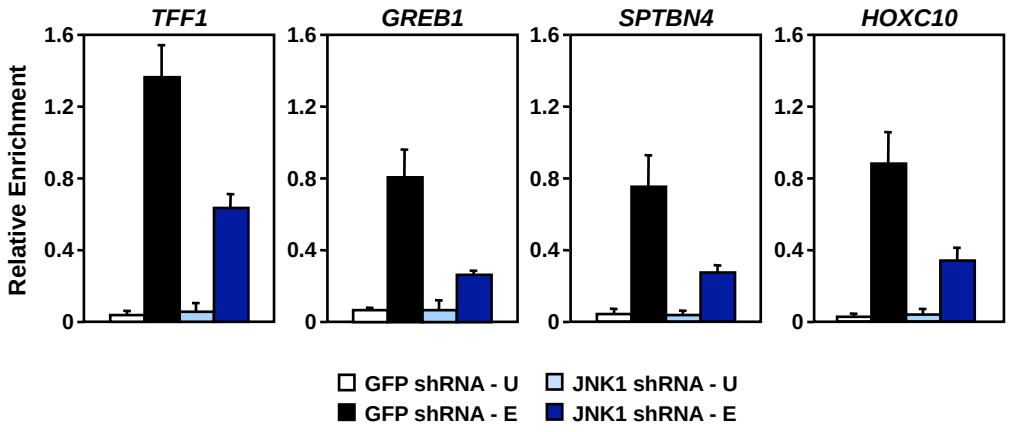


Figure 2.8 Loss of JNK1 occupancy at target gene promoters in JNK1 knockdown cells.

JNK1 was stably knocked down in MCF-7 cells by retroviral-mediated delivery of an shRNA targeting JNK1 into the cells, followed by drug selection. Knockdown of JNK1 mRNA and protein was confirmed as shown in Fig. 6A. ChIP-qPCR using a JNK1 antibody was performed in JNK1 knockdown and control (GFP shRNA) cell lines in the absence (U) and presence (E) of E2 treatment for the four "JNK1-recruited" gene promoters shown (*TFF1*, *GREB1*, *SPTBN4*, and *HOXC10*). The results demonstrate a reduction of E2-induced JNK1 binding at gene promoters upon JNK1 knockdown, as expected.

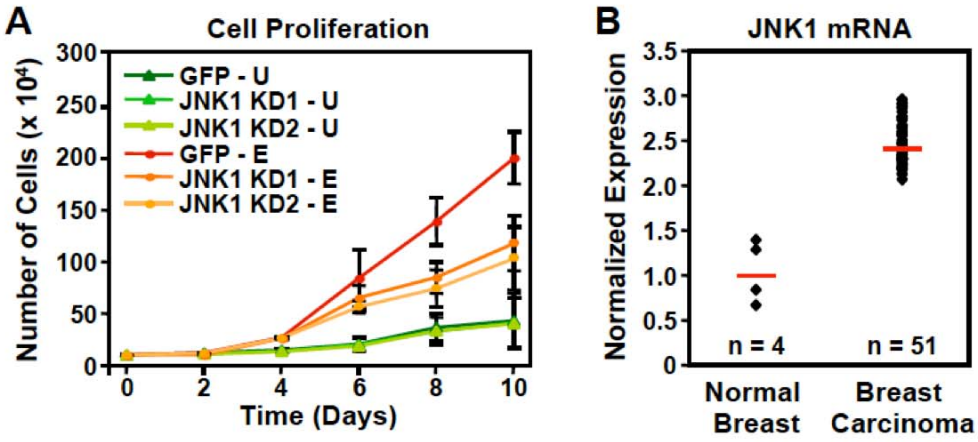


Figure 2.9 JNK1 is required for full estrogen-dependent growth responses in MCF-7 cells.

(A) Analysis of E2-dependent cell growth in control (GFP) and JNK1 knockdown MCF-7 cells grown for 10 days in the absence (U) or presence (E) of E2. Two independent JNK1 shRNA (KD1 and KD2) constructs were used, as shown, and the number of cells in each condition was counted every two days. Each point represents the mean ± SEM, n = 3.

(B) JNK1 expression is elevated in breast carcinomas. Data obtained from gene expression analyses were analyzed using Oncomine. The relative expression of MAPK8 (i.e., JNK1) from 4 normal breast stroma samples and 51 breast tumor samples is shown. The Oncomine-reported p-value was $<3.0 \times 10^{-4}$. The values were normalized to an average expression level of 1 for the normal breast samples. Red lines represent the average signal in each category.

reflected in a corresponding loss of cell growth (Fig. 2.9A).

The observed link between JNK1 and estrogen signaling may have relevance for the growth and clinical outcomes of estrogen-dependent breast cancers. In this regard, note that the expression of JNK1 is upregulated in breast cancers (Fig. 2.9B). In addition, the expression of the JNK1 phosphatase, MPK-1, a negative regulator of JNK1 activity, is reduced in high grade malignant breast cancers (Fig. 2.10). Both of these cancer-related changes would increase the net JNK1 activity and, hence, have the potential to modulate estrogen-dependent growth responses in those cells.

2.4. Discussion

Our genomic and gene-specific analyses of the nuclear functions of JNK1 have revealed new facets of JNK1 biology, including functional interplay with the estrogen signaling pathway. Collectively, our results indicate that (1) activated nuclear JNK1 binds to specific sites in the genome (Fig. 2.1), (2) E2 induces a redistribution of JNK1 binding at promoters (Fig. 2.1), (3) E2-induced binding of JNK1 at many target genes is mediated by the E2-induced formation of promoter-bound complexes containing ERα and, in some cases, tethering proteins such as AP-1 (Figs. 2.2 through 2.6), (4) JNK1 can act as a coregulator of ERα-dependent transcriptional outcomes (Fig. 2.7), and (5) the estrogen and JNK1 signaling pathways collaborate to control the proliferation of breast cancer cells (Fig. 2.9). Thus, the functional interplay between the estrogen and MAPK signaling pathways that has been observed previously is manifested in a molecular crosstalk at the genomic level. These results help to define the molecular mechanisms underlying estrogen signaling in the nucleus and estrogen-dependent gene regulation.

Our results support a model for the estrogen- and ERα-dependent recruitment of pre-

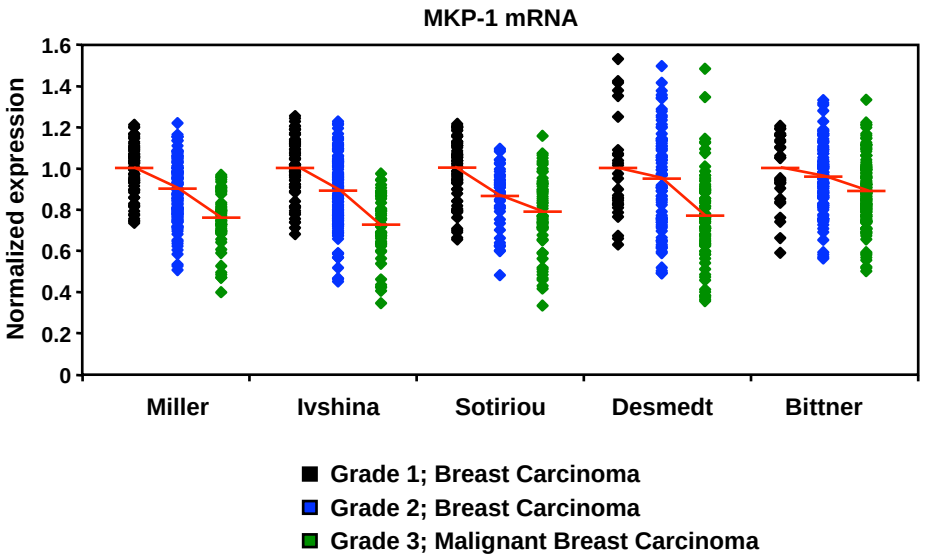
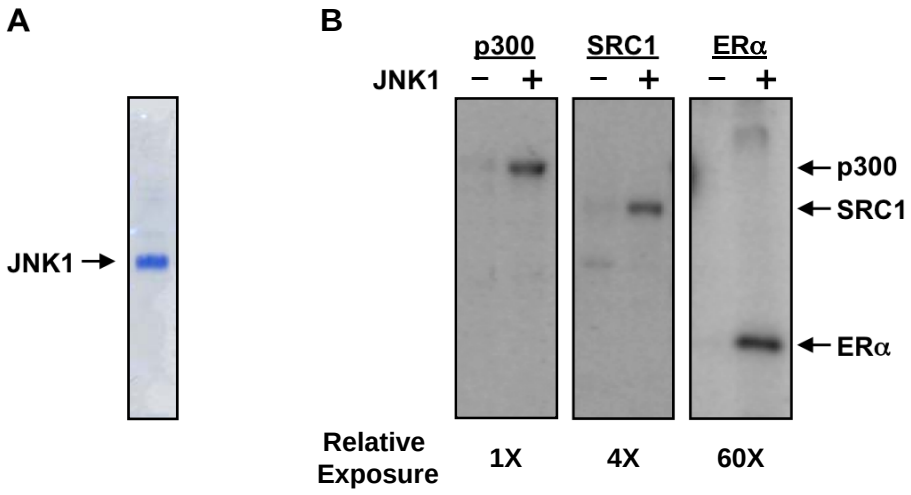


Figure 2.10 Expression of the JNK phosphatase, MKP-1, decreases with breast cancer progression.

Data obtained from the Oncomine database were plotted as shown. The relative expression of MKP-1 across three breast carcinoma grades is shown from five independent studies (indicated by the last names of the first authors). The p-values for negative correlation were <0.001 for all five studies. The values were normalized so that the average expression level for the Grade 1 sample from each study was 1. Red bars represent the average signal in each category.

activated JNK1 from the nuclear compartment (i.e., nucleoplasm or chromatin) to the promoters of estrogen target genes. JNK1, in turn, serves a coregulator function required for efficient estrogen-dependent transcription of these genes. This role of JNK1 in the genomic estrogen signaling pathway is supported by JNK1's kinase activity, which likely targets histones or other proteins in the promoter-assembled transcription complexes, as described for other cellular kinases in the nucleus (Baek 2011). Indeed, two well-characterized ERα coregulators, p300 and SRC1, are strongly phosphorylated by JNK1 in vitro, whereas ERα is only weakly phosphorylated (Fig. 2.11). Nucleosomal histone H3 is also phosphorylated by JNK1, but only when JNK1 is recruited to the nucleosomes by a DNA-bound transcription factor, such as AP-1 (Fig. 2.12). Determining the functional relevance of these and other potential targets of JNK1 will be an important question to address in future studies.

Our results fit well with the growing evidence supporting a role for cellular kinases in the nucleus and across the genome (Baek 2011). A recent study by Madak-Erdogan et al. showed that ERK2 is recruited to ERα binding sites across the genome, where it supports E2-induced gene expression (Madak-Erdogan, Lupien et al.). There are a number of parallels between this study and ours, which suggests that there may be some universal features for cellular kinase actions across the genome, at least in the estrogen signaling pathway. These include the following: (1) estrogen-induced binding and colocalization of the kinase and ERα at specific sites in the genome, (2) a requirement for ERα to drive the recruitment of the kinase to chromatin, (3) interplay with other transcription factors (e.g., AP-1 for JNK1 and CREB1 for ERK2), possibly through a tethering mechanism, (4) a requirement for the kinase to be activated and (5) a role for the kinase in estrogen-dependent gene regulation. The extent to which these features apply to other kinases and other transcription factors, especially other nuclear receptors,



C

JNK Binding Motif ++xxxxTΦxΦ		
1635	RRxxxxTMxM	p300
1013	RxKxxxxTMxV	SRC1
	None	ERα

Figure 2.11 JNK1 phosphorylates ERα coactivators in vitro.

(A) Activated recombinant JNK1 was expressed in bacteria, purified as described previously (Khokhlatchev, Xu et al. 1997), and analyzed by SDS-PAGE with staining using Coomassie Brilliant Blue.

(B) In vitro kinase assays using recombinant JNK1. JNK1 was incubated with ERα, p300, or SRC1 in the presence of ^{32}P-ATP, and phosphorylation of the target proteins was detected by autoradiography. p300 was strongly phosphorylated by JNK1 (note the relative exposure times for each autoradiogram).

(C) p300 and SRC-1, but not ERα, contain a putative JNK interaction domain, similar to the JNK-interacting sequence in c-Jun. Key: + = basic amino acid; x = any amino acid; T = Threonine; Φ = hydrophobic amino acid; R = Arginine; K = Lysine; M = Methionine; V = Valine. Numbers represent the first amino acid position in the motif.

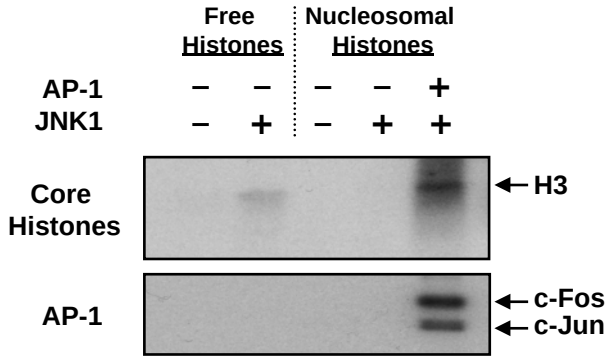


Figure 2.12 JNK1 phosphorylates nucleosomal histone H3 in vitro.

Kinase reactions were performed using recombinant JNK1, as shown in Fig. S3A. JNK1 was incubated with core histones or an equivalent amount of core histones assembled into mononucleosomes containing an AP-1 binding site by salt dialysis. The nucleosome reactions were performed in the absence or presence of recombinant AP-1 (c-Fos/c-Jun heterodimers) to target JNK1 specifically to the nucleosome. ^{32}P phosphorylation of histones, c-Fos, and c-Jun was detected by autoradiography. Nucleosomal H3 is a target of JNK1 enzymatic activity when JNK1 is specifically targeted to the nucleosome.

remains to be determined.

In summary, our studies have identified significant molecular crosstalk between the estrogen and JNK1 signaling pathways that regulates target gene expression and downstream cell growth responses. As noted above, similar genomic systems are likely to integrate the signaling pathways for other steroid hormones and signal-regulated nuclear kinases.

2.5. Methods and Materials

Cell culture and treatments. MCF-7 cells were maintained in MEM with Hank's salts (Sigma; M1018) supplemented with 5% calf serum. Prior to all experimental procedures and treatments, the cells were grown for at least 3 days in phenol red-free MEM Eagle medium with Earle's salts (Sigma; M3024) supplemented with 5% charcoal-dextran-treated calf serum, as described previously (Kininis, Chen et al. 2007). Adherent HeLa and HeLa-ERα cells were maintained in DMEM/F12 (Sigma, D2906) supplemented with 10% charcoal-dextran stripped calf serum, as described previously (Heldring, Isaacs et al. 2011). The cells were treated with control vehicle (ethanol) or E2 (100 nM) for the times specified in the figure legends. For the JNK inhibition experiments, the cells were pre-treated with control vehicle or 20 μM SP600125 (SP; Biomol) for 1 hour before treatment with E2.

Antibodies. The antibodies used were as follows: JNK1 (Santa Cruz, sc-474), pan-JNK (Santa Cruz, sc-7345), phosphorylated pan-JNK (Santa Cruz, sc-6254), ERα (rabbit polyclonal generated in the Kraus lab), and c-Fos (rabbit polyclonal generated in the Kraus lab).

JNK1 subcellular localization. Estrogen-starved MCF-7 cells were treated with ethanol or 100

nM E2 for 45 min. and then fractionated into cytoplasmic and nuclear extracts in the presence of phosphatase inhibitors (5 mM NaF, 1 mM sodium vanadate). The extracts were subjected to immunoblotting using antibodies against JNK1, phosphorylated JNK1, ERα, and GAPDH.

Immunofluorescent staining of cells for JNK1. Estrogen-starved MCF-7 cells were grown on coverslips and treated with ethanol or 100 nM E2 for 45 min. The cells were fixed with 3% formaldehyde, permeabilized with 0.1% Triton X-100, blocked with 5% BSA, and subjected to staining with primary (anti-JNK1) and secondary (fluorescein-conjugated anti-goat IgG) antibodies. The coverslips were then washed 5 times with TBST, mounted on slides using Vectashield (Vector Laboratories; H-1000), and visualized using a Leica Confocal Microscope System.

Chromatin immunoprecipitation (ChIP). ChIP assays for JNK1, ERα, and c-Fos were performed using a ChIP protocol described previously (Kininis, Chen et al. 2007), with minor modifications. The key difference in the protocol was the inclusion of a crosslinking step with 10 mM dimethyl suberimidate•HCl (DMS; Pierce, 20700) for 10 min. at room temperature prior to crosslinking with 1% formaldehyde for 10 min at 37°C. The ChIP DNA was dissolved in water and analyzed by qPCR using a set of gene-specific primers. Each ChIP experiment was conducted with at least three independent chromatin isolates to ensure reproducibility.

ChIP-reChIP. Following the primary ChIP, JNK1- and ERα-precipitated complexes were eluted with 10 mM DTT twice for 20 min at 37°C. Eluates were diluted 20 times in ChIP dilution buffer, incubated with a second antibody at 4°C overnight, followed by the addition of

the protein-A/G-agarose bead mixture. After this secondary ChIP, washing, elution, reversal of the crosslinks and analysis by qPCR were carried out as described for the standard ChIP protocol described above.

ChIP-chip. JNK1- and ERα-precipitated genomic DNA was blunted, amplified by LM-PCR, and labeled as described previously (Krishnakumar, Gamble et al. 2008). The labeled samples were combined and hybridized to human HG18 RefSeq Promoter Arrays (Nimblegen; C4226-00-01), which contain ~19,000 well-characterized RefSeq promoters tiled with 50-mer to 75-mer probes every 100 bp. The tiled regions cover ~2200 bp upstream and ~500 bp downstream of each TSS. The ChIP-chip experiments were performed using three independent ChIP DNA isolates from cells treated with or without E2.

ChIP-chip data analysis. Data processing was done essentially as described previously (Krishnakumar, Gamble et al. 2008) using the statistical programming language R (Harrow, Denoeud et al. 2006), All R scripts are available upon request. The analysis included three components: (1) a moving window analysis using a 1000 bp moving window with 250 bp steps in which both the mean probe $\log_2$ ratio and p-value from a nonparametric Wilcoxon signed-rank test were calculated for each window, (2) definition of regions with significant JNK1 or ERα binding or significant fold changes in JNK1 or ERα binding according to a set of definitions and criteria elaborated in Supplemental Materials, and (3) visual representation of the data by generating TSS-anchored heat maps using Java Treeview (Saldanha 2004). The ChIP-chip data have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE13200

(http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13200).

Knockdown of ERα and JNK1 in MCF-7 cells. Transient RNAi-mediated knockdown of ERα was performed using transient transfection of siRNAs purchased from Dharmacon with an appropriate control siRNA pool. The control or ERα siRNAs pools were transfected into MCF-7 cells as recommended by Dharmacon. Sixty hours after transfection, the cells were treated with E2 and collected for experiments. Stable RNAi-mediated knockdown of JNK1 was performed using retroviral-mediated gene transfer of short hairpin RNA (shRNA) sequences specifically targeting the JNK1 mRNA with the pSUPER.retro system under appropriate drug selection (Oligoengine). The JNK1 target sequences are as follows: 5'-CAGAGAGCTAGTTCTTATGAA-3' and 5'-CCTACAGAGAGCTAGTTCTTA-3'. As a control, we used an shRNA sequence directed against GFP. Knockdown was verified by immunoblotting and RT-qPCR.

Gene-specific expression analyses by RT-qPCR. The expression of endogenous target genes was determined by reverse transcription-quantitative PCR (RT-qPCR), as described previously (Kininis, Chen et al. 2007), with minor modifications. The cDNA products form the reverse transcription reactions were analyzed by qPCR using a set of gene-specific primers. Each experiment was conducted with at least three independent RNA isolates to ensure reproducibility.

Bioinformatic analyses. De novo motif predictions were performed on gene lists that show JNK1 binding at their promoters in the E2-treated condition. These lists were formulated using

the tools on the Galaxy browser (Elnitski, King et al. 2006) so genomic locations from JNK1-bound regions would not be present in the background regions. De novo motif detection was carried out using MEME (Multiple Em for Motif Elicitation) (Bailey, Williams et al. 2006) on repeat-masked sequences. The top 20 motifs in each peak class were retained for further analysis. MAST (Motif Alignment and Search Tool) (Bailey, Williams et al. 2006) was used to scan for the locations of all motif instances within both bound and unbound sequences, using a p-value threshold of 1.5×10^{-4}, as previously reported (Bailey, Williams et al. 2006). Fisher's exact tests were used to determine enrichments relative to background with p-values corrected for multiple testing using the Holm method in R.

TESS (Transcription Element Search Software) (Schug 2008) was used to predict the transcription factors that might bind to the enriched sequences from MEME. Position weight matrices for the predicted transcription factors were obtained from the TRANSFAC database (Wingender, Chen et al. 2001). Adjusted matrices for the predicted transcription factors were mapped to the JNK1-bound and JNK1-negative regions with MAST using a 6th order Markov model. Fisher's exact tests were used to determine the enrichments for each motif, as described above. In addition, promoters were scanned for the presence of EREs in the same manner and the enrichment was calculated.

Gene ontology (GO) analyses. Gene ontologies were obtained using Genecodis (Nogales-Cadenas, Carmona-Saez et al. 2009) for the "All JNK1-bound", "JNK1-released", "JNK1-constitutive", and "JNK1-recruited" gene sets. The entire gene list represented on the ChIP-chip array was used as the background reference. GO terms representing less than 5 genes were not considered. p-values were determined by Genecodis using Chi-square tests. Randomized gene

lists (of equal size to each gene set analyzed) were generated from the genes present on the ChIP-chip array to determine a significance threshold and demonstrate the specificity of ontology assignments. Five random gene sets were generated using the programming language R from the total number of genes present on the ChIP-chip array. No GO terms were enriched (i.e., all p-values were >0.001) in the random lists using the criteria described above.

In vitro kinase assays. Kinase reactions were conducted using recombinant JNK1 purified from *E. coli* using an activated MAP kinase purification system kindly provided by Dr. Melanie Cobb (University of Texas Southwestern Medical Center at Dallas) and described previously (Khokhlatchev, Xu et al. 1997) with two major modifications: (1) the human JNK1 cDNA (JNK1α1) was cloned in to replace the rat JNK2 cDNA sequence and (2) the final cation exchange purification was omitted. Purified activated JNK1 (300 nM) was incubated with various substrates for 30 min. at 30°C in kinase buffer (25 mM HEPES pH 7.5, 10 mM magnesium acetate, 50 μM ATP, 2 μCi γ-^{32}P-ATP). The labeled proteins were resolved using SDS-polyacrylamide gel electrophoresis, the gels were dried on filter paper, and the ^{32}P signal was detected using a phosphorimager system. The substrates tested were as follows: core histones from HeLa cells (140 ng), salt-dialyzed mononucleosomes with an AP-1 binding site containing approximately 140 ng of HeLa cell core histones assembled as described (Jeong, Lauderdale et al. 1991), FLAG-tagged SRC1 (160 nM) purified as described (Thackray and Nordeen 2002), FLAG-tagged ERα (160 nM) and 6xHis-tagged p300 (160 nM) purified as described (Kraus, Manning et al. 1999), and c-Fos/c-Jun dimers (300 nm) purified as described (Ferguson and Goodrich 2001).

Primers for qPCR. The following oligonucleotide primers were used for the ChIP-qPCR and RT-qPCR assays.

<u>ChIP-qPCR:</u>

ACO2 forward	5’- CTTGCACCAGGCCCGTCT-3’
ACO2 reverse	5’- AAGATGTTTTACCCAAGAACAAAT - 3’
Blk44 forward	5’- GGGAAAATATGCAGAAGAAAACGA -3’
Blk44 reverse	5’- CATTTATTCAACACCTCTGATGTCCTA -3’
CYP1B1 forward	5’- CGTGCGGCCTCGATTG -3’
CYP1B1 reverse	5’- AGGTGCCCACGTTTCCATT -3’
GREB1 forward	5’- AGTGTGGCAACTGGGTCATTCTGA -3’
GREB1 reverse	5’- GGTATGATTCATCATTGTCTGCTGCG -3’
PCYT1A forward	5’- CCCTCGCTGTCACTTACCA -3’
PCYT1A reverse	5’- GTTGCAGGTGTGTGCCTATC -3’
PLAC1 forward	5’- TGACAGAACTCATTCACAGGAAG -3’
PLAC1 reverse	5’- GGCAACAGCAAGCACTACAA -3’
SPTBN4 forward	5’- GACTACACGTGCGTGACACC -3’
SPTBN4 reverse	5’- ACGTCCCACACCCTATCGTA -3’
TFF1 forward	5’- ATAACATTTGCCTAAGGAGGCCCG -3’
TFF1 reverse	5’- TCAGCCAAGATGACCTCACCACAT -3’
UGT2B15 forward	5’- TGAACTGTACACACTAATTGGTGAGTCA -3’
UGT2B15 reverse	5’- TCGTGGTGCAAGTAATGTCTTCTAA -3’

RT-qPCR:

ACTB forward	5’- AGCTACGAGCTGCCTGAC -3
ACTB reverse	5’- AAGGTAGTTTCGTGGATGC -3’
GREB1 forward	5’- GCCGTTGACAAGAGGTTC -3’
GREB1 reverse	5’- GGGTTGAGTGGTCAGTTTC -3’
ELOVL2 forward	5’- AGAGGGTGGTTCATGTTGGA -3’
ELOVL2 reverse	5’- CAAGGTGAGGATACCCCTGA -3’
HOXC10 forward	5’- GACACCTCGGATAACGAAGC -3’
HOXC10 reverse	5’- TTTCTCCAATTCCAGCGTCT -3’
PLAC1 forward	5’- CAGTGAGCACAAAGCCACAT -3’
PLAC1 reverse	5’- AACCACAGGAAACAGGAAGC -3’

REFERENCES

Acevedo, M. L. and W. L. Kraus (2004). "Transcriptional activation by nuclear receptors." Essays in biochemistry 40: 73-88.

Baek, S. H. (2011). "When signaling kinases meet histones and histone modifiers in the nucleus." Molecular cell 42(3): 274-284.

Bailey, T. L., N. Williams, et al. (2006). "MEME: discovering and analyzing DNA and protein sequence motifs." Nucleic acids research 34(Web Server issue): W369-373.

Bruna, A., M. Nicolas, et al. (2003). "Glucocorticoid receptor-JNK interaction mediates inhibition of the JNK pathway by glucocorticoids." The EMBO journal 22(22): 6035-6044.

Bungard, D., B. J. Fuerth, et al. (2010). "Signaling kinase AMPK activates stress-promoted transcription via histone H2B phosphorylation." Science 329(5996): 1201-1205.

Carroll, J. S., X. S. Liu, et al. (2005). "Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1." Cell 122(1): 33-43.

Chang, L. and M. Karin (2001). "Mammalian MAP kinase signalling cascades." Nature 410(6824): 37-40.

Cheung, E. and W. L. Kraus (2010). "Genomic analyses of hormone signaling and gene regulation." Annu Rev Physiol 72: 191-218.

Couse, J. F. and K. S. Korach (1999). "Estrogen receptor null mice: what have we learned and where will they lead us?" Endocrine reviews 20(3): 358-417.

Dai, T., E. Rubie, et al. (1995). "Stress-activated protein kinases bind directly to the delta domain of c-Jun in resting cells: implications for repression of c-Jun function." Oncogene 10(5): 849-855.

Davis, R. J. (2000). "Signal transduction by the JNK group of MAP kinases." Cell 103(2): 239-252.

Dawson, M. A., A. J. Bannister, et al. (2009). "JAK2 phosphorylates histone H3Y41 and excludes HP1alpha from chromatin." Nature 461(7265): 819-822.

Edmunds, J. W. and L. C. Mahadevan (2004). "MAP kinases as structural adaptors and enzymatic activators in transcription complexes." Journal of cell science 117(Pt 17): 3715-3723.

Elnitski, L., D. King, et al. (2006). "Computational prediction of cis-regulatory modules from multispecies alignments using Galaxy, Table Browser, and GALA." Methods in molecular biology 338: 91-103.

Ferguson, H. A. and J. A. Goodrich (2001). "Expression and purification of recombinant human c-Fos/c-Jun that is highly active in DNA binding and transcriptional activation in vitro." Nucleic Acids Res 29(20): E98.

Findlay, J. K., S. H. Liew, et al. (2010). "Estrogen signaling in the regulation of female reproductive functions." Handbook of experimental pharmacology(198): 29-35.

Foster, J. S., D. C. Henley, et al. (2001). "Estrogens and cell-cycle regulation in breast cancer." Trends in endocrinology and metabolism: TEM 12(7): 320-327.

Gaub, M. P., M. Bellard, et al. (1990). "Activation of the ovalbumin gene by the estrogen receptor involves the fos-jun complex." Cell 63(6): 1267-1276.

Glass, C. K., D. W. Rose, et al. (1997). "Nuclear receptor coactivators." Current opinion in cell biology 9(2): 222-232.

Harrow, J., F. Denoeud, et al. (2006). "GENCODE: producing a reference annotation for ENCODE." Genome Biol 7 Suppl 1: S4 1-9.

Heldring, N., G. D. Isaacs, et al. (2011). "Multiple sequence-specific DNA-binding proteins mediate estrogen receptor signaling through a tethering pathway." Molecular endocrinology 25(4): 564-574.

Heldring, N., A. Pike, et al. (2007). "Estrogen receptors: how do they signal and what are their targets." Physiological reviews 87(3): 905-931.

Hess, J., P. Angel, et al. (2004). "AP-1 subunits: quarrel and harmony among siblings." Journal of cell science 117(Pt 25): 5965-5973.

Hess, R. A. (2003). "Estrogen in the adult male reproductive tract: a review." Reproductive biology and endocrinology : RB&E 1: 52.

Hewitt, S. C., J. C. Harrell, et al. (2005). "Lessons in estrogen biology from knockout and transgenic animals." Annual review of physiology 67: 285-308.

Hibi, M., A. Lin, et al. (1993). "Identification of an oncoprotein- and UV-responsive protein kinase that binds and potentiates the c-Jun activation domain." Genes Dev 7(11): 2135-2148.

Hu, S., Z. Xie, et al. (2009). "Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling." Cell 139(3): 610-622.

Ip, Y. T. and R. J. Davis (1998). "Signal transduction by the c-Jun N-terminal kinase (JNK)--from inflammation to development." Current opinion in cell biology 10(2): 205-219.

Jeong, S. W., J. D. Lauderdale, et al. (1991). "Chromatin assembly on plasmid DNA in vitro. Apparent spreading of nucleosome alignment from one region of pBR327 by histone H5." J Mol Biol 222(4): 1131-1147.

Johnson, G. L. and R. Lapadat (2002). "Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases." Science 298(5600): 1911-1912.

Johnston, S. R. (2001). "Endocrine manipulation in advanced breast cancer: recent advances with SERM therapies." Clinical cancer research : an official journal of the American Association for Cancer Research 7(12 Suppl): 4376s-4387s; discussion 4411s-4412s.

Karin, M. (1995). "The regulation of AP-1 activity by mitogen-activated protein kinases." The Journal of biological chemistry 270(28): 16483-16486.

Khokhlatchev, A., S. Xu, et al. (1997). "Reconstitution of mitogen-activated protein kinase phosphorylation cascades in bacteria. Efficient synthesis of active protein kinases." J Biol Chem 272(17): 11057-11062.

Kininis, M., B. S. Chen, et al. (2007). "Genomic analyses of transcription factor binding, histone acetylation, and gene expression reveal mechanistically distinct classes of estrogen-regulated promoters." Mol Cell Biol 27(14): 5090-5104.

Kininis, M. and W. L. Kraus (2008). "A global view of transcriptional regulation by nuclear receptors: gene expression, factor localization, and DNA sequence analysis." Nuclear receptor signaling 6: e005.

Kraus, W. L., E. T. Manning, et al. (1999). "Biochemical analysis of distinct activation functions in p300 that enhance transcription initiation with chromatin templates." Mol Cell Biol 19(12): 8123-8135.

Krishnakumar, R., M. J. Gamble, et al. (2008). "Reciprocal binding of PARP-1 and histone H1 at promoters specifies transcriptional outcomes." Science 319(5864): 819-821.

Kuiper, G. G., G. J. van den Bemd, et al. (1999). "Estrogen receptor and the SERM concept." Journal of endocrinological investigation 22(8): 594-603.

Kumar, V. and P. Chambon (1988). "The estrogen receptor binds tightly to its responsive element as a ligand-induced homodimer." Cell 55(1): 145-156.

Kushner, P. J., D. Agard, et al. (2000). "Oestrogen receptor function at classical and alternative response elements." Novartis Found Symp 230: 20-26; discussion 27-40.

Kushner, P. J., D. A. Agard, et al. (2000). "Estrogen receptor pathways to AP-1." The Journal of steroid biochemistry and molecular biology 74(5): 311-317.

Lange, C. A. (2004). "Making sense of cross-talk between steroid hormone receptors and intracellular signaling pathways: who will have the last word?" Mol Endocrinol 18(2): 269-278.

Li, R. and Y. Shen (2005). "Estrogen and brain: synthesis, function and diseases." Frontiers in bioscience : a journal and virtual library 10: 257-267.

Madak-Erdogan, Z., M. Lupien, et al. (2011). "Genomic collaboration of estrogen receptor alpha and extracellular signal-regulated kinase 2 in regulating gene and proliferation programs." Mol Cell Biol 31(1): 226-236.

Mangelsdorf, D. J., C. Thummel, et al. (1995). "The nuclear receptor superfamily: the second decade." Cell 83(6): 835-839.

McDonnell, D. P., C. Y. Chang, et al. (2001). "Capitalizing on the complexities of estrogen receptor pharmacology in the quest for the perfect SERM." Annals of the New York Academy of Sciences 949: 16-35.

Murphy, E. and K. S. Korach (2006). "Actions of estrogen and estrogen receptors in nonclassical target tissues." Ernst Schering Foundation symposium proceedings(1): 13-24.

Narayanan, R., A. A. Adigun, et al. (2005). "Cyclin-dependent kinase activity is required for progesterone receptor function: novel role for cyclin A/Cdk2 as a progesterone receptor coactivator." Molecular and cellular biology 25(1): 264-277.

Nogales-Cadenas, R., P. Carmona-Saez, et al. (2009). "GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information." Nucleic Acids Res 37(Web Server issue): W317-322.

O'Lone, R., M. C. Frith, et al. (2004). "Genomic targets of nuclear estrogen receptors." Molecular endocrinology 18(8): 1859-1875.

Oh, A. S., L. A. Lorant, et al. (2001). "Hyperactivation of MAPK induces loss of ERalpha expression in breast cancer cells." Molecular endocrinology 15(8): 1344-1359.

Pallottini, V., P. Bulzomi, et al. (2008). "Estrogen regulation of adipose tissue functions: involvement of estrogen receptor isoforms." Infectious disorders drug targets 8(1): 52-60.

Pascual-Ahuir, A., K. Struhl, et al. (2006). "Genome-wide location analysis of the stress-activated MAP kinase Hog1 in yeast." Methods 40(3): 272-278.

Pokholok, D. K., J. Zeitlinger, et al. (2006). "Activated signal transduction kinases frequently occupy target genes." Science 313(5786): 533-536.

Prall, O. W., E. M. Rogan, et al. (1998). "Estrogen regulation of cell cycle progression in breast cancer cells." The Journal of steroid biochemistry and molecular biology 65(1-6): 169-174.

Qi, X., S. Borowicz, et al. (2004). "Estrogen receptor inhibits c-Jun-dependent stress-induced cell death by binding and modifying c-Jun activity in human breast cancer cells." The Journal of biological chemistry 279(8): 6769-6777.

Saldanha, A. J. (2004). "Java Treeview--extensible visualization of microarray data." Bioinformatics 20(17): 3246-3248.

Santen, R. J., R. X. Song, et al. (2002). "The role of mitogen-activated protein (MAP) kinase in breast cancer." The Journal of steroid biochemistry and molecular biology 80(2): 239-256.

Schug, J. (2008). "Using TESS to predict transcription factor binding sites in DNA sequence." Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.] Chapter 2: Unit 2 6.

Smith, C. L. (1998). "Cross-talk between peptide growth factor and estrogen receptor signaling pathways." Biol Reprod 58(3): 627-632.

Sommer, S. and S. A. Fuqua (2001). "Estrogen receptor and breast cancer." Seminars in cancer biology 11(5): 339-352.

Suganuma, T., A. Mushegian, et al. (2010). "The ATAC acetyltransferase complex coordinates MAP kinases to regulate JNK target genes." Cell 142(5): 726-736.

Teyssier, C., K. Belguise, et al. (2001). "Characterization of the physical interaction between estrogen receptor alpha and JUN proteins." The Journal of biological chemistry 276(39): 36361-36369.

Thackray, V. G. and S. K. Nordeen (2002). "High-yield purification of functional, full-length steroid receptor coactivator 1 expressed in insect cells." Biotechniques 32(2): 260, 262-263.

Turjanski, A. G., J. P. Vaque, et al. (2007). "MAP kinases and the control of nuclear events." Oncogene 26(22): 3240-3253.

Umayahara, Y., R. Kawamori, et al. (1994). "Estrogen regulation of the insulin-like growth factor I gene transcription involves an AP-1 enhancer." J Biol Chem 269(23): 16433-16442.

Vicent, G. P., C. Ballare, et al. (2006). "Induction of progesterone target genes requires activation of Erk and Msk kinases and phosphorylation of histone H3." Molecular cell 24(3): 367-381.

Vlahopoulos, S. and V. C. Zoumpourlis (2004). "JNK: a key modulator of intracellular signaling." Biochemistry. Biokhimiia 69(8): 844-854.

Warner, M., S. Nilsson, et al. (1999). "The estrogen receptor family." Current opinion in obstetrics & gynecology 11(3): 249-254.

Webb, P., G. N. Lopez, et al. (1995). "Tamoxifen activation of the estrogen receptor/AP-1 pathway: potential origin for the cell-specific estrogen-like effects of antiestrogens." Molecular endocrinology 9(4): 443-456.

Webb, P., P. Nguyen, et al. (2003). "Differential SERM effects on corepressor binding dictate ERalpha activity in vivo." The Journal of biological chemistry 278(9): 6912-6920.

Weisz, A. and R. Rosales (1990). "Identification of an estrogen response element upstream of the human c-fos gene that binds the estrogen receptor and the AP-1 transcription factor." Nucleic Acids Res 18(17): 5097-5106.

Wingender, E., X. Chen, et al. (2001). "The TRANSFAC system on gene expression regulation." Nucleic acids research 29(1): 281-283.

CHAPTER 3

Integrative Annotation and Functional Characterization of Long Noncoding RNAs in Human Breast Cancer Cells*

* This research was done with contributions from Gadad SS.. He assisted with the generation of RNA-seq libraries, carried out siRNA-mediated knockdown of the two candidate lncRNAs, and performed the corresponding cell proliferation assay.

3.1. Summary

Long noncoding RNAs (lncRNAs) that are emerging as key regulators in a wide variety of cellular processes, including breast cancer, but the extent of their implications is just beginning to be elucidated. Therefore, I developed a computational approach that integrates information from multiple high-throughput sequencing datasets, and derived at a comprehensive catalog of 1888 expressed lncRNA genes in the estrogen-responsive MCF-7 breast cancer cells. More than 40% of them are first annotated in this study. Close examination of these lncRNAs revealed many interesting features, including their distribution along the genome, their subcellular localization and the associated stability, as well as their chromatin signatures. More than a quarter of these lncRNAs are regulated by estrogen, either transcriptionally or post-transcriptionally. In addition, cell type-specific expression of lncRNAs predicts the instrinsic molecular subtypes of breast cancer cells, suggesting its potential utility as prognostic marker. Lastly, by selecting lncRNAs with elevated expression in breast tumors, and whose differential expression across a whole spectrum of tissues and cell types correlates with important cell viability genes, we identified a number of lncRNAs that are required for the normal growth of MCF-7 cells. Collectively, these results expanded our knowledge in the implications of lncRNAs in breast cancer biology, and suggested new targets for therapeutic interventions.

3.2. Introduction

Genome-wide transcriptome analyses conducted over the past decade, including recent studies by the ENCODE (Encyclopedia of DNA Elements) Consortium, have revealed that mammalian genomes are pervasively, but not indiscriminately, transcribed, giving rise to a wide variety of coding and non-coding RNA (ncRNA) transcripts (Okazaki, Furuno et al. 2002;

Birney, Stamatoyannopoulos et al. 2007; Djebali, Davis et al. 2012). The cellular repertoire of non-coding RNAs consists of small housekeeping RNAs such as ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs), microRNAs, and long noncoding RNAs (lncRNAs) including antisense RNAs (asRNAs) and enhancer RNAs (eRNAs). The functions of many of these ncRNAs are poorly understood, but interests in uncovering their biological functions and molecular mechanisms of action are intense. In this review, we focus on lncRNAs, presenting the most current information on their discovery, annotation, molecular actions, and biological functions, especially as they relate to hormonal signaling systems.

3.2.1. Defining lncRNAs

lncRNAs, defined as non-protein coding RNA transcripts longer than 200 nucleotides (nt), are emerging as key regulators of diverse cellular processes. To date, a limited, but fast growing number of lncRNAs have been functionally characterized through gene-specific studies. To further expand our understanding of lncRNAs, rapid advancements in genomic methods and analyses have spearheaded recent efforts in the large-scale identification of lncRNAs across multiple biological systems. Nevertheless, accurate identification demands a clear definition and sufficient knowledge of the features of lncRNAs.

3.2.1.1. An Evolving Definition of LncRNAs

The definition of lncRNAs continues to evolve. A universal classification scheme does not exist, and there have been various synonyms describing either very similar or slightly differing lncRNA-like molecules, adding to the confusion. The basic features are represented in the name “lncRNA”: they are obligate “non-coding” RNAs (Fig. 3.2.1, top) and are relatively

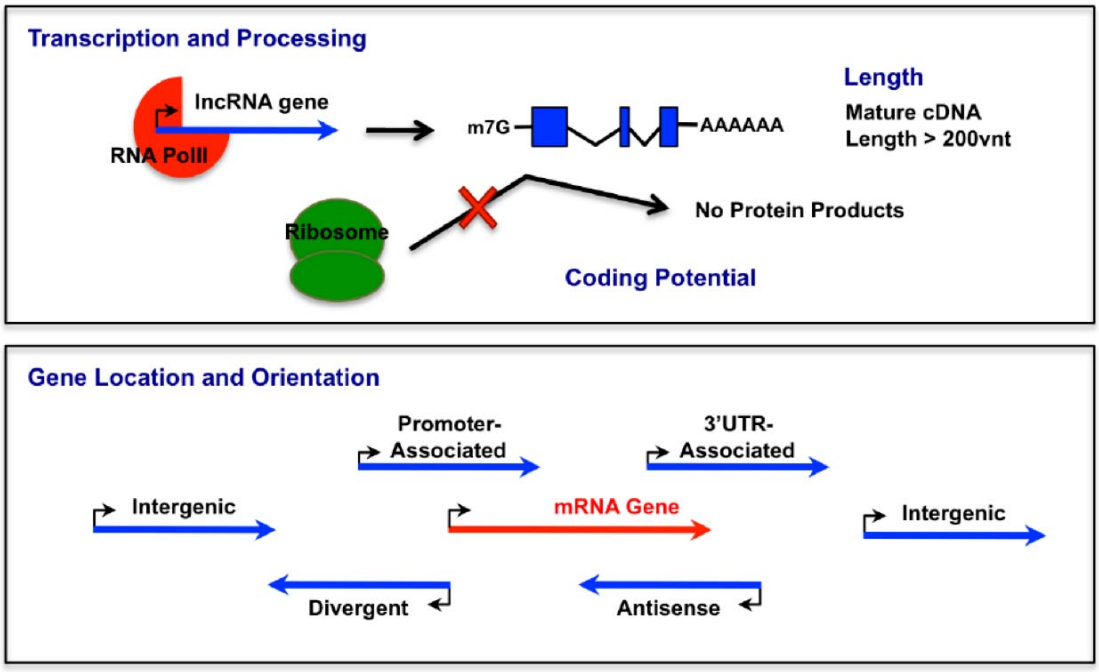


Figure 3.2.1 Definition and key features of lncRNAs.

Graphical representation of the definition and key features of lncRNAs, including the length of mature cDNA > 200 nt, the lack of coding capability, being mostly transcribed by RNA Pol II, spliced, 5'-capped and 3'-polyadenylated, and displaying a variety of gene location and orientation relative to a protein-coding gene.

"long" (>200 nt) (Fig. 3.2.1, top). Some definitions include an "intergenic" feature (i.e., "lincRNAs"; by definition, they do not overlap in any way with annotated protein coding transcription units) (Guttman, Amit et al. 2009; Khalil, Guttman et al. 2009; Guttman, Garber et al. 2010; Cabili, Trapnell et al. 2011).

Length. While the current pool of known lncRNAs display a wide range of transcript length, the lower bound for "long" is somewhat arbitrarily set to be greater than 200 nt in an attempt to facilitate distinction from most other well-characterized groups of small ncRNA trancripts, such as rRNAs, tRNAs, small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), and microRNAs. This length was chosen for practical considerations as well, since this threshold allows empirical separation of RNAs in common experimental procedures. The 200 nt cutoff, however, does not make clear biological distinctions, creating potential grey areas in our understanding.

Coding potential. The absolute requirement for "noncoding" also invites controversy. Recent results from ribosome profiling studies have shown that some previously classified lncRNAs are detected in association with ribosomes. To some, such results would preclude these transcripts from consideration as lncRNAs, since they are likely to be translated and, therefore, coding. It raises the question of whether protein-coding capacity completely prevents an RNA transcript from being defined as a lncRNA. Some RNAs may function both as a structural RNA and a coding RNA (see the steroid receptor RNA activator below; SRA). The SRA gene produces a functional ncRNA, as well as a protein-coding variant(Lanz, McKenna et al. 1999; Chooniedass-Kothari, Emberley et al. 2004; Chooniedass-Kothari, Hamedani et al. 2006). Although the SRA gene fails to satisfy the definition of "noncoding", we can still argue this is a case where the functional lncRNA isoform can be unambiguously distinguished from its

coding isoforms. Nevertheless, cases of functional lncRNA-like transcripts with a detectable level of coding potential have been described. Their protein-coding capability can be reminiscent of degenerate evolution as the noncoding function prevails, or the RNA transcript can simply be bifunctional, being both a protein-coding RNA and a functional lncRNA. Instead of excluding any lncRNA-like transcripts due to a potential to code for a polypeptide product, a more reasonable approach may be to use a definition of "noncoding" that focuses on a coding-independent functional role of the untranslated RNA transcript. Thus, the key feature of is a lncRNA must function as a RNA transcript, whether or not it also codes for a polypeptide.

Transcription and processing. In many respects, lncRNAs resemble protein-coding mRNAs: mostly spliced, and polyadenylated RNA polymerase II (Pol II) transcripts (Fig. 3.2.1, top). Also, although not explicitly tested in most cases, lncRNAs are thought to be 5'-capped like mRNAs (Fig. 3.2.1, top). Pol II is more likely to be responsible for these long RNA transcripts due to its higher processivity, and RNA Pols I and III are generally limited to the transcription of short housekeeping RNA transcripts. The polyadenylation of lncRNAs is consistent with transcription by Pol II, and it helps to stabilize the transcripts to preserve their functional roles. Nonetheless, non-polyadenylated, Pol III-transcribed, noncoding RNA transcripts, such as BC200(Mus, Hof et al. 2007), and asOct4-pg5(Hawkins and Morris 2010), have been identified. Both are functional RNAs, playing roles in the regulation of translation and chromatin structure respectively, and are commonly referred to as lncRNAs in the literature. While BC200 is 200 nt long, barely fulfilling the minimum length requirement of lncRNAs, the actual length of asOct4-pg5 has not been evaluated and may be even shorter than 200 nt. Thus, the notion that Pol I and Pol III transcripts are too short too meet the criteria of a lncRNA still

holds true; BC200 may just be a rare exception that marginally escapes the arbitrary length cutoff.

Gene location and orientation. Historically, the focus has been on those lncRNAs encoded by genes that are well separated from genes encoding known protein-coding transcripts, hence the name "long intergenic noncoding RNAs (lincRNAs)," as noted above(Guttman, Amit et al. 2009; Khalil, Guttman et al. 2009; Guttman, Garber et al. 2010; Cabili, Trapnell et al. 2011). Nonetheless, as discovered in the large-scale discovery efforts noted below, "genic" lncRNAs are emerging as a prevalent class, with approximately one third to one half of lncRNAs overlapping protein-coding genes(Jia, Osak et al. 2010; Derrien, Johnson et al. 2012). They can be further divided into (1) natural antisense transcripts (NATs) (Zhang, Liu et al. 2006; Li, Zhang et al. 2008), which run in the opposite direction to known mRNA genes and overlap with their gene bodies either in the exonic or intronic regions, (2) divergent lncRNAs, which originate from the opposite DNA strand, but use the same promoters as mRNAs, extending in the opposite direction, and (3) promoter-associated lncRNAs and 3'-UTR-associated lncRNAs, grouped based on their proximal location relative to the start and end of mRNAs genes (Fig. 3.2.1, bottom). Various hypotheses have linked each of these classes of lncRNAs to a specific mode of action, but the extent to which gene location and molecular function is associated, and whether such location-based classification holds real biological meaning, still remains to be determined.

3.2.1.2. A Working Definition of LncRNAs

As illustrated here, questions remain regarding a unifying definition for lncRNAs. The field, however, has reached the point of having a solid working definition for lncRNAs. For the purpose of convenience and simplicity in identifying lncRNAs and distinguishing them from

other major classes of RNA transcripts, RNA molecules longer than 200 nt and having little coding potential are often classified as lncRNAs. They are very likely transcribed by Pol II, and in many cases, are capped, spliced, and polyadenylated.

3.2.2. Identifying and Cataloging lncRNAs

The earliest efforts to identify lncRNAs were mostly gene-specific, starting with the discovery of a novel transcript associated with a specific biological function and followed by the surprising realization that the function of the transcript is independent of the production of a protein product. More recently, significant advances in high-throughput sequencing technology and bioinformatics have revolutionized non-coding RNA discovery. In recent studies, the combined use of genomic and bioinformatic approaches have led to the large-scale identification of a whole host of novel lncRNAs. Consistent with the definition of lncRNAs, the general strategy involves two major steps: (1) the identification of novel transcripts that pass the 200 nt length threshold and (2) evaluation of their coding potential (Fig. 3.2.2). Large-scale lncRNA discovery efforts are often followed by gene-specific validations. The newly acquired information can then be consolidated into public databases, thus feeding back into the discovery process to facilitate identification of greater number of lncRNAs with higher confidence. In the following section, we highlight the major approaches used for lncRNA discovery.

3.2.2.1. Identification of LncRNA Transcripts – "Omics" Approaches

A number of different groups and consortia have used high-throughput sequencing technology and bioinformatics to facilitate non-coding RNA discovery.

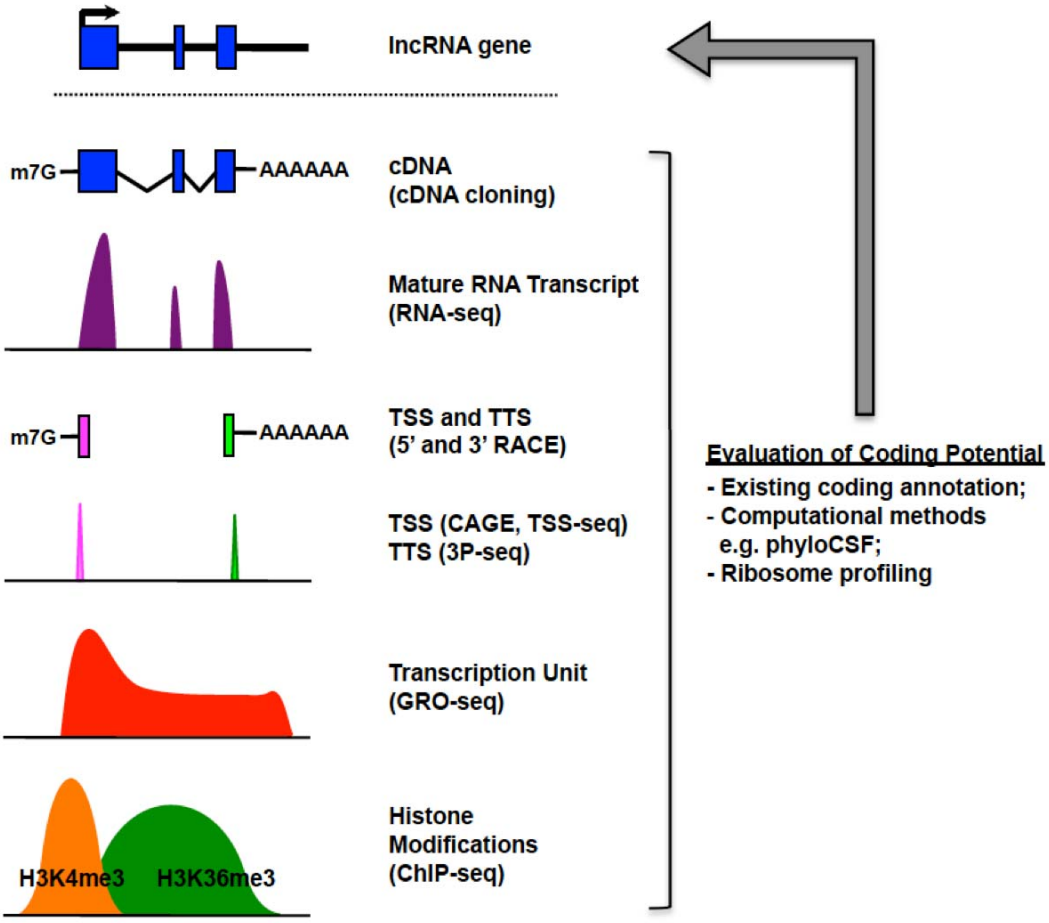


Figure 3.2.2 Methods for the identification of lncRNAs.

Schematics showing the strategies and approaches being used for the large-scale discovery and annotation of lncRNAs. The methods for the identification of novel transcripts include cDNA cloning, detection of mature cDNAs by RNA-seq, detection of TSS and TTS by CAGE, TSS-seq and 3P-seq as well as gene-specific RACE experiments, detection of transcription units by nascent transcript profiling using GRO-seq and by inference from histone modifications using ChIP-seq, and an integration of the abovementioned methods. Experimental and computational methods are then used to evaluate the coding potential of the identified novel transcripts to determine if they fit the definition of lncRNAs.

cDNA cloning. RIKEN's FANTOM (Functional Annotation of the Mammalian Genome) consortium pioneered the genome-wide discovery of lncRNAs, publishing a set of 34,030 polyadenylated lncRNAs from mouse in 2005(Carninci, Kasukawa et al. 2005). In addition, they isolated and cloned mouse full-length cDNA libraries for 5'- and 3'sequencing, and developed their own bioinformatics methods to map these transcripts to the mouse genome, resulting in 102,281 cDNAs as the starting point of lncRNA identification. To evaluate the coding potential of these cDNA transcripts, they searched for the presence of (1) protein-domain-like regions from Pfam(Finn, Mistry et al. 2006) and SUPERFAMILY databases(Wilson, Madera et al. 2007; Wilson, Pethica et al. 2009) and (2) transmembrane regions predicted by the TMHMM program(Krogh, Larsson et al. 2001), coiled coil regions predicted by the NCOIL program, and signal peptides predicted by the SignalP program(Nielsen, Engelbrecht et al. 1997). The absence of such protein-domain-like regions and the lack of open reading frame (ORF) longer than 100 amino acids were used to annotate one third of the cDNA transcripts as lncRNAs.

Histone modification signatures. In 2009, Guttman and colleagues proposed a different strategy that used global histone modificaton signatures to identify novel lncRNAs(Guttman, Amit et al. 2009). Using this approach, lncRNAs were defined as polyadenylated Pol II transcripts whose entire transcription units are longer than 5 kb and are well-separated from known protein-coding and microRNAs genes (the "lincRNA" definition note above). Using chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq), the authors generated genome-wide histone modificaton maps, focusing on those signatures associated with Pol II transcription (i.e., H3K4me3 at the promoter and H3K4me36 along the gene body), which served as markers of transcription units genome-wide. Polyadenylated exons were identified by microarray analyses with polyadenylated RNA across a random sample of 350

regions out of all 1675 transcription units identified. Due to a lack of splicing information, a conservative length cutoff of 5 kb was used to fulfill the length requirement of >200 nt. Codon substitution frequency (CSF)(Clamp, Fry et al. 2007), a measure of coding potential that examines evolutionary signatures characteristic to alignments of conserved coding regions, was evaluated for all intergenic transcripts in all three reading frames to confirm that the majority of the putative transcripts lack significant coding potential. Using this approach, the authors were able to identify approximately 1600 putative lncRNAs across four mouse cell types and about 3300 lncRNAs across six human cell types.

Identification of transcription units. Another approach that can be used to identify novel ncRNAs is the identification of transcription units. Methods that detect the densities of elongating RNA polymerases along the genome, such as GRO-seq(Core, Waterfall et al. 2008; Hah, Danko et al. 2011) or native elongating transcript sequencing (NET-seq)(Churchman and Weissman 2011), can be used to define the transcription units and serve as the basis for transcript discovery. A modified version of GRO-seq that incorporates initial steps of Rapid Amplification of 5'-end (5' RACE)(Frohman, Dush et al. 1988) is useful in determining the exact transcription start sites (TSS) of all transcripts. Additional genome-wide methods that facilitate the determination TSS include new cap-analysis gene expression (CAGE) (Shiraki, Kondo et al. 2003) and TSS-seq(Wakaguri, Yamashita et al. 2008; Tsuchihara, Suzuki et al. 2009). Gene identification signature and gene signature cloning ditag technologies, as shown in FANTOM3(Carninci, Kasukawa et al. 2005), can be used for the identification of sequences corresponding to both the TSS and the transcription termination sites. A method known as poly(A) position profiling by sequencing (3P-seq) can also be used to more precisely determine the directionality and end position of the polyadenylated transcription units(Jan, Friedman et al.

2011; Ulitsky, Shkumatava et al. 2011; Nam and Bartel 2012). Nevertheless, while these methods delineate the transcriptional landscape of potential lncRNA genes, information from RNA-seq will still be essential in elucidating the structure of the mature RNA transcripts contained within the corresponding transcription units, will in turn reveal the exact reading frame, allowing subsequent evaluation of coding potential.

Mature RNA structure. Characterization of the exon structure of lncRNAs has been facilitated by the development of bioinformatics algorithms that perform *ab initio* transcriptome reconstruction. Using programs such as Cufflinks(Trapnell, Williams et al. 2010) or Scripture(Guttman, Garber et al. 2010), entire transcriptomes of mammalian cells can be reconstructed using only RNA-seq reads and the genome sequence. RNA-seq reads directly reflect the position and structure of mature RNA transcripts. Compared to histone modification signature-based transcript determination, RNA-seq analysis gives a more accurate measurement of the length of mature RNA transcripts and information about exon-intron structure reveals the actual reading frame, allowing for more accurate calculation of coding potential. In the initial report using Scripture for transcript annotation, the authors identified over a thousand novel lncRNAs in three mouse cell types(Guttman, Garber et al. 2010). These lncRNAs are polyadenylated and multi-exonic, and have an average mature transcript length of 859 nt with very low coding potential.

More recently, both Cufflinks and Scripture have been used to assemble transcripts from RNA-seq datasets of very high sequencing depth in an attempt to accurately identify comprehensive lists of lncRNAs. One study examined lncRNAs in 24 human tissues and cell types, and cataloged the results in the Human Body Map lncRNA database(Cabili, Trapnell et al. 2011). Another study looked across eight time points during zebrafish embryogeneisis(Pauli,

Valen et al. 2012). The combined use of two independent assembly programs, together with high sequencing depth on multiple cells types or across multiple developmental stages, strengthens the confidence of the discovery process, especially because lncRNAs, as a group, have low expression levels, are highly cell-type specific, and are tightly developmentally regulated. In both studies, low CSF scores and the absence of Pfam domains were absolutely required for designation as a lncRNA, introducing extra criteria to ensure the non-coding status of identified lncRNAs.

Integration of approaches. Researchers have developed and improvised a variety of strategies to identify and annotate lncRNAs genome-wide. Moreover, they have integrated elements from these pipelines to facilitate lncRNA discovery. For example, in an effort to identify lncRNAs genome-wide in zebrafish, Ulitsky and colleagues also used H3K4me3 and H3K36me3 to mark promoters and gene bodies, but supplemented the histone modification maps with 3P-seq to more precisely map the polyadenylated end positions(Ulitsky, Shkumatava et al. 2011). They also incorporated existing transcriptome datasets, such as RNA-seq, annotated ESTs, and full-length cDNAs, to partially compensate for the lack of accurate mature RNA structures. A coding potential calculator (CPC)(Kong, Zhang et al. 2007) was used to determine the coding potential of each transcript. Collectively, the authors bioinformatically integrated multiple genomic datasets and identified 550 distinct lncRNAs in zebrafish.

3.2.2.2. Evaluation of Coding Potential

By definition, lncRNAs are unable to code for proteins. Determining the coding potential of a lncRNA, however, can be difficult. Three determinants have commonly been used for distinguishing noncoding RNAs from all identified RNAs: the length of longest ORF,

bioinformatically calculated coding potential, and the presence of coding potential for conserved protein domains. Among them, calculation of coding potential is the least straightforward. It involves the analysis of DNA alignments and codon usage across multiple species, favoring changes in amino acids that will preserve structural similarity versus changes that may lead to dramatic alterations in protein structure. In addition to the CSF and CPC scores mentioned above, other computational approaches examining coding potential include CSTminer(Castrignano, Canali et al. 2004), QRNA(Rivas and Eddy 2001) and CRITICA(Badger and Olsen 1999). CONC (Coding Or NonCoding) is a program that was developed based on support vector machines and can be used to classify transcripts according to features including peptide length, amino acid composition, predicted secondary structure content, predicted percentage of exposed residues, compositional entropy, number of homologs from database searches, and alignment entropy(Liu, Gough et al. 2006). The identification and characterization of a growing set of lncRNAs has allowed experimental validation of these bioinformatic approaches.

While most studies of lncRNAs have used the aforementioned bioinformatic approaches to evaluate their coding potential, ribosomal profiling is a direct experimental approach that can be used to address this issue. It was first developed to investigate the process of translation with sub-codon resolution, involves deep sequencing of ribosome-protected RNA fragments(Ingolia, Ghaemmaghami et al. 2009). It was then adapted to distinguish polyribosome-associated RNAs that are likely being translated from others that are more likely to be noncoding. Nam and Bartel identified polyadenylated transcripts in *C. elegans* using both RNA-seq and 3P-seq(Nam and Bartel 2012). Over 300 lncRNAs were identified from these transcripts after filtering through the coding potential threshold calculated from the CPC program and removing those that can be

detected in ribosome profiling experiments. However, ribosomal profiling requires further testing and validation, since association with the ribosome alone cannot be taken as the absolute evidence of protein coding potential. For example, both H19 and TUG1, two well-characterized lncRNAs, can be detected in association with the ribosome(Li, Franklin et al. 1998; Ingolia, Lareau et al. 2011). Some researchers have argued that instead of simply eliminating all transcripts identified in association with ribosomes (i.e., from ribosome profiling experiments), a more carefully examination of the preferential usage of a specific coding frames and features conferred by the release of the ribosomal complex at the site of the stop codon should be used to determine whether the transcript is productively translated.

3.2.2.3. Gene-Specific Validations

High-throughput sequencing and bioinformatics methods have led to tremendous progress in the large-scale identification of lncRNAs. Nevertheless, empirical validation of lncRNAs using a set of classical molecular biology techniques is still required. After learning the approximate location of a potential lncRNA transcript using global approaches, 5'- and 3'-RACE experiments can be carried out to determine the exact transcription initiation and termination sites, and to examine the presence or the absence of 5' cap and 3' poly(A) tail. PCR-based approaches can be used to isolate full-length cDNAs for those lncRNAs whose cDNAs are not available from public repositories, followed by traditional Sanger sequencing to obtain precise information on the exact exon-intron structure of the mature lncRNA transcript. Validation of the noncoding status of a putative lncRNA is less straightforward. In vitro transcription-translation assays have been used, but may give inconclusive results. In the case of SRA, functional outcomes associated with the RNA transcript were monitored after the

introduction of different missense and frameshift mutations, illustrating how one can prove that a lncRNA functions in a coding-independent manner(Lanz, McKenna et al. 1999). Nevertheless, this approaches demands prior knowledge of the functions of the identified lncRNAs.

3.2.2.4. Cataloging LncRNAs in Public Databases

The identification and characterization of a growing set of lncRNAs has provided additional insights into the properties of lncRNAs as a group, which facilitate subsequent efforts in lncRNA research. To make better use of the power of recursion, a number of lncRNA databases have been developed to consolidate and summarize the growing body of information.

Early lncRNA databases. In 2003, Barciszewski's group developed one of the first databases, NcRNAdb, which focuses on functional noncoding RNA transcripts that perform regulatory roles in the cell(Szymanski, Erdmann et al. 2003). In addition to small noncoding RNAs, this database consolidated the limited number of lncRNAs that had been identified at that time, but later gathered sequences from other relatively earlier databases including FANTOM3 (described above), GenBank, and H-Invitational Databases (H-InvDB). It was last updated in 2006, and has over 30,000 individual sequences from 99 species of Bacteria, Archaea and Eukaryota(Szymanski, Erdmann et al. 2007).

Second generation lncRNA databases. fRNAdb (2009)(Mituyama, Yamada et al. 2009) and NONCODE (2011)(Bu, Yu et al. 2012) are more recent databases that compile and integrate existing information of ncRNAs, including lncRNAs. They also provide support for associated computational analysis. fRNAbd currently contains 509,795 sequences and NONCODE has 423,976. A total of 1635 lncRNAs in NONCODE have been annotated with potential functions.

A number of public repositories for lncRNAs, such as Noncoding RNA Expression Database (NRED, 2009)(Dinger, Pang et al. 2009) and lncRNAdb (2011)(Amaral, Clark et al. 2011), have also been developed. NRED is linked to lncRNAdb. Both repositories integrate gene expression information with evolutionary conservation, secondary structure, genomic context, and antisense relationships for the cataloged lncRNAs. In the case of lncRNAdb, features such as subcellular localization and functional evidence associated with these lncRNAs have also been included.

Other catalogs collect lncRNAs using in-house annotation pipelines, such the Human Body Map lincRNA database (described above)(Cabili, Trapnell et al. 2011) and GENCODE(Derrien, Johnson et al. 2012), which are the most current and comprehensive. GENCODE, which is part of the ENCODE project, attempts to annotate all evidence-based gene features (cDNA, EST sequences) in the entire human genome at a high accuracy, and generates annotations of both protein-coding and noncoding genes, including a large number of lncRNAs. The latest version of GENCODE (February 2013 freeze, GRCh37) contains 13333 lncRNA genes from 22631 lncRNA loci transcripts.

Looking forward. At this point, a substantial proportion of all polyadenylated lncRNAs expressed in human have already been annotated, but these annotations need additional refinement and validation. In addition, give the tissue- and species-specificity of lncRNAs, there are most certainly more to be discovered. The majority of identified lncRNAs remains uncharacterized, but holds great promises for novel biological discoveries. With the existing annotations and functional databases, molecular biologists interested in the functional characterization of lncRNAs are no longer tied to the requirement of bioinformatics expertise and high cost of deep sequencing associated with de novo identification of lncRNAs. In many

cases, information extracted from existing databases may be a good starting point for characterizing previously identified lncRNAs with unknown functions and gain additional insights into the general features of lncRNAs as a group.

3.2.3. Functional Characterization of LncRNAs

Assigning molecular, cellular, and physiological functions to well annotated lncRNAs is the next great challenge in the field. Classical biochemical and molecular biology techniques have been instrumental in gene-specific functional characterization of lncRNAs. Gain-of-function and loss-of-function experiments can be used to validate the role of lncRNAs in modulating specific cellular processes. But, it is often challenging to determine whether an uncharacterized lncRNA plays an important functional role or which cellular process can be probed to yield an observable phenotype. Indeed, more efficient functional analyses, including high-throughput approaches linking lncRNAs to their probable functions, are required to keep pace with the tremendous progress made in lncRNA discovery. Below, we review a number of genomic strategies that facilitate large-scale functional characterization of lncRNAs.

3.2.3.1. Expression Profiling across Spatial and Temporal Gradients

The expression of lncRNAs is often cell type-, tissue-, and context-dependent. Therefore, the involvement of lncRNAs in specific cellular processes may be inferred by their differential expression patterns across tissues and across different developmental- or signal-regulated time points. For instance, Klattenhoff et al. identified lncRNAs that play critical roles in cardiovascular lineage commitment by reasoning that such candidates should demonstrate expression patterns restricted to specific cell types during embryonic stem cell (ESC)

differentiation (Klattenhoff, Scheuermann et al. 2013). They measured lncRNA expression in mouse ESCs and in differentiated tissues using RNA-seq and focused on 47 candidates whose expression levels were elevated in ESCs compared to other differentiated tissues. Among them, Braveheart, a lncRNA with higher expression in the heart relative to other tissues, was selected and characterized as a mediator of epigenetic regulation of cardiac commitment. Similarly, Kretz at el. focused on lncRNAs in keratinocyte differentiation, performing RNA-seq in primary keratinocytes during a calcium-induced differentiation time course (Kretz, Siprashvili et al. 2013). TINCR was one of the candidates chosen for further characterization, as it is among the most highly induced annotated lncRNAs whose expression changes during differentiation. Consistently, TINCR was found to be a key lncRNA required for somatic tissue differentiation by binding to differentiation mRNAs to stabilize their expression. These are just two examples where transcriptome profiling experiments across either spatial or temporal gradients generate clues to the functions of annotated lncRNAs. Since most lncRNA discovery approaches incorporate transcriptome profiling, it can be easily envisaged that when carefully designed, such efforts will not only yield information on the annotation and expression of novel and existing lncRNAs genes, but also shed light on the probable functions of a selected group of newly annotated lncRNAs.

3.2.3.2. Coding-Noncoding Coexpression Relationships – "Guilt-by-Association"

While the spatial and temporal gradients are helpful in choosing and characterizing a selected group of lncRNAs, additional approaches are needed for other situations. Guttman and colleagues have proposed a genomic approach to allow global functional characterization of lncRNAs, also known "guilt-by-association", which relies on correlation and clustering analysis

performed on mRNA expression profiling data and gene ontology or functional pathway analyses (Fig. 3.2.3) (Guttman, Amit et al. 2009). In this approach, groups of lncRNAs of unknown function are associated with groups of protein-coding mRNAs known to be involved in a specific cellular process based on a common expression pattern across cell-types and tissues. A positive correlation between the expression profile of a lncRNA and mRNAs suggests common function in the same cellular process. In their original paper, lincRNA-p21 was predicted to associate with p53-mediated DNA damage responses, with lincRNA-p21 later validated as a p53 target that modulates apoptotic responses upon DNA damage (Huarte, Guttman et al. 2010). The guilt-by-association approach is a useful first pass in assigning putative biological functions to lncRNAs and provides a working hypothesis for targeted perturbation experiments.

Zhao's group has expanded the analysis of gene coexpression relationships into a coding-noncoding coexpression network (CNC), making computational prediction of lncRNA functions through the evaluation of network characteristics (Liao, Liu et al. 2011). In addition to the coexpression network, co-localization relationships were also taken into consideration in their analysis. They focused on mouse lncRNAs annotated by FANTOM3 and extracted gene expression information from re-annotated Affymetrix Mouse Genome Array data. Ultimately, they predicted functions for 349 lncRNAs and further streamlined the application into a practical user interface called the Non-coding RNA Function Annotation Server (ncFANs) (Liao, Xiao et al. 2011). ncFANs is a useful tool for global prediction of lncRNA function, forming the basis of functional annotation in the NONCODE database, but its application is limited to annotated lncRNAs associated with corresponding microarray-based gene expression data.

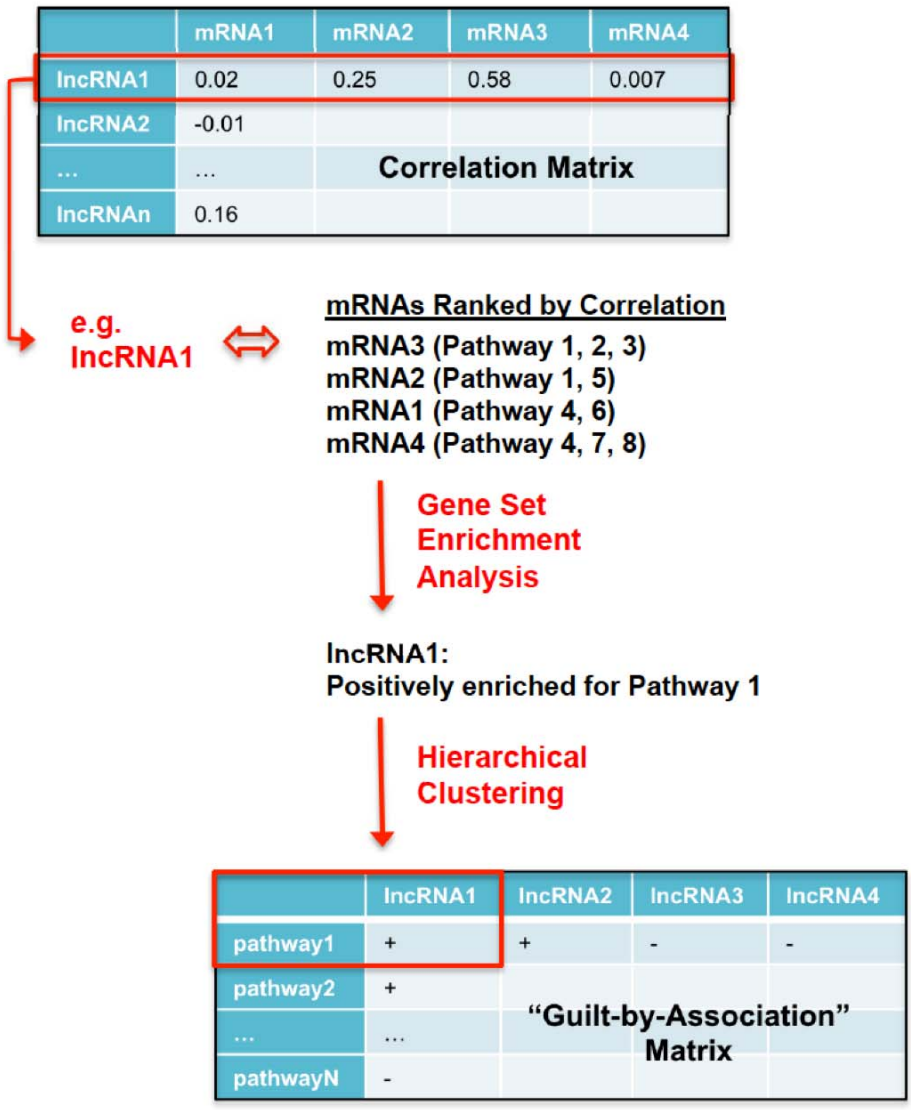


Figure 3.2.3 "Guilt-by-association" approach.

Schematics showing the "guilt-by-association" approach that relies on correlation and clustering analysis to serve as a first-pass method in determining the possible functions of uncharacterized lncRNAs. A correlation matrix is generated based on the correlation of expression between every pair of lncRNAs and mRNAs across many tissue samples and cell types. For each lncRNA, the corresponding list of mRNAs are ranked based on their correlation coefficients and it goes through the gene set enrichment analysis to produce statistically enriched functional pathways that are associated with the lncRNA. The procedure is performed for all lncRNAs of interest, and a "guilt-by-association" matrix can be generated through hieratical clustering of the enrichment relationships, resulting in groups of lncRNAs associated with distinct functional roles.

3.2.3.3. A Role for LncRNAs in Cis-Regulation of Gene Expression

One rationale behind the use of co-localization relationships in CNC-based functional characterization is that many lncRNAs have been shown to play a cis-regulatory role in the expression of nearby genes (Fig. 3.2.4A). For example, the gene for the ANRIL lncRNA overlaps and runs antisense to the gene encoding p15, mediating its gene silencing (Kotake, Nakagawa et al. 2011). In contrast, a chromatin-associated lncRNA CAR intergenic 10 is coexpressed with its flanking coding genes, FANK1 and Adam12, and helps to maintain their expression by establishing active chromatin structures (Mondal, Rasmussen et al. 2010).

The cis-regulatory function of HOTTIP involves an additional element. It is a lncRNA transcribed from the 5' end of the HoxA cluster and functions to activate the expression of neighbouring genes (Wang, Yang et al. 2011). Nevertheless, its influence extends to multiple distal HoxA genes owing to chromosome looping. The authors used chromosome conformation capture carbon copy (5C), a high throughput method to identify physical chromatin interaction, suggesting a model of how a cis-acting lncRNA can affect distal genes.

In a separate study, Ørom and colleagues selected candidates from the GENCODE database and suggested an enhancer-like function for several lncRNAs, which they termed ncRNA-activating (ncRNA-a), in activating the expression of neighbouring coding genes using heterologous transcription assays (Orom, Derrien et al. 2010). Similar to HOTTIP, some of the ncRNA-a were connected to their target genes through long-range chromatin loops. Lai and colleagues further demonstrated that these lncRNAs recruit the Mediator complex to their targets genes, and the Mediator complex plays an important role in forming DNA loops between the lncRNAs and their targets, and in mediating ncRNA-a-dependent gene activation (Lai, Orom et al. 2013).

The co-localization relationship has been exploited even further. To study lncRNAs involved in cell cycle regulation, Hung and colleagues looked in the proximity of known cell cycle genes and designed their approaches based on both "guilt-by-association" strategy and the cis-regulatory model (Hung, Wang et al. 2011). They used an ultra high-density array that tiles the promoters of 56 cell-cycle genes to interrogate 108 samples representing diverse conditions and perturbations, identifying 216 putative lncRNA transcripts originating proximal to these cell cycle gene promoters. Subsequently, they examined the coding-noncoding coexpression map across the conditions and clustered lncRNAs into different cell cycle-associated functions. The lncRNA PANDA (P21 associated ncRNA DNA damage activated) was selected for further analysis and was shown regulate apoptosis, consistent with the prediction.

3.2.3.4. A Role for LncRNAs in Trans-Regulation of Gene Expression

When coexpression and co-localization relationships are used as basis for functional prediction, direct perturbation experiments are required to validate the prediction. Therefore, Guttman and colleagues suggested a more direct approach for the functional characterization of lncRNAs, performing RNAi-based loss-of-function experiments and monitoring consequent changes in global gene expression (Guttman, Donaghey et al. 2011). They focused on previously identified lncRNAs expressed in embryonic stem cells and were able to successfully knock down the expression of 147 lncRNAs using custom designed short hairpin RNAs. For 137 lncRNAs, knockdown resulted in significant global changes in gene expression as shown in microarray analysis, and the majority had little effect on neighbouring genes, suggesting that these lncRNAs most likely affect gene expression in trans (Fig. 3.2.4B).

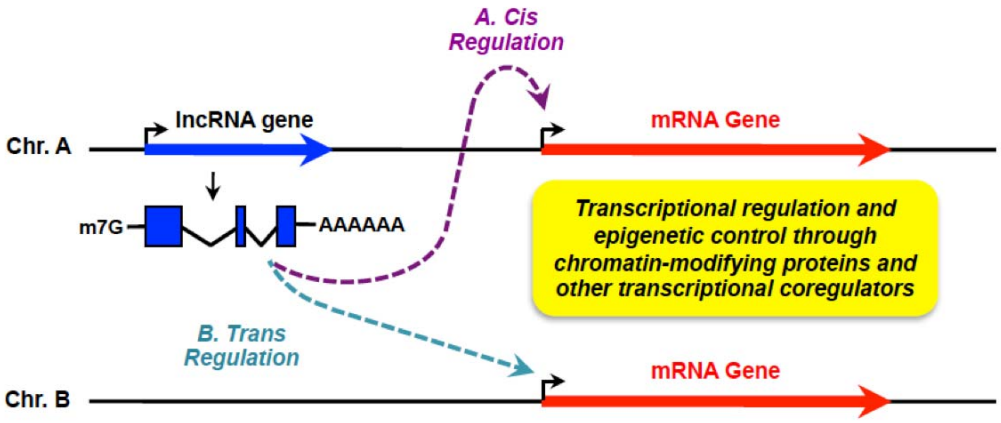


Figure 3.2.4 Cis and trans gene regulation by lncRNAs.

(A and B) LncRNAs can mediate transcription regulation and epigenetic control through chromatin-modifying proteins and other transcriptional coregulators in cis or trans. (A) In cis regulation, the lncRNA acts on target genes that are either located near the lncRNA gene or can be looped to the proximity of the lncRNA gene through higher-order chromatin structures. (B) In trans regulation, the lncRNA acts on target genes located distally to the lncRNA gene, possibly on another chromosome. Blue and red arrows indicate lncRNA and mRNA genes, respectively. Bent arrows indicate the TSSs of the genes.

These were not the first lncRNAs that have been associated with trans-regulation. HOTAIR, a well characterized lncRNA involved in developmental processes, is co-expressed with the HoxC genes, interacts with the chromatin-modifying PRC2 complex, and functions in trans to repress HoxD expression (Rinn, Kertesz et al. 2007). Interactions between HOTAIR and PRC2 proteins have been verified in both RNA-pulldown (captures proteins associated with a RNA bait) and RNA immunoprecipitation (RIP) (captures RNAs that are associated with proteins of interest using specific antibodies).

Indeed, there are many other lncRNAs that have been shown to interact with PRC2, including Braveheart (described earlier) (Klattenhoff, Scheuermann et al. 2013) and XIST (Zhao, Ohsumi et al. 2010), which coats the X chromosome to initiate and propagate X-inactivation (Penny, Kay et al. 1996; Marahrens, Panning et al. 1997). Other lncRNAs have been shown to interact with additional chromatin-modifying complexes. For example, HOTTIP binds and targets the WDR5/MLL complex across the HoxA to maintain active chromatin and coordinate homeotic gene expression (Wang, Yang et al. 2011). In addition, the tissue-specific lncRNA Fendrr has been shown to bind both the PRC2 and TrxG/MLL complexes, modulating chromatin signatures and gene activities to ensure the proper development of heart and body wall in mouse (Grote, Wittler et al. 2013).

Taken together, the connection between lncRNAs and chromatin-modifying complexes forms an appealing model of trans-regulation (Fig. 3.2.5A). Therefore, Lander and Rinn groups coupled the RIP assay to a microarray analysis (RIP-chip) to query many lncRNAs simultaneously (Khalil, Guttman et al. 2009). Among the 3300 human lncRNAs being queried, PRC2 or CoREST complexes were found to associate with 38% of them, suggesting that lncRNAs interacting with chromatin-associated complexes could be a common mechanism. In

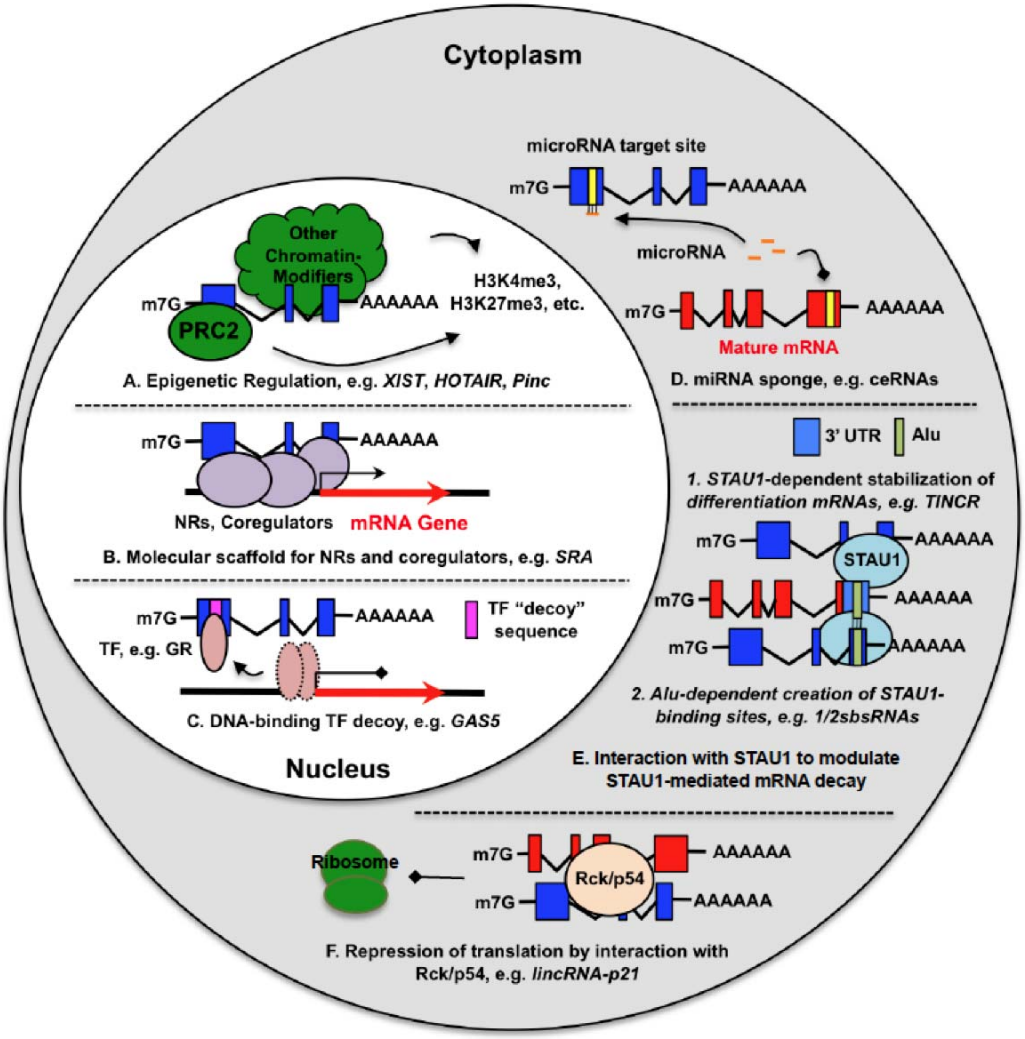


Figure 3.2.5 A broader view of lncRNA functions.

(A-F) LncRNAs mediate their functional roles through the regulation of gene expression, via a variety of molecular mechanisms both in the nucleus and in the cytoplasm. The nuclear functions of lncRNAs include (A) interaction with chromatin modifying complexes to alter epigenetic modifications, (B) interaction with transcription factors (TFs) such as nuclear receptors (NRs), and additional transcriptional coregulators to act as transcriptional cofactors themselves, and (C) having a decoy sequence to titrate away and repress the activity of DNA-binding TFs. Their cytoplasmic functions include (D) acting as a microRNA sponge as in the case of ceRNAs, (E) interaction with STAU1 and the regulation of STAU1-dependdent mRNA stability, and (F) interaction with the cytoplasmic RNA-binding protein Rck/p54 to inhibit translation.

addition, while RIP-chip requires prior knowledge of lncRNA sequences, Zhao and colleagues improved the method by replacing the microarray analysis with high throughput sequencing, which allows for unbiased identification of lncRNAs that interacts with candidate proteins (Zhao, Ohsumi et al. 2010). In this case, they tested their method on PRC2 and identified a genome-wide pool of >9000 PRC2-interacting RNAs in mouse embryonic stem cells. Not surprisingly, XIST is highly enriched in PRC2 RIPseq experiments, serving as a good positive control.

RIP-based experiments have helped to establish direct interactions between lncRNAs and proteins, suggesting that lncRNAs can act as molecular scaffolds to guide chromatin modifying complexes to their target genomic locations. Coupled with profiles of changes in chromatin signatures by ChIP-seq, the target sites of lncRNA action can be deduced. For example, changes in H3K4me3 and H3K27me3 marks were observed in HOTAIR knockdown foreskin fibroblasts, consistent with the modes of action of HOTAIR in targeting LSD1 and PRC2 to specific genomic locations to affect histone modifications (Rinn, Kertesz et al. 2007). Nevertheless, direct methods that capture the interaction between lncRNAs and chromatin sites have been developed recently: (1) chromatin isolation by RNA purification (ChIRP) (Chu, Qu et al. 2011) and (2) capture hybridization analysis of RNA targets (CHART) (Simon, Wang et al. 2011). In both methods, chromatin is crosslinked to lncRNA:protein adducts in vivo, followed by affinity capture of target lncRNA:chromain complexes using tiling antisense oligonucleotides in ChIRP or pre-selected oligonucleotides targeting RNase-H sensitive regions of the lncRNA in CHART. LncRNA-bound DNA were then isolated and sequenced to generate a genomic map of lncRNA binding sites. Such methods have been applied to trans-acting lncRNAs such as the Drosophila lncRNA roX2 and human HOTAIR to confirm their genomic binding sites.

3.2.3.5. Beyond the Nucleus: A Broader View of LncRNA Functions

LncRNAs play important roles in both cis- and tran-regulation of transcription, but continued studies are needed to determine the relative contributions of cis and trans mechanisms of lncRNA function. There is a strong bias in the field for this potential aspect of lncRNA function, leading to the common belief that lncRNAs as a group are mostly involved in transcriptional regulation. Although lncRNAs as a group may show a slight enrichment for the nuclear compartment, many lncRNAs are predominantly or even exclusively cytoplasmicly localized. Inherent biases in some previous analytical approaches, however, have propagated the emphasis on nuclear functions for lncRNAs. For example, PRC2 RIP-based methods have suggested that a large number of lncRNAs are involved in PRC2-mediated transcriptional repression. Nevertheless, the RIP protocol used limited the analysis to nucleus-retained RNAs, leaving open the possibility that a larger proportion of lncRNAs interact with cytoplasmic proteins. Indeed, there have been an increasing number of examples of cytoplasmic lncRNAs. Among them, half-STAU1-binding site RNAs have been shown to transactivate the binding of STAU1 protein to its target mRNAs to facilitate mRNA decay (Fig. 3.2.5E) (Gong and Maquat 2011). On the other hand, TINCR, another lncRNA that has been shown to bind STAU1, interacts with differentiation mRNAs to mediate their stabilization in a STAU1-dependent manner (Fig. 3.2.5E) (Kretz, Siprashvili et al. 2013).

Furthermore, when Huarte et al. attempted to delineate the mechanism of action of lincRNA-p21 by identifying its interaction partners using RNA pull down, nuclear extract was used, and the nuclear RNA binding protein hnRNP was found to associate with this lncRNA to facilitate its action in mediating gene repression (Huarte, Guttman et al. 2010). Yoon and colleagues confirmed this interaction in an anti-hnRNP RIP experiment (Yoon, Abdelmohsen et

al. 2012). More interestingly, as they searched for RNA partners for the cytoplasmic RNA binding protein HuR in anti-HuR RIP experiment using whole cell lysates, lincRNA-p21 was readily enriched as well (Yoon, Abdelmohsen et al. 2012). This interaction accelerates the degradation of lincRNA-p21, which in turn ameliorates its interaction with additional cytoplasmic RNA-binding proteins Rck/p54, and derepresses the expression of a subset of target mRNAs, elucidating an additional role of cytoplasmic lincRNA-p21 as a post-transcriptional inhibitor of translation (Fig. 3.2.5F).

LincRNA-p21 is just one example showing that methods limited to the characterization of nucleus-retained lncRNAs are thus not sufficient to provide us with a complete spectrum of functional roles played by lncRNAs. Delineating the cellular localization of lncRNAs in an unbiased manner should be one of the first steps in used for gathering more clues on their possible functional roles. Nucleus-retained lncRNAs are more likely to be involved in transcriptional regulation, while cytoplasmic lncRNAs may have other functions. RNA fluorescence in situ hybridization (FISH) is a popular method that has been used to visualize the cellular localization of lncRNAs, but challenges remain for a high-throughput FISH approach that examines many lncRNAs simultaneously. Alternatively, lncRNAs can be extracted from each of the physically defined cellular compartments and then sequenced, revealing the relative amount of each lncRNA in the various cellular fractions. With modifications as described in Yoon et al., RIP-based methods can also be used with key cytoplasmic proteins that act in important cellular pathways to identify and characterize cytoplasmic lncRNAs involved in those pathways (Yoon, Abdelmohsen et al. 2012). Furthermore, Kretz et al, who characterized TINCR, utilized a protein microarray analysis containing approximately 9400 recombinant

human proteins (Human Protoarray) to identify the TINCR-STAU1 interaction in the cytoplasm (Kretz, Siprashvili et al. 2013).

3.2.3.6. Lessons Learned from the Best-Characterized LncRNAs

Using methods described above and additional strategies, a growing number of lncRNAs have been characterized molecularly and functionally. A limited few are as well characterized as some protein-coding RNAs. Below, we summarize the current status of the few best-characterized lncRNAs to date and highlight the lessons learned from these examples.

XIST. The X-inactive-specific transcript (XIST) was one of the first lncRNAs to be discovered in mammals(Borsani, Tonlorenzi et al. 1991; Brockdorff, Ashworth et al. 1991; Brockdorff, Ashworth et al. 1992; Brown, Hendrich et al. 1992). It is responsible for the initiation and spreading of X-chromosome inactivation (XCI) in female somatic cells(Penny, Kay et al. 1996; Marahrens, Panning et al. 1997; Wutz and Jaenisch 2000; Wutz, Rasmussen et al. 2002). XIST is transcribed from the XCI loci and acts in concert with the transcription factor YY1 and several other lncRNAs from the same locus (e.g., RepA, Tsix, Jpx/Enox) to facilitate the loading of PRC2 and initiate DNA methylation and the subsequent chromosome-wide silencing(Lee, Davidow et al. 1999; Lee and Lu 1999; Sado, Hoki et al. 2005; Ogawa, Sun et al. 2008; Zhao, Sun et al. 2008; Tian, Sun et al. 2010; Jeon and Lee 2011). It is one of the best examples of multiple lncRNAs utilizing their base complementarity properties to collaborate with each other and with proteins to achieve a common cellular function. This could be a recurring theme with lncRNAs, which may base pair with DNA in the genome or RNA elements in the transcriptome, creating unique interfaces for RNA-protein interactions. LncRNAs

encompass RNA motifs with variable lengths, offering advantages over small protein motifs and allowing more specificity in targeting to unique addresses.

Even after more than two decades of extensive research, the exact mechanism of XIST-mediated spreading of XCI is yet to be fully elucidated. This is due, in part, to the lack of high throughput approaches of sufficient resolution to distinguish allelic differences of the X chromosomes. To address this, Pinter et al. developed allele-specific ChIP-seq, mapping the positions of the PRC2 component EZH2 and XCI-associated histone marks on the inactive (Xi) and active (Xa) X chromosomes separately over a developmental time course(Pinter, Sadreyev et al. 2012). The authors presented a model in which XCI is governed by a hierarchy of defined PRC2 stations that spread H3K27 methylation in cis. Following in the path of allele-specific ChIP-seq, allele-specific ChIRP or CHART of XIST on the X chromosomes could be used to visualize the spreading of XIST on the inactive X chromosome along the developmental axis more directly.

MALAT1. MALAT1 is one of the first cancer-associated lncRNAs discovered, hence its name metastasis-associated lung adenocarcinoma transcript 1(Ji, Diederichs et al. 2003). It is extremely abundant and highly conserved over its full length across all mammalian species, both properties that highlight its likely importance(Ji, Diederichs et al. 2003; Hutchinson, Ensminger et al. 2007; Bernard, Prasanth et al. 2010). MALAT1 localizes to nuclear bodies known as nuclear speckles(Yang, Lin et al. 2011), suggesting functions in the nucleus. In cell-based models, MALAT1 has been shown to regulate alternative splicing and gene expression at the molecular level(Bernard, Prasanth et al. 2010; Tripathi, Ellis et al. 2010; Yang, Lin et al. 2011), contributing to its association with metastatic lung adenocarcinoma. Given these preliminary results, the observation that MALAT1 knockout mice display little observable phenotype,

especially with respect to splicing or gene expression, is surprising(Zhang, Arun et al. 2012). The field must address why lncRNAs that show cell-based phenotypes are not functional in vivo, as is the case with MALAT1. Parallels between lncRNAs and the better understood class of microRNAs may help to explain such conundrums. Phenotypic evaluation of microRNA knockout mice has revealed similar disappointing phenotypes(Liu, Bezprozvannaya et al. 2008; Jin, Hirokawa et al. 2009; Williams, Valdez et al. 2009; Park, Jeker et al. 2012). But new studies suggest that the most dramatic phenotypes often arise in response to specific cellular signals, such as special diet or stress, or under a compromised genetic background(van Rooij, Sutherland et al. 2007; Callis, Pandya et al. 2009). In other words, the appropriate cellular context is essential.

In this regard, given the association between MALAT1 and lung adenocarcinoma, it will be interesting to cross the MALAT1 knockout mice with genetic models of lung cancer, or to generate lung-specific MALAT1 knockouts and treat them with oncogenic agents, to determine whether any transcriptional and phenotypic consequences arise. In this regard, Gutschner et al. diminished MALAT1 expression in A549 human lung adenocarcinoma cells using a zinc finger nuclease-mediated KO approach(Gutschner, Hammerle et al. 2013). They observed changes in gene expression and impairment in the metastatic potential of these MALAT1-deficient cells in mice xenograph experiments, once again established a critical role of MALAT1 as a regulator of gene expression governing hallmarks of lung cancer metastasis(Gutschner, Hammerle et al. 2013). Not unlike the situation with microRNAs, when probing the in vivo functions of lncRNAs, it is important to find the right context in order to uncover the observable phenotype.

HOTAIR. HOTAIR is a 2.2 kb lncRNA transcribed from the HOXC locus that functions to repress transcription in trans across 40 kilobases of the HOXD locus(Rinn, Kertesz et al.

2007). Similar to MALAT1, HOTAIR is a cancer-associated lncRNA, including breast, colorectal, nasopharyngeal, and hepatocellular cancers(Gupta, Shah et al. 2010; Geng, Xie et al. 2011; Kogo, Shimamura et al. 2011; Nie, Liu et al. 2013), although its prognostic value in clinical oncology is still undetermined. Mechanistically, it is the first lncRNA to be found to associate with PRC2 complexes(Rinn, Kertesz et al. 2007), initiating the subsequent characterization of a large number of PRC2-interacting RNAs later known as the PRC2 transcriptome(Khalil, Guttman et al. 2009; Zhao, Ohsumi et al. 2010). It is also the first mammalian lncRNA to be screened by ChIRP, demonstrating its direct association with GA-rich regions of chromatin that nucleate broad domains of Polycomb and H3K27me3 occupancy(Chu, Qu et al. 2011). Tsai and colleagues showed that not only does the 5' domain of HOTAIR binds to PRC2, but the 3' domain binds LSD1, a chromatin modifying complex that promotes H3K4me3 demethylation, suggesting a role for HOTAIR as a molecular scaffold possessing distinct RNA domains for protein interactions (Tsai, Manor et al. 2010).

The characterization of HOTAIR illustrates two aspects of lncRNA function. First, it provides a model for the function of a lncRNA that regulates transcription in trans, through tethering to chromatin regions and recruiting chromatin modifying complexes. Second, it shows that lncRNAs can be modular, not unlike proteins. Its functions can be separated into independent molecular domains that act in collaboration. These results suggest ways of studying lncRNAs in a manner similar to studying proteins. We can draw hints and insights from the prediction or biochemical mapping of RNA structures, as well as from information of evolutionary conversation, and perhaps can even work towards building a database of lncRNA domains or motifs, which will help to elucidate the functions of lncRNAs, much in the same way

PFAM(Finn, Mistry et al. 2006) and PROSITE(Hulo, Bairoch et al. 2006) have done for proteins.

3.2.4. Emerging Roles of lncRNAs in Molecular Endocrinology

As described in the previous sections, efforts to understand the biology of lncRNAs are beginning to shed new light on their roles in physiological and pathological processes. Accumulating evidence has pointed to key roles of lncRNAs in development and differentiation (e.g., XIST, HOTAIR, TINCR, Braveheart, Fendrr), as well as cell proliferation and cell death (e.g., PANDA, lincRNA-p21, ANRIL), but lncRNAs are likely to be functionally involved in many more, if not all, cellular processes. Emerging evidence has suggested roles for a number of lncRNAs in various endocrine functions of the reproductive and metabolic tissues. Most of them are involved in hormonal-regulated signaling pathways.

Steroid receptor activator (SRA). The first link between a lncRNA and hormone receptor-associated pathways was established more than two decades ago with the discovery of steroid receptor activator (SRA) by O'Malley and colleagues(Lanz, McKenna et al. 1999). SRA was initially described as an RNA transcript specifically expressed in steroid hormone target tissues, which functions as a steroid receptor coactivator. The SRA RNA interacts with steroid receptor coactivators 1 and 2 (SRC-1, SRC-2) and facilitates ligand-dependent transactivation in reporter gene assays. In a careful series of biochemical and cellular experiments, the authors used mutations that introduce early stop codons in SRA and inhibitors of protein synthesis to convincingly demonstrate that the coactivator function of SRA is independent of of translated protein products. Subsequent studies have substantiated the earlier findings and identified additional interaction partners involving both coactivators (e.g., p68, p72, Pus1p and

Pus3p)(Watanabe, Yanagisawa et al. 2001; Zhao, Patton et al. 2007) and corepressors (e.g., SHARP and SLIRP)(Shi, Downes et al. 2001; Hatchell, Colley et al. 2006), thus expanding the role of SRA as a transcriptional coregulator (Fig. 3.2.5B). More recently, protein-coding isoforms of the SRA gene, containing an extended exon-1, have also been identified(Hube, Velasco et al. 2011), making it an interesting case of an RNA with dual roles as both a lncRNA and a protein-coding RNA. Nevertheless, the noncoding isoform displays differential expression patterns across different breast cancer cell lines and appears to play oncogenic roles in breast cancer tumorigenesis (Chooniedass-Kothari, Hamedani et al. 2006; Leygue 2007; Cooper, Guo et al. 2009), making the studies of such lncRNAs highly relevant to endocrine cancer research.

Growth arrest-specific 5 (GAS5). Growth arrest-specific 5 (GAS5) is another lncRNA that has been shown to regulate the activity and function of multiple receptors, including the glucocorticoid, androgen, mineralcorticoid, and progesterone receptors (Kino, Hurt et al. 2010). Unlike SRA, which participates in steroid coactivator complexes as a scaffold, GAS5 forms an RNA stem-loop structure to mimic a DNA response element. In the context of glucocorticoid receptor (GR), it interacts with the GR DNA-binding domain and acts as a decoy GR response element, titrating GR away from its sites of transcriptional activity in a ligand-dependent manner (Fig. 3.2.5C). The GAS5 RNA accumulates in fasting and growth arrested cells, thus functioning as a starvation- or growth arrest–linked riborepressor for GR and possibly other nuclear receptors that share the same DNA response element sequence, facilitating steroid-modulated cell survival and metabolism(Kino, Hurt et al. 2010; Williams, Mourtada-Maarabouni et al. 2011). In human adeherent cell lines including 293T and MCF-10A, Mourtada-Maarabouni and colleagues have shown that overexpresssion of GAS5 suppresses cell growth and promotes apoptosis, and it is found at reduced levels in human breast carcinoma samples compared to their

matched controls, suggesting a role of GAS5 as a tumour suppressor(Mourtada-Maarabouni, Pickard et al. 2009).

Progesterone receptor gene antisense transcripts. Some receptor-related lncRNAs may be more receptor-specific. Corey's lab has examined the transcriptional landscape of the progesterone receptor (PR) gene and showed the existence of antisense RNA transcripts overlapping the PR gene promoter(Schwartz, Younger et al. 2008). They are likely to be lncRNAs, and at least one of them is spliced and polyadenylated. Although the coding potential of these transcripts have not been explicitly evaluated, they appear to be acting at the RNA level through base complementarities to other RNA molecules. Specifically, duplex RNAs, or antigene RNAs (agRNAs), complementary to the PR gene promoter increase expression of PR mRNA and protein levels after transfection into human breast cancer cells(Janowski, Younger et al. 2007; Schwartz, Younger et al. 2008). Interestingly, the antisense lncRNAs are required for the agRNA-mediated activation of the PR gene, possibly through base pairing with the agRNAs(Schwartz, Younger et al. 2008; Janowski and Corey 2010). The possibility that these PR antisense lncRNAs are involved in the modulation of PR gene expression is an attractive one, and Corey and colleagues continue to search for endogenous agRNA-like molecules that might mediate these effects. microRNAs are possible candidates. Indeed, the inhibitory effects of mir123b on PR gene expression can be inhibited by PR antisense lncRNAs(Janowski and Corey 2010), suggesting a role of lncRNAs in acting as competing endogenous RNAs (ceRNAs) to sequester microRNAs, thus adding to the growing list of lncRNAs acting as ceRNAs in multiple cellular models (Fig. 3.2.5D)(Cesana, Cacchiarelli et al. 2011; Karreth, Tay et al. 2011; Salmena, Poliseno et al. 2011; Tay, Kats et al. 2011).

Pregnancy-induced noncoding RNA (PINC): A hormone-regulated lncRNA. The examples noted above illustrate how direct or indirect interactions between lncRNAs and nuclear receptors (or their genes) can affect receptor activity or expression. Other lncRNAs function as downstream targets of the gene-regulating activities nuclear receptors. Rosen and colleagues have identified pregnancy-induced noncoding RNA (PINC) as a lncRNA that is persistently up-regulated in the involuted mammary glands of estrogen- and progesterone-treated rodents (Shore, Kabotyanski et al. 2012). Although its function during early pregnancy is unclear, it may play a key role in regulating the development of lactating mammary glands. The levels of the PINC transcript decline as mammary alveolar cells undergo terminal secretory differentiation. In HC11 mouse mammary epithelial cells, PINC levels decreases upon lactogenic differentiation following hormone treatment. Reduction in PINC levels enhances the lactogenic differentiation of HC11 cells, as shown in knockdown and overexpression experiments. Mechanistically, PINC has been shown to interact with the chromatin-modifying PRC2 complex in RNA immumoprecipitation assays. The PRC2 complex is also known to bind many other lncRNAs. In established molecular models, lncRNAs such as HOTAIR guide the PRC2 complex to its target genomic loci to exert epigenetic modification and transcriptional regulation (Fig. 3.2.5A). Whether PINC plays similar roles has yet to be determined.

Estrogen-regulated lncRNAs. Global nuclear run-on and sequencing (GRO-seq) was recently used to explore the rapid effects of estrogen signaling on the entire transcriptome in MCF-7 human breast carcinoma cells, as well as identify thousands of novel estrogen-regulated ncRNAs, including lncRNAs(Hah, Danko et al. 2011). Like some previously characterize estrogen-regulated protein coding genes, many estrogen-upregulated lncRNAs show estrogen-induced ER binding in their proximal promoter regions, suggesting direct regulation by ER.

Those lncRNAs that show rapid and robust regulation in response to estrogen signaling are likely to play important roles in the estrogen signaling pathway. Of course, further investigation is required to fully annotate and accurately determine the complete set of lncRNAs associated with the estrogen signaling pathway, as well as to functionally characterize the molecular mechanisms and biological roles played by these lncRNAs.

lncRNAs regulating adipogenesis. By evaluating the differential expression of lncRNAs across primary brown and white adipocytes, preadipocytes, and cultured adipocytes, Sun and colleagues identified 175 lncRNAs that are specifically regulated during adipogenesis. Out of the 175, they selected twenty lncRNAs that are likely regulated by PPARγ and CEBPα, the master regulators of adipogenesis, to perform a loss-of-function screen, from which they showed that ten of them function to modulate the progression of adipocyte differentiation(Sun, Goff et al. 2013). The adipocyte is an active endocrine cell. Therefore, lncRNAs regulating adipocyte differentiation have important implications in endocrine-mediated metabolic functions.

3.2.5. Conclusions and Perspectives

The introduction of whole transcriptome sequencing methods, such as RNA-seq and GRO-seq, as well as the FANTOM and ENCODE transcript mapping projects, have transformed our perspectives on the variety and dynamic nature of lncRNAs. A growing number of lncRNAs is being characterized and has been shown to play central roles in various biological processes, including cancer, metabolism, and endocrinology. The list is growing with increasing interest and efforts in the field.

Apart from broadening our perspectives on fundamental aspects of biology, lncRNAs offer possibilities in medicine. Given their tissue-specific expression patterns, lncRNAs can be

superior biomarkers for certain diseases. In the case of prostate cancer, PCA3 test has been developed and used clinically, utilizing the observation that the PCA3 lncRNA is specifically overexpressed in prostate cancer(Lee, Dobi et al. 2011). Moreover, there is a transition from developing lncRNA-based diagnostics to exploring lncRNA-based therapies. For example, antisense oligos (ASO) that attenuate the expression of MALAT1 in EBC-1 lung cancer cells inhibits metastasis to the lung(Gutschner, Hammerle et al. 2013). While still in its infancy, the therapeutic promise of lncRNAs could be tremendous. The growing interest in the identification and characterization of endocrine-associated lncRNAs could lead to new developments in the diagnostic, prognostic, and even therapeutics of endocrine diseases. Nonetheless, a more thorough understanding of the biological functions of lncRNAs along with their detailed mechanisms of action, are required before we can fully exploit their potential utility.

Therefore, in the current study, I explored the implications of lncRNAs in breast cancer, which is another area where the clinical utility of lncRNAs is waiting to be exploited. There are already a number of lncRNAs, such as *SRA*, *GAS5* and *HOTAIR*, that have been associated with the development and treatment outcomes of breast cancer. In addition, transcriptome analysis in MCF-7 breast cancer cells reported the abundance of lncRNAs and E2-regulated lncRNAs (Hah, Danko et al. 2011), hence raising the possibility that they are playing a previously under-appreciated role in breast caner biology. In this study, I developed an integrative approach that incorporates the analysis of multiple high-throughput sequencing datasets and existing resources, to generate a comprehensive catalog of lncRNAs in basal and E2-stimulated MCF-7 cells. I revealed many interesting features of lncRNAs, and showed that the differential expression of lncRNAs predicts the intrinsic molecular subtype of breast cancer cells, suggesting their potential utility as prognostic biomarkers. Furthermore, by examining the tumor-specific

expression of lncRNAs and the “guilt-by-association” approach, I am able identify lncRNAs that are required for the normal growth of MCF-7 breast cancer cells. Whether these lncRNAs can potentially serve as new targets for therapeutic interventions will be the subject for future investigation.

3.3. Results

Integrative analysis of RNA-seq and GRO-seq generates a comprehensive catalog of lncRNA genes in MCF-7 cells

To identify and annotate lncRNA genes in the estrogen-responsive MCF-7 human breast cancer cells, we developed a computational approach that incorporates evidence of RNA transcripts from multiple high-throughput sequencing datasets over a time course following treatments with E2. In brief, our pipeline consists of three major parts: (1) mapping and assembly of RNA transcripts from polyadenylated (polyA+) RNA-seq; (2) integration of nascent RNA profiles from GRO-seq; (3) processing and filtering of transcripts based on length, coverage, expression levels, and coding potential. Schematics and detailed steps of the analysis were illustrated in Fig. 3.3.1A and Fig. 3.3.2A. Sensitivity and specificity along the pipeline were reported in Fig. 3.3.2B.

We sequenced polyA+ RNAs extracted from the whole cell, and from each of the cytoplasmic, nucleoplasmic and chromatin-associated fractions to evaluate mature RNA contents (Fig. 3.3.1C). Upon alignment to the reference human genome and transcript reconstruction as described previously (Cabili, Trapnell et al. 2011), we observed that cytoplasmic RNAs are more completely spliced, while RNAs in the nucleus often contain varying amounts of unspliced introns (Fig. 3.3.1D). It provides evidence for post-transcriptional splicing, and suggests that

Figure 3.3.1 Integrative analysis of RNA-seq and GRO-seq generates a comprehensive catalog of lncRNA genes in MCF-7 cells.

(A) An overview of the experimental and analysis pipeline for the identification of lncRNA genes in MCF-7 cells (lncM).

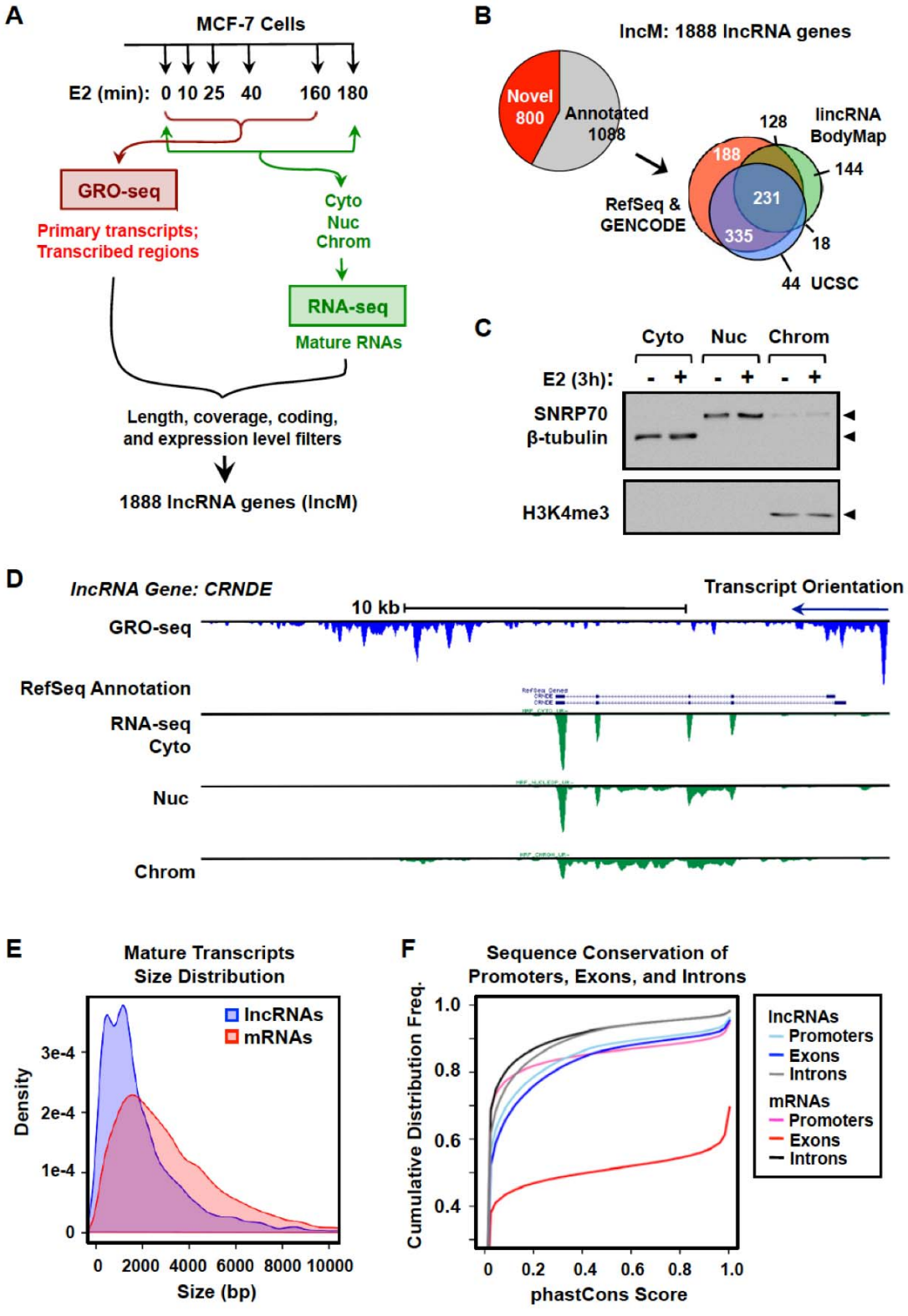
(B) The Venn diagrams show the fraction of lncM genes that are novel (red) and annotated (grey), and their overlap with annotations from RefSeq and GENCODE (orange), UCSC (purple), and lincRNA BodyMap (green) databases.

(C) Western blotting shows the purity of the subcellular fractions used for RNA-seq. β-tubulin, SNRP70 and H3K4me3 are used as fraction-specific markers.

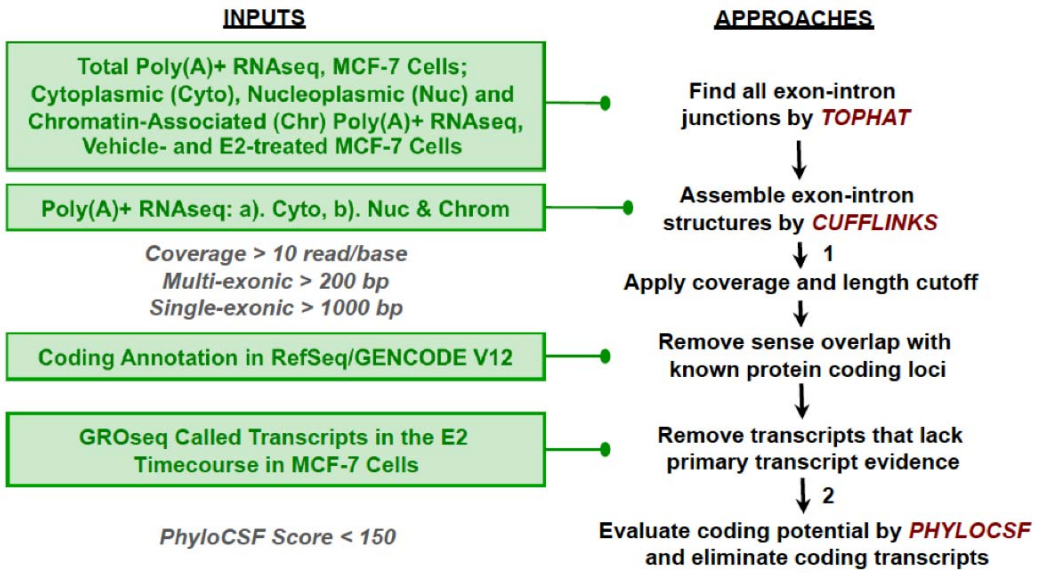
(D) Genome browser view for the locus of an annotated lncRNA gene, CRNDE, showing RefSeq annotation, GRO-seq (blue) and fractionated RNA-seq (green) data.

(E) Comparisons of mature transcript size distributions between lncRNA (blue) and protein-coding genes (red) assembled from cytoplasmic RNA-seq data.

(F) Cumulative distribution frequency curves show the sequence conservation of the promoters, exons and introns of lncRNA and protein-coding mRNAs assembled from cytoplasmic RNAs.


A
MCF-7 Cells
E2 (min): 0 10 25 40 160 180
GRO-seq
Primary transcripts;
Transcribed regions
Cyto
Nuc
Chrom
RNA-seq
Mature RNAs
Length, coverage, coding,
and expression level filters
1888 lncRNA genes (lncM)
B
lncM: 1888 lncRNA genes
Novel
800
Annotated
1088
RefSeq &
GENCODE
128
lincRNA
BodyMap
188
144
231
335
18
44 UCSC
C
Cyto
Nuc
Chrom
E2 (3h): - + - + - +
SNRP70
β-tubulin
H3K4me3
D
lncRNA Gene: CRNDE
Transcript Orientation
10 kb
GRO-seq
RefSeq Annotation
RNA-seq
Cyto
Nuc
Chrom
E
Mature Transcripts
Size Distribution
lncRNAs
mRNAs
Density
3e-4
2e-4
1e-4
0
0 2000 4000 6000 8000 10000
Size (bp)
F
Sequence Conservation of
Promoters, Exons, and Introns
Cumulative Distribution Freq.
1.0
0.8
0.6
0.4
0 0.2 0.4 0.6 0.8 1.0
phastCons Score
lncRNAs
Promoters
Exons
Introns
mRNAs
Promoters
Exons
Introns

A



B

	Sensitivity	Specificity
1 (CYTO - NR)	4.5 - 9.0	4.6 - 12.3
2 (CYTO - NR)	4.1 - 9.5	28.8 - 21.5
3 (CYTO - NR)	4.0 - 9.1	30.5 - 21.7
lncM	5.1	31.5
lincRNA BodyMap	3.7	40.6

Figure 3.3.2 Generation of the lncRNA catalog.

(A) The pipeline takes as inputs (left) polyadenylated RNA-seq data from total and fractionated, basal and E2-stimulated MCF-7 cells, existing coding annotation from RefSeq and GENCODE for humans, and previously available GRO-seq datasets across an estrogen treatment timecourse. RNA-seq data were aligned to the genome and assembled by Tophat and Cufflinks respectively. Cytoplasmic-based RNA-seq data (lncCYTO) were evaluated separately from nucleus-based RNA-seq (lncNR). Transcription units were called de novo from GRO-seq data using previously described methods. RNA transcripts with both RNA-seq and GRO-seq information were further filtered by coverage, size, and coding capabilities. The remaining transcripts pass through a FPKM threshold (>1) to derive at a set of 1888 expressed lncRNA genes (lncM).

(B) Sensitivity and specificity at the base level measured by Cuffcompare, comparing transcripts from lncCYTO and lncNR at indicated points along the pipeline, lncM and lincRNA Bodymap annotations, to GENCODE lncRNA annotations.

mature RNA structures assembled based on the cytoplasmic RNAs will be more accurate. Therefore, we relied on cytoplasmic RNAs for the determination of exon-intron structure when possible, but also included transcripts that can only assembled from nuclear RNAs, to achieve a comprehensive list of lncRNAs. To capture the regulation of estrogens, we performed fractionation followed by RNA-seq in MCF-7 cells following 0 and 3 hours of E2 treatment. We also incorporated into our pipeline the nascent RNA profiles of these cells, as measured by GRO-seq, with 0, 10, 25, 40 and 160 min. of E2 treatment, calling transcription units de novo as previously described (Danko, Hah et al. 2013), (Hah, Danko et al. 2011). It requires evidence of both mature and nascent RNAs to be included in our lncRNA catalog.

Furthermore, lncRNAs in our catalog need to be reasonably "long", most likely "noncoding" and reliably expressed. We applied the widely used length cutoff of 200 nt to multi-exonic RNA transcripts. In the case of single-exonic transcripts, which are more susceptible to the limitations on the resolution of transcript assembly, a 1000-nt cutoff was used. To ensure the noncoding status of lncRNAs, we excluded transcripts with any overlap to known protein coding loci (RefSeq and GENCODE) running in the same direction, and eliminated transcripts scoring relatively high codon substitution frequency (i.e. phyloCSF score > 150) (Lin, Jungreis et al. 2011), a bioinformatically computed measure of coding potential. All transcripts were required to pass a coverage threshold of 10 reads/base. Moreover, we calculated the steady state expression levels for the resulting lncRNA genes, as measured either by RNA-seq FPKM in the untreated whole cell, or by combined RNA-seq FPKM of the subcellular fractions in basal and E2-treated conditions (explained in later sections). Together, we collected an expressed set of 1888 lncRNAs genes (lncM) with observed or combined RNA-seq FPKM > 1. Among the lncM genes, 42% (800) of them do not overlap with lncRNA genes annotated previously in RefSeq,

GENCODE, UCSC or lincRNA BodyMap databases (Fig. 3.3.1B) (Harrow, Denoeud et al. 2006; Hsu, Kent et al. 2006; Cabili, Trapnell et al. 2011). They are novel lncRNAs first annotated in the current study.

We focused on lncRNA genes assembled from cytoplasmic RNAs (lncCYTO, Fig. 3.3.2A) for the examination of their gene structures, since they represent more accurate exon-intron calls. Similar to what has been reported in the lncRNA study from the GENCODE project (Derrien, Johnson et al. 2012), the length of their mature transcripts is generally shorter, which can be attributed to the reduced number of exons per transcript (Fig. 3.3.1E). Also consistent with previous reports (Cabili, Trapnell et al. 2011; Derrien, Johnson et al. 2012), the underlying sequences of lncCYTO genes are evolutionarily less conserved compared to their protein-coding counterparts, yet they display local areas of modest conservation, as measured by the phastCons scores, in the exon and promoter regions relative to the intron regions (Fig. 3.3.1F).

LncRNAs are only slightly enriched in the nuclear compartments, and nucleus-retained lncRNAs are less stable at steady state

We compared the transcript abundance in each subcellular fraction of annotated protein-coding mRNAs (codA) and lncRNAs (lncA), and lncM RNAs, and showed that regardless of where they are localized, lncRNAs are expressed at much lower steady-state levels than mRNAs (Fig. 3.3.3A, 3.3.4B). This difference could be partly attributed to the stability of the different types of transcripts. While lncRNAs are transcribed at similar levels, they appear to be less stable than the protein-coding mRNAs (Fig. 3.3.4A, C). We estimated the stability of RNA transcripts from the ratio of steady-state RNA level (RNA-seq FPKM) to nascent transcript level (GRO-seq

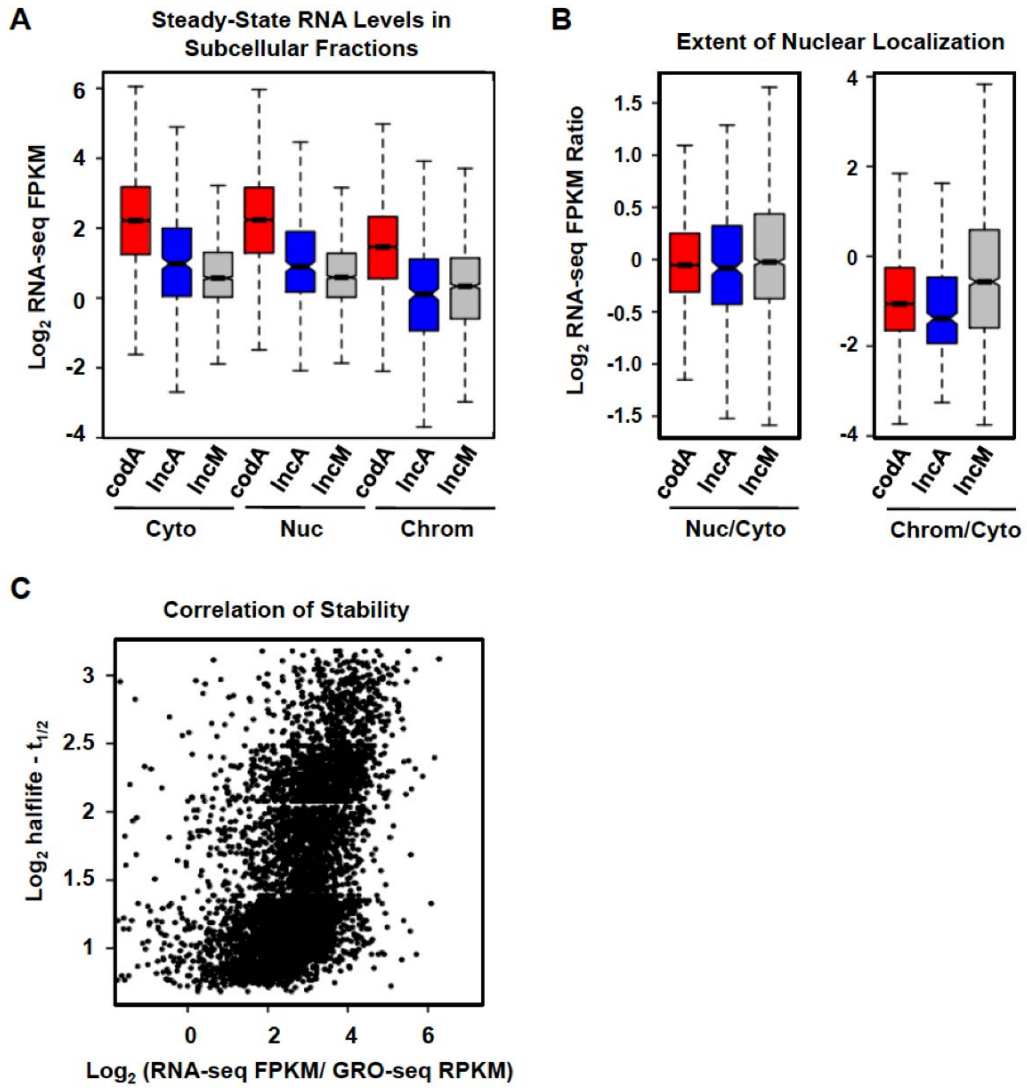


Figure 3.3.3 Subcellular localization of lncRNAs and protein-coding mRNAs.

(A and B) Boxplots show (A) steady-state RNA levels in each of the subcellular fractions, and (B) the extent of nuclear localization, of codA (red), lncA (blue), and lncM transcripts (grey).

(C) Tree diagram shows the significant ontological terms (red) associated with protein-coding mRNAs that are extracted from nuclear fractions.

(D) Correlation of RNA stability measured by two independent approaches. *Tani et. al.* determined the half lives of RNA transcripts using BRIC-seq (vertical axis), and we estimated the stability of RNA transcripts based on the ratio of steady-state RNA levels to nascent transcript levels (horizonal axis).

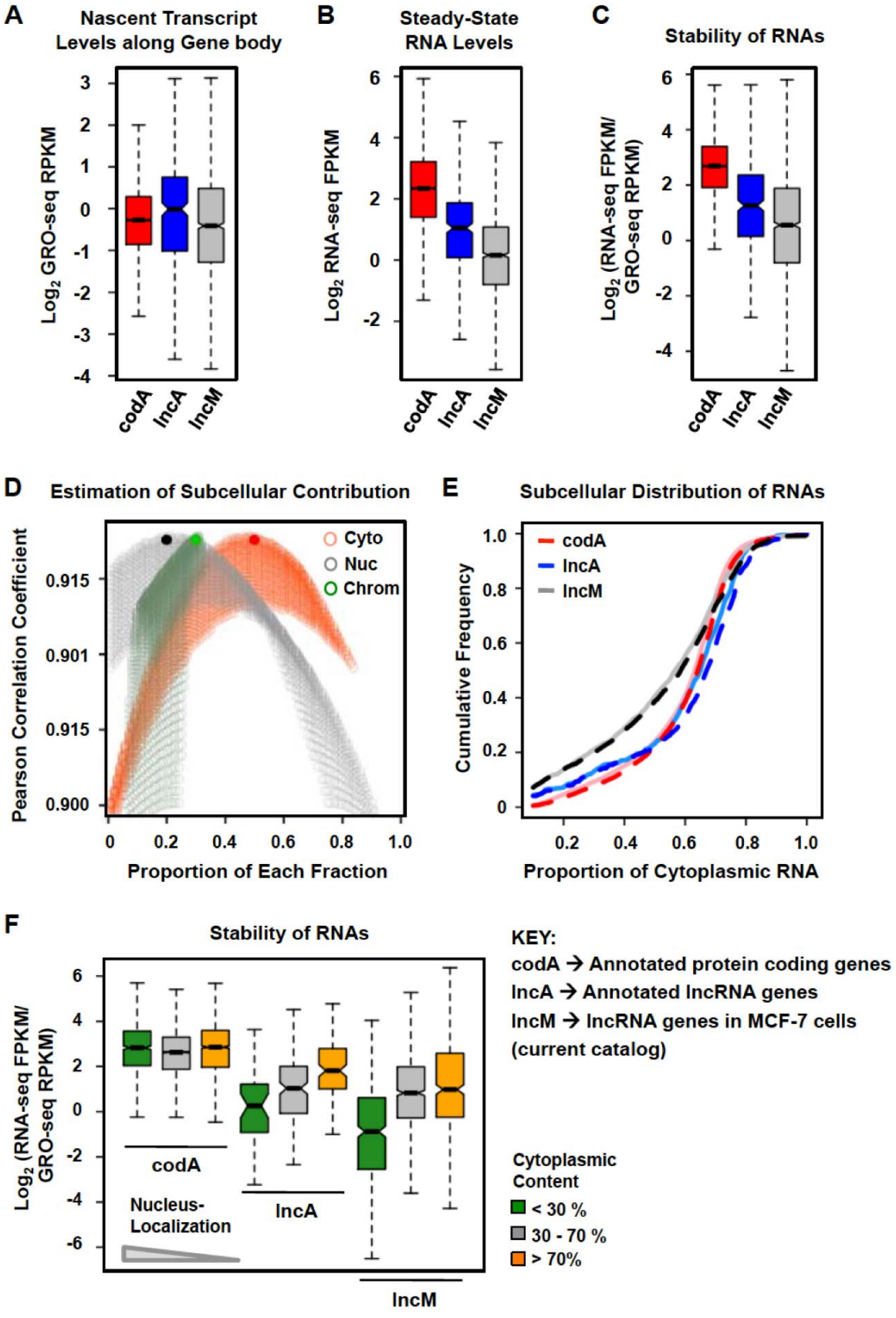
Figure 3.3.4 Nuclear-retained lncRNAs are less stable than cytoplasmic lncRNAs.

(A - C) Nascent transcript levels (A), steady-state RNA levels (B), and stability (C) of annotated protein-coding mRNAs (codA, red), annotated lncRNAs (lncA, blue) and lncM transcripts (grey). Stability of RNAs is measured by the ratio of steady-state RNA levels to nascent transcript levels.

(D) Estimation of the contribution of each subcellular fraction to the total RNA content, based on the relationship a x *Cyto* $+b$ x *Nuc* $+c$ x *Chrom* = *Total*, where *Cyto*, *Nuc*, *Chrom*, and *Total* represent RNA-seq FPKM values from cytoplasmic, nucleoplasmic, chromatin-associated and whole cell samples, and a, b and c indicate their corresponding contributions. Pearson correlation coefficients were calculated and plotted for every pair of a x *Cyto* $+b$ x *Nuc* $+c$ x *Chr* and *Total*, calculated for all assembled transcripts, as we sample the proportion of each fraction, a (orange red), b (grey) and c (green) from 0.01 to 0.99. The combination, a = 0.5, b = 0.2, and c = 0.3 gives the highest correlation and is highlighted in the plot.

(E) Cumulative frequency curves show the extent of cytoplasmic localization of coda (red), lncA (blue) and lncM transcripts (grey and black), in basal (solid) and E2-treated (dotted) conditions.

(F) Boxplot shows the stability of coda (red), lncA (blue) and lncM (grey) transcripts in basal MCF-7 cells, grouped by the extent of cytoplasmic localization.


A
Nascent Transcript Levels along Gene body
Log2 GRO-seq RPKM
codA
lncA
lncM
B
Steady-State RNA Levels
Log2 RNA-seq FPKM
C
Stability of RNAs
Log2 (RNA-seq FPKM/ GRO-seq RPKM)
D
Estimation of Subcellular Contribution
Pearson Correlation Coefficient
Cyto
Nuc
Chrom
Proportion of Each Fraction
E
Subcellular Distribution of RNAs
Cumulative Frequency
Proportion of Cytoplasmic RNA
F
Stability of RNAs
Nucleus-Localization
KEY:
codA → Annotated protein coding genes
lncA → Annotated lncRNA genes
lncM → lncRNA genes in MCF-7 cells (current catalog)
Cytoplasmic Content
< 30 %
30 - 70 %
> 70%

RPKM), which correlates with the half times of RNAs measured by an independent method (Fig. 3.3.3D) (Tani, Mizutani et al. 2012).

Previous literature suggests that lncRNAs predominantly localize to the nucleus, and often get recruited to the chromatin to mediate epigenetic control and transcriptional regulation (Kapranov, Cheng et al. 2007; Mondal, Rasmussen et al. 2010; Derrien, Johnson et al. 2012). Nevertheless, a growing list of lncRNAs is found in the cytoplasm and characterized to be involved in cytoplasmic functions (Willingham, Orth et al. 2005; Gong and Maquat 2011; Carrieri, Cimatti et al. 2012; Clark, Johnston et al. 2012; Yoon, Abdelmohsen et al. 2012; Kretz, Siprashvili et al. 2013). In our pipeline, we are able to detect and annotate the majority of lncM genes, 800/1888, using RNA-seq reads obtained from the cytoplasmic fraction, which again raises the question to what extent lncRNAs are inside the nucleus and associated with the chromatin.

To address this, we displayed the ratios of transcript abundance in the nuclear fractions over the cytoplasm in Fig. 3.3.3B, and observed that there is a slight but significant enrichment of the lncM RNAs, but not the previously annotated lncRNAs, relative to coding mRNAs in the nucleoplasm, and to a greater extent on the chromatin. Moreover, we estimated the contribution of each of the subcellular fractions to the total RNA pool as described in material and methods (Fig. 3.3.4D) to calculate the actual cytoplasmic content of every transcript, and we derived at the same conclusion that a smaller fraction of the lncM RNAs, in comparison to codA and lncA RNAs, are found in the cytoplasm (Fig. 3.3.4E). Nevertheless, our results support that in contrast to the traditional view, lncRNAs are only slightly enriched in the nuclear compartments.

When the codA, lncA and lncM transcripts were stratified into nuclear (< 30% cytoplasmic), intermediate, and cytoplasmic (>70% cytoplasmic), and their corresponding stability is displayed (Fig. 3.3.4F), we observed an interesting pattern that the nuclear lncRNAs are significantly less stable than the cytoplasmic lncRNAs. Indeed, a recent study that measured the half-life of about 100 mouse lncRNAs in a "stability microarray" has reached the same conclusion (Clark, Johnston et al. 2012). The authors have further speculated that this subset lncRNAs is turned over very rapidly to facilitate the dynamic regulation of cellular process in the nucleus. Moreover, our findings suggest that the lower stability of nuclear RNAs could be the reason for them to evade earlier annotation, and our approach takes advantage of its improved sensitivity in all subcellular fractions and allows for the identification of the less stable nuclear lncRNAs.

Divergent and antisense lncRNAs are highly transcribed and contribute predominantly to the chromatin signatures associated with lncRNAs

Historically, researchers have focused on lncRNAs that are well separated from existing coding loci, collectively known as long intergenic noncoding RNAs, or lincRNAs (Guttman, Amit et al. 2009; Khalil, Guttman et al. 2009; Guttman, Garber et al. 2010; Cabili, Trapnell et al. 2011; Guttman, Donaghey et al. 2011; Ulitsky, Shkumatava et al. 2011). Nevertheless, the list of characterized natural antisense lncRNAs, those transcribed from the antisense strand and overlap in part with well-defined sense RNA transcripts, has been growing, some of them are involved in important cellular functions (Krystal, Armstrong et al. 1990; Munroe and Lazar 1991; Yan, Hong et al. 2005; Beltran, Puig et al. 2008; Carrieri, Cimatti et al. 2012). In addition, the phenomenon of divergent (bidirectional) transcription commonly occurs across the mammalian genome,

producing unstable transcripts that are likely lncRNAs (Core, Waterfall et al. 2008; Preker, Nielsen et al. 2008; Hah, Danko et al. 2011). Indeed, divergently transcribed lncRNA/mRNA gene pairs have been described in gene-specific and genomic studies (Rinn, Kertesz et al. 2007; Wang, Yang et al. 2011), the latter reporting varying degree of dominance as a group (Cabili, Trapnell et al. 2011; Sigova, Mullen et al. 2013). When we examined the distribution of lncM genes along the genome, we observed that about 21% of them are transcribed from the antisense strand, and overlap either the promoter (243 divergent lncRNAs) or the gene body (159 antisense lncRNAs) of a sense protein-coding mRNA or another lncRNA (Fig. 3.3.5A). To evaluate whether lncRNAs as a group favors the divergent and antisense arrangements, we examined the genomic distribution of codA genes. We found that the proportion of codA genes that overlap another coding locus, 15.1% divergent and 6.3% antisense, adds up to a similar fraction (data not shown). It suggests that similar to protein-coding genes, a considerable proportion of lncRNA genes originate from the genic regions.

Moreover, we examined whether these genic lncRNA genes display any specific features that distinguish them from intergenic lncRNAs. Indeed, while the steady-state levels of lncRNAs appear comparable regardless of their gene location, divergent and antisense lncRNAs are more highly transcribed at the TSS and across the gene body as measured by GRO-seq (Fig. 3.3.5C, D), and it is likely contributed by their sense mRNA partner, which are often associated with the production of divergent transcripts even when an overlapping lncRNA is absent. Correspondingly, genic lncRNAs display much higher levels of H3K4me3, which marks active promoters, and H3K36me3, which is associated with actively elongating polymerases along the gene body than intergenic lncRNAs (Fig. 3.3.5C). Chromatin signatures are not strand specific. Therefore, most of the divergent RNA pairs share the same H3K4me3 domain, and antisense

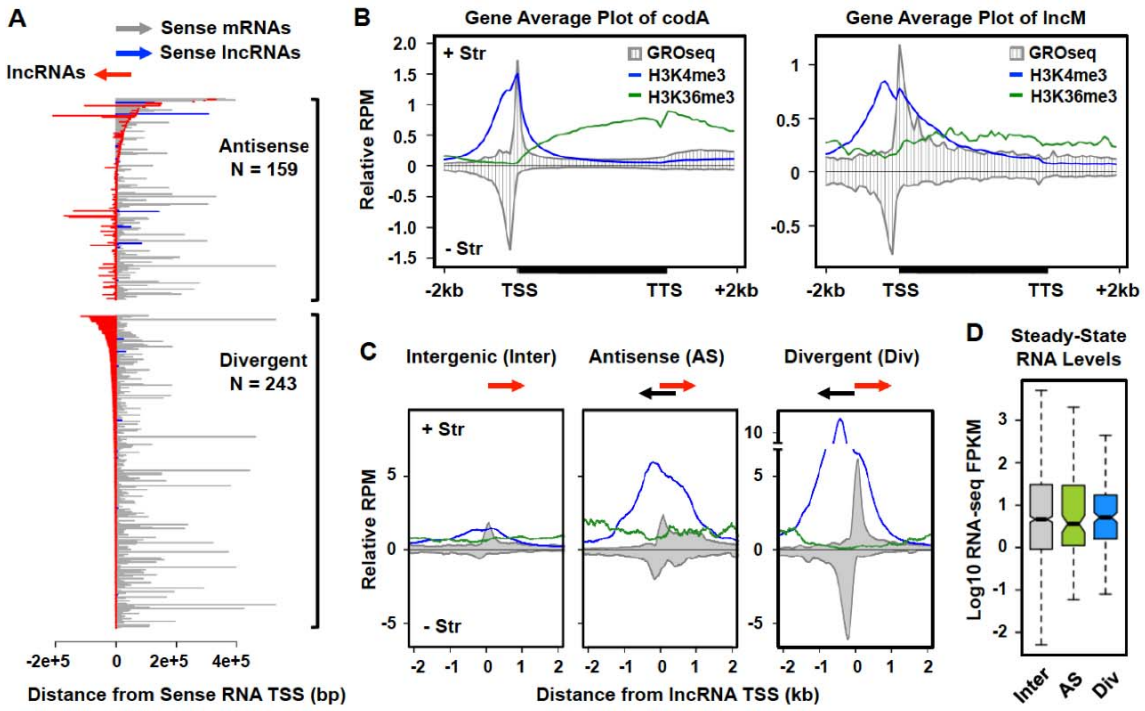


Figure 3.3.5 Divergent and Antisense lncRNA genes are associated with higher levels of transcriptional activity and chromatin signatures.

(A) Graphical representation of the orientation, position and length of antisense (top) and divergent (bottom) lncRNA genes (red) relative to their sense RNA genes (grey or blue).

(B) Average profiles of GRO-seq reads (grey), and ChIP-seq reads of H3K4me3 (blue) and H3K36me3 (green), for codA (left) and lncM (right) genes. All gene bodies are scaled to 4 kb. ChIP-seq RPM of H3K36me3 are scaled to 3x its original value in order to show it clearly on the same plot.

(C) Similar to (B), average profiles of GRO-seq reads, and ChIP-seq reads of H3K4me3 and H3K36me3, centered on the TSS of intergenic (Inter), antisense (AS), and divergent (Div) lncM genes.

(D) Boxplot shows the steady-state RNA levels of Inter, AS and Div lncM genes.

lncRNA/mRNA pairs share at least a fraction of the H3K36me3 domain. The average gene profiles of transcription activity (GRO-seq) and chromatin signatures (H3K4me3 and H3K36me3) for lncM genes and codA genes share many similarities (Fig. 3.3.5B). Nevertheless, sense mRNAs with divergent or antisense lncRNAs on their opposite strand, likely contribute to the overall profile, and intergenic lncRNAs that are devoid of the influence of mRNAs show markedly lower level of transcriptional activity and its associated chromatin marks.

LncRNAs display lower levels of promoter H3K4me3 and gene body H3K36me3

Among the methods for the identification of lncRNA genes, the presence of H3K4me3-H3K36me3 domain demarcates the location of the transcription units and is frequently used in previous studies (Guttman, Amit et al. 2009; Ulitsky, Shkumatava et al. 2011). Nevertheless, we observed very low levels of these marks associated with intergenic lncRNAs, where the influences of protein-coding genes are absent. It suggests that lncRNAs may be by nature displaying lower levels of these chromatin marks and hence questions the applicability of using these marks in identifying lncRNAs.

To address this possibility, we focused on intergenic lncM genes and examined the levels of the associated H3K4me3 and H3K36me3 marks in comparison to codA genes that do not overlap with any antisense or divergent gene loci. Even when we controlled for the level of transcriptional activity, sampling codA genes that are transcribed at similar levels as lncM genes, they showed significantly lower levels of H3K4me3 at the TSS and H3K36me3 along the gene body (Fig. 3.3.6A). We then searched for additional properties that differs codA genes from lncM genes, such as their higher steady-state RNA levels, and asked whether it may be correlated with the observed differences in the levels of histone marks. Nevertheless, when we

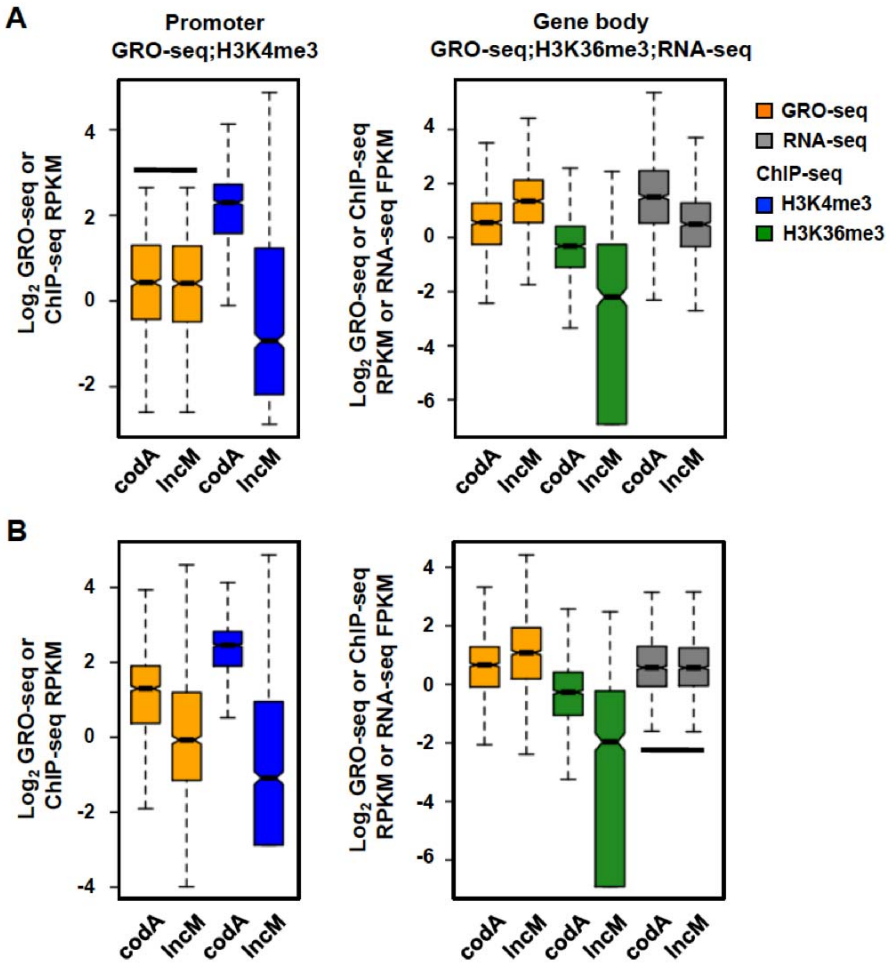


Figure 3.3.6 Intergenic lncRNA genes display significantly lower levels of H3K4me3 at the promoter and H3K36me3 along the gene body than equally expressed protein-coding genes.

(A and B) Boxplots comparing the levels of GRO-seq (orange) and H3K4me3 (blue) at the promoter, GRO-seq (orange) and H3K36me3 (green) along the gene body, and the steady-state RNA levels (grey) of selected codA genes with intergenic lncM genes. We sampled intergenic codA genes that have either the same distribution of GRO-seq levels at the promoter (A), or the same distribution of steady-state RNA levels (B), as indicated by the black bars.

controlled for the level of steady-state gene expression, lncM genes as a group still show much lower chromatin signatures than the sampled codA genes (Fig. 3.3.6B). In addition, we performed a similar analysis comparing coda genes with annotated intergenic lncRNAs (lncA), and the conclusion holds true (Fig. 3.3.7). We have also evaluated the contribution of the length of the transcript, the length of coding sequences and the number exons to the levels of H3K4me3 and H3K36me3 marks associated with protein-coding genes, and it is apparent that these factors cannot sufficiently explain the significant differences in histone marks as we compare lncRNA to protein-coding genes (Fig. 3.3.8).

In conclusion, intergenic lncRNAs that are devoid of the influences of coding mRNAs display lower levels promoter H3K4me3 and gene body H3K36me3, suggesting that using these marks to define the transcription units of intergenic lncRNAs likely suffer from this inherent limitation.

ERα localizes proximal to the promoters of E2-upregulated lncRNA genes, some of them are associated with an elevated level of enhancer features.

It has been recently reported that E2 stimulation regulates an under-appreciated large fraction of the whole transcriptome, likely including the lncRNAs (Hah, Danko et al. 2011). Indeed, in the current study, we observed that more than a quarter (531 lncRNA genes, 28.1%) of the lncM genes are statistically regulated by E2. By comparing the regulation calls based on GRO-seq and RNA-seq datasets, we are able to distinguish regulation that occurs transcriptionally or post-transcriptionally (Fig. 3.3.9). At steady-state level, 158 lncRNA genes are upregulated, and 164 are downregulated. While half of the upregulated lncRNAs and a third of the downregulated lncRNAs show corresponding changes at nascent transcript level (Fig.

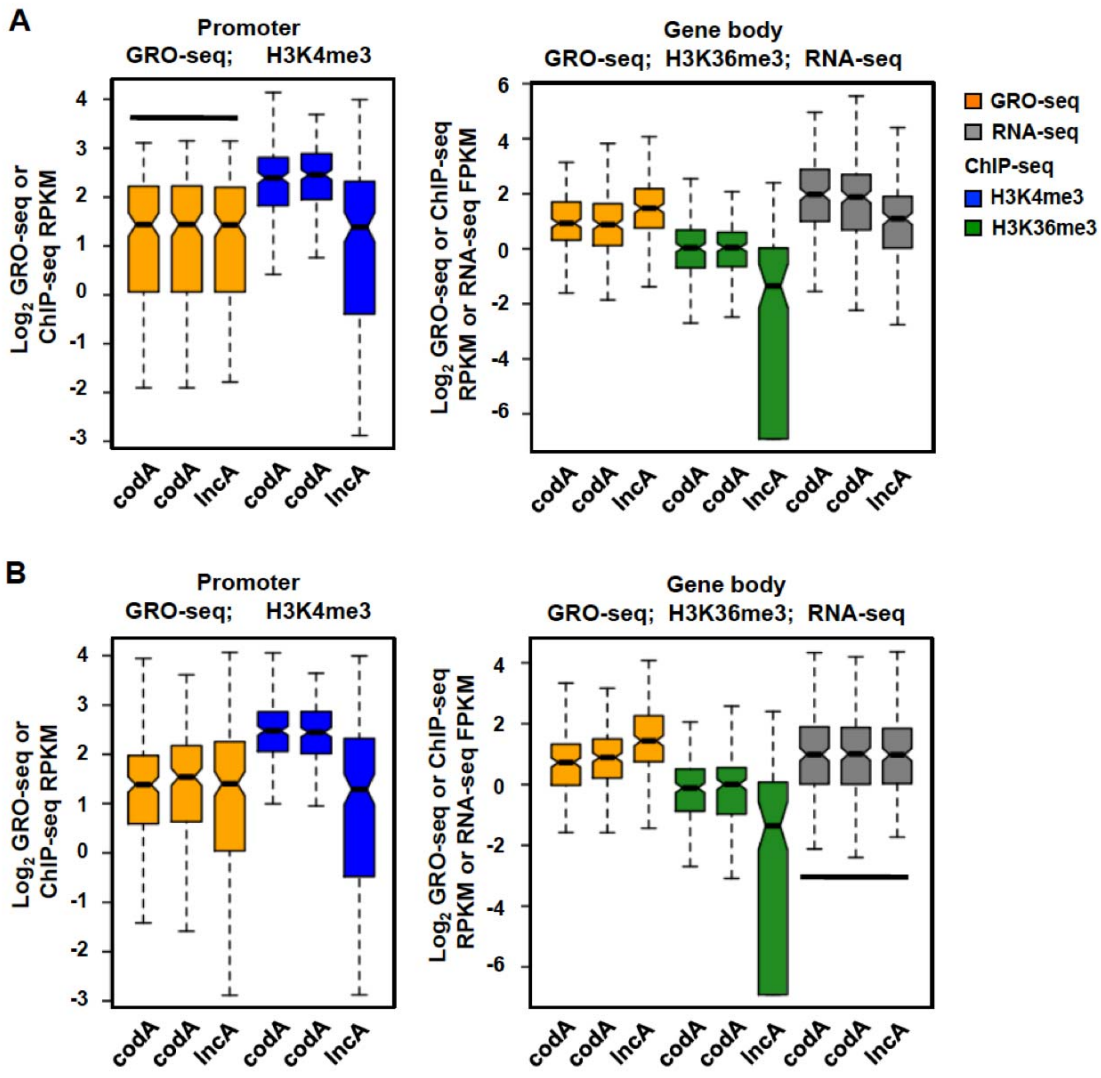
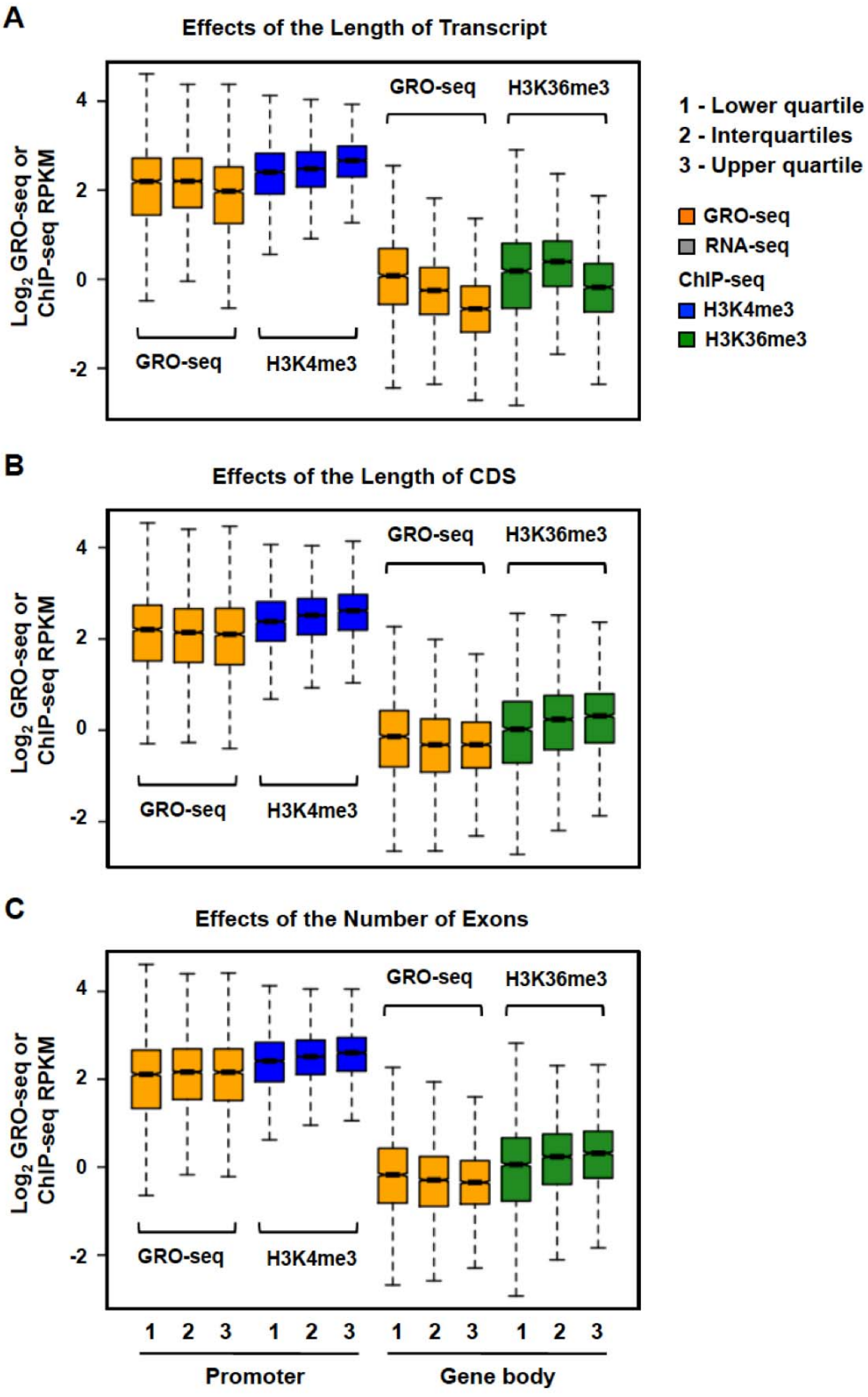


Figure 3.3.7 Intergenic lncA genes display significantly lower levels of H3K4me3 at the promoter and H3K36me3 along the gene body than equally expressed protein-coding genes.

(A and B) Boxplots similar to Fig. 3.3.5, but comparing the levels of nascent transcription and chromatin signatures of selected codA genes with intergenic lncA genes. Two non-overlapping groups of codA genes were sampled, each having either the same distribution of GRO-seq levels at the promoter (A), or the same distribution of steady-state RNA levels with lncA genes (B).

Figure 3.3.8 Chromatin signatures of protein-coding genes are only minutely affected by the length of transcript, length of coding sequence (CDS) and the number of exons.

(A - C) Boxplots show the levels of nascent transcription (GRO-seq) and chromatin signatures (H3K4me3, H3K36me3) of all codA genes, separated into lower quartile, interquartiles and upper quartile, based on the length of transcript (A), length of CDS (B), and number of exons (C).


A
Effects of the Length of Transcript
Log2 GRO-seq or ChIP-seq RPKM
4
2
0
-2
GRO-seq
H3K36me3
GRO-seq
H3K4me3
1 - Lower quartile
2 - Interquartiles
3 - Upper quartile
GRO-seq
RNA-seq
ChIP-seq
H3K4me3
H3K36me3
B
Effects of the Length of CDS
Log2 GRO-seq or ChIP-seq RPKM
4
2
0
-2
GRO-seq
H3K36me3
GRO-seq
H3K4me3
C
Effects of the Number of Exons
Log2 GRO-seq or ChIP-seq RPKM
4
2
0
-2
GRO-seq
H3K36me3
GRO-seq
H3K4me3
1 2 3 1 2 3 1 2 3 1 2 3
Promoter
Gene body

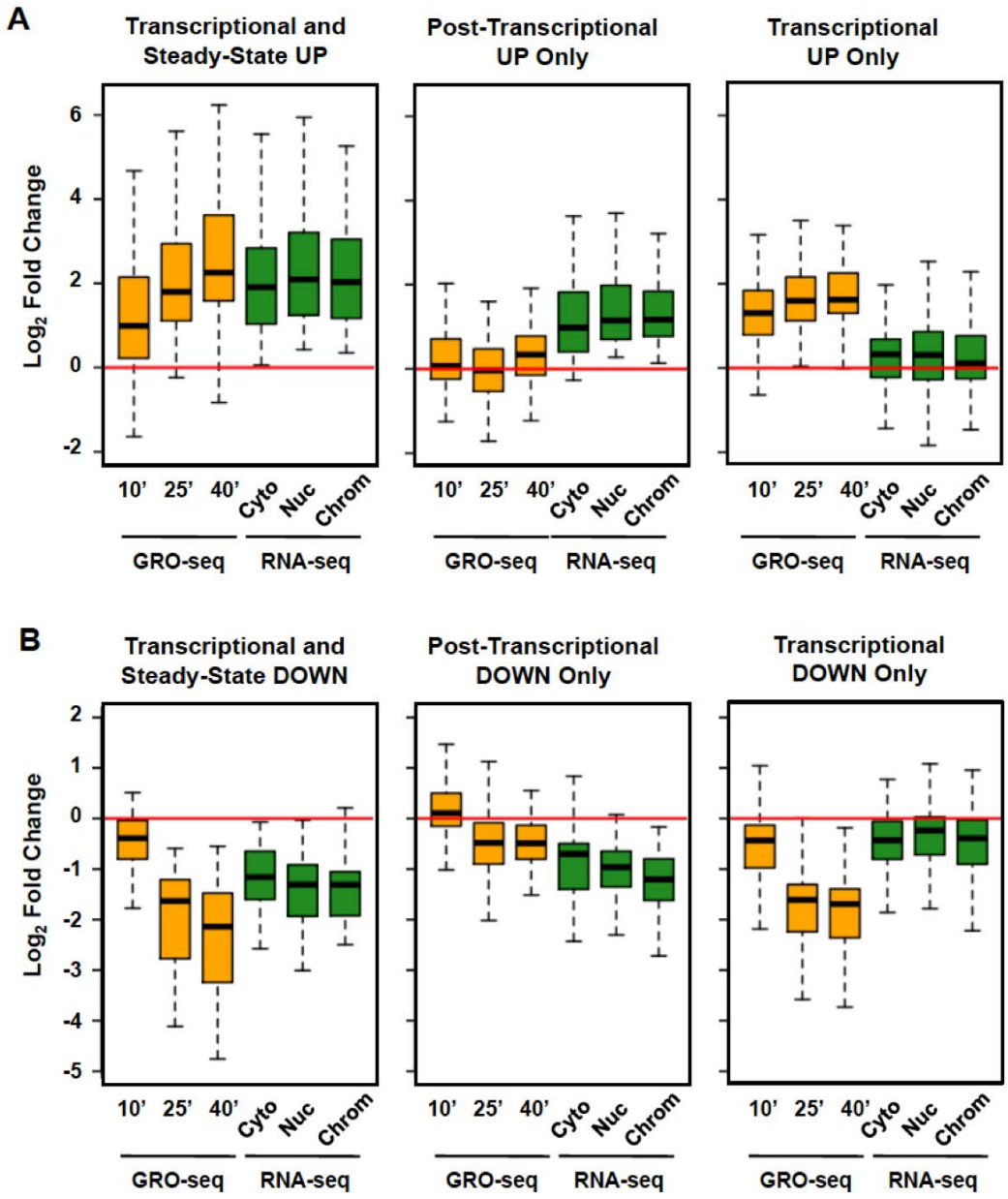


Figure 3.3.9 LncRNA genes are regulated by E2 transcriptionally and post-transcriptionally.

(A and B) Boxplots showing E2-induced fold changes of lncM genes that are upregulated (A), and downregulated (B), at both transcriptional and steady-state RNA levels (left), at steady state only (middle), and at transcriptional level only (right). Red lines indicate the level where there is no fold change.

3.3.9A, B, left), the remaining ones are probably regulated by E2 post-transcriptionally (Fig. 3.3.9A, B, right). On the other hand, there are also 375 lncRNA genes that are E2-regulated only at the transcriptional level. 66 of them are E2-induced (Fig. 3.3.9A, middle) and 209 of them are E2-repressed (Fig. 3.3.9B, middle), and they do not show statistically significant changes at steady-state level. Perhaps not surprisingly, lncRNAs with coordinated regulation at both levels are also associated with the highest degree of regulation, either up or down (Fig. 3.3.9A,B).

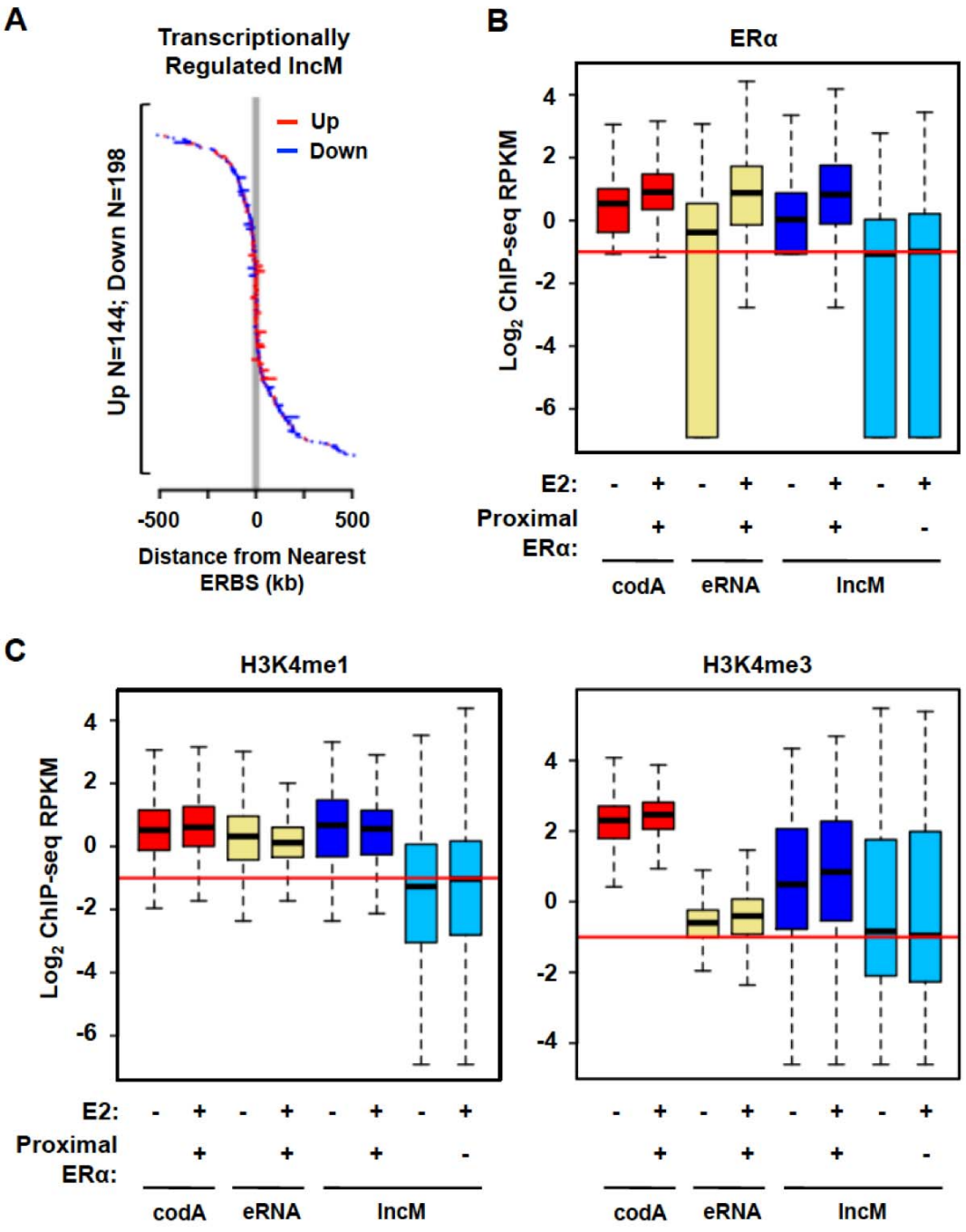
Similar to protein-coding genes regulated by E2 at the transcriptional level, many E2-regulated lncM genes show E2-induced ERα binding at their promoters as well. We measured the distance from the TSS of transcriptionally regulated lncRNA genes to their nearest ERα binding sites (ERBS) in the E2-treated condition, and observed a distinct segregation of E2-upregulated lncRNA genes in close proximity to ERBS (Fig. 3.3.10A), suggesting that ERα is involved in the E2-induced transcriptional upregulation of these lncRNA genes.

Moreover, lncRNA genes with ERα binding at their promoters remind us of a recent report on ERα enhancer RNAs (eRNAs) (Hah, Murakami et al. 2013), which are transiently-expressed, short RNAs transcribed proximal to an ERBS. In addition, reports from the Shiekhattar lab have proposed an enhancer function for some of the lncRNAs, collectively known as ncRNA-activating (Orom, Derrien et al. 2010). Together, it raises the question of whether lncRNAs and eRNAs are distinct concepts. Therefore, we closely examined the profile of H3K4me1, a commonly used histone mark to define enhancers, and of H3K4me3, a mark of active gene promoter, at the promoter of lncM genes, in comparison to previously defined eRNAs (Hah, Murakami et al. 2013). We observed that lncM genes that display promoter proximal ERα binding in the E2-treated condition are associated with a comparable level of H3K4me1 to the promoter of eRNAs (Fig. 3.3.10B, C). When we compared the profiles of

Figure 3.3.10 ERα localizes to the promoters of a subset of lncRNA genes, which are associated with an elevated level of enhancer features.

(A) Graphical representation of the length and distance from the nearest ERα-binding site (ERBS) of transcriptional regulated lncM genes.

(B) Boxplots comparing the levels of ERα, as well as enhancer (H3K4me1) and promoter signatures (H3K4me3) at promoter regions of codA genes (red) and eRNAs (yellow) with promoter proximal ERα binding, and of lncM genes with (blue) and without proximal ERα binding (light blue).


A
Transcriptionally Regulated lncM
Up
Down
Up N=144; Down N=198
-500
0
500
Distance from Nearest ERBS (kb)
B
ERα
Log2 ChIP-seq RPKM
E2:
Proximal ERα:
codA
eRNA
lncM
C
H3K4me1
H3K4me3
Log2 ChIP-seq RPKM
E2:
Proximal ERα:
codA
eRNA
lncM

promoter signature, ERα-bound lncM genes show an intermediate level of H3K4me3 in comparison to ERα-bound eRNAs and protein-coding genes (Fig. 3.3.10C). Therefore, a subset of lncRNA genes are associated with promoter proximal ERα binding and an elevated level of enhancer features.

LncRNAs carry important cellular information and the cell type-specific expression of lncRNAs predict the intrinsic molecular subtype of breast cancer cells

As we examined the differential expression of lncRNA and protein-coding genes across a panel of 304 samples encompassing a whole spectrum of cancer and normal tissue samples and cell lines, our results agreed with a previous report (Cabili, Trapnell et al. 2011) that lncRNAs show more tissue- and cell type-specific expression than protein-coding mRNAs (Fig. 3.3.11A). The differential expression of protein-coding genes has been reasonably explored by researchers for its utility in the clinics as diagnostic and prognostic biomarkers (Alizadeh, Eisen et al. 2000; Perou, Sorlie et al. 2000; Nielsen, West et al. 2002), such as in the case of using PAM50 to subtype breast cancers into their intrinsic molecular subtypes that are associated with different prognostic outcomes (Parker, Mullins et al. 2009). The expanding landscape of lncRNAs is a relatively untouched ground, and their tissue- and cell type-specific expression patterns make them highly attractive as new and alternative biomarkers. Indeed, the differential expression of lncRNAs allowed them to accurately cluster the hundreds of samples into their respective tissue types in an unsupervised manner (Fig. 3.3.11B). More interestingly, they are able predict the intrinsic molecular subtypes of a panel of 45 human breast cancer cell lines, to a similar accuracy as the performance of protein-coding genes (Fig. 3.3.11C), suggesting that they carry a similar amount of important cellular information. Therefore, lncRNAs, or a selected subset of them, or

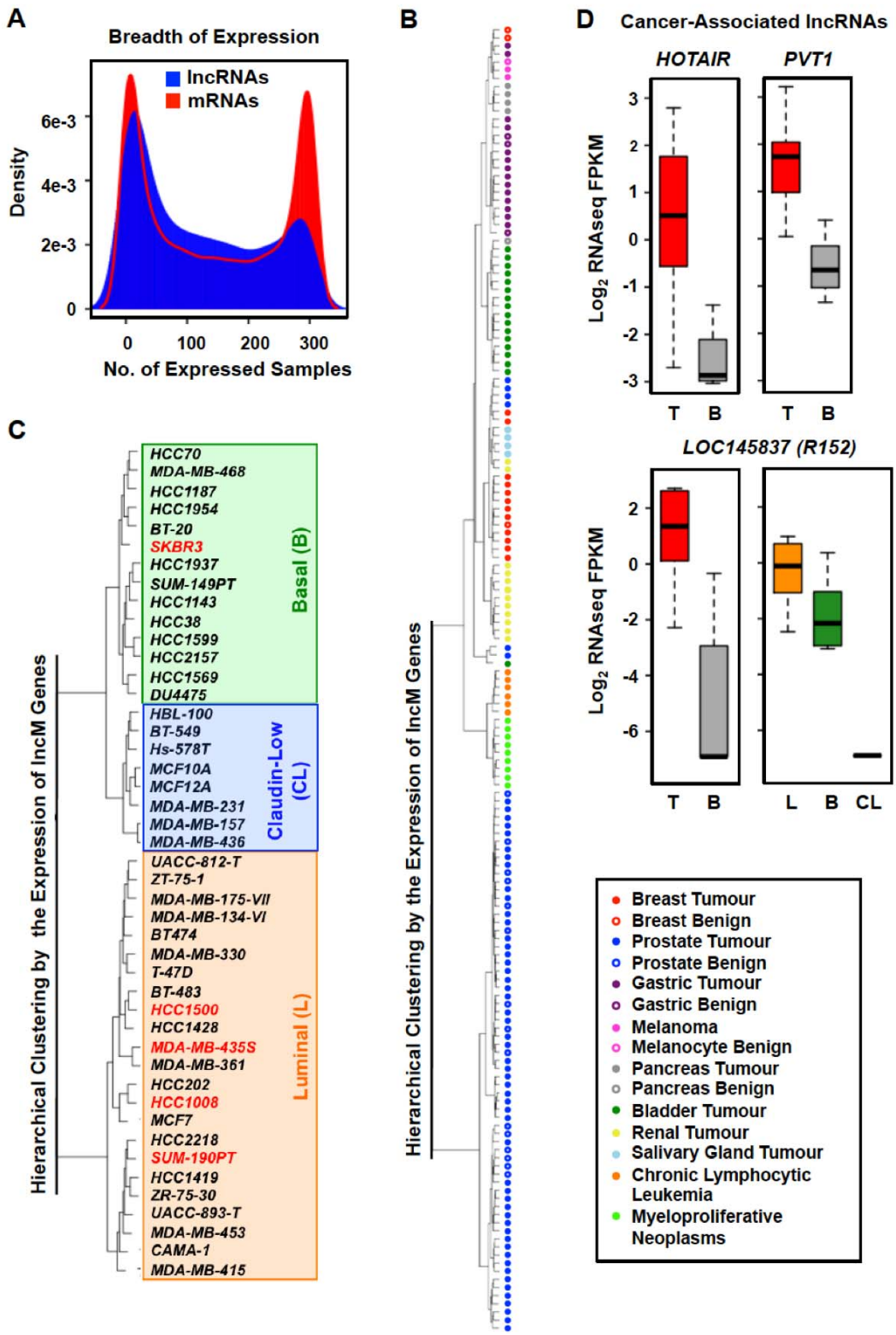
Figure 3.3.11 Tissue- and cell type-specific expression of lncRNA genes informs tissue identity and predicts the intrinsic molecular subtype of breast cancer cells.

(A) Density plot shows the breadth of expression of lncM (blue) and protein-coding genes (red) across a panel of 304 tissue samples and cell types.

(B) Hierarchical clustering of 150 tissue samples from ten different tissues/organs of origin, including tumor (solid circle) and benign (open circle) samples, based on the differential expression of lncM genes.

(C) Hierarchical clustering of 45 breast cancer cell lines into their instinct molecular subtypes, basal (B), claudin-low (CL), and luminal (L), based on the differential expression of lncM genes. Mis-classified cell lines are shown in red.

(D) Boxplots show examples of lncRNA genes with elevated expression in breast tumors (T, red) in comparison to benign breast tissues (B, grey). *HOTAIR* and *PVT1* are previously characterized lncRNAs that have been implicated in breast cancer. LOC145837 (R152) is an uncharacterized lncRNA that we picked for subsequent study. The expression of R152 (LOC145837) is cancer-specific and subtype-specific, showing elevated levels in luminal (L, orange) and basal (B, green) breast cancer cells.


A
Breadth of Expression
lncRNAs
mRNAs
Density
6e-3
4e-3
2e-3
0
0
100
200
300
No. of Expressed Samples
B
Hierarchical Clustering by the Expression of lncM Genes
C
Hierarchical Clustering by the Expression of lncM Genes
HCC70
MDA-MB-468
HCC1187
HCC1954
BT-20
SKBR3
HCC1937
SUM-149PT
HCC1143
HCC38
HCC1599
HCC2157
HCC1569
DU4475
Basal (B)
HBL-100
BT-549
Hs-578T
MCF10A
MCF12A
MDA-MB-231
MDA-MB-157
MDA-MB-436
Claudin-Low (CL)
UACC-812-T
ZT-75-1
MDA-MB-175-VII
MDA-MB-134-VI
BT474
MDA-MB-330
T-47D
BT-483
HCC1500
HCC1428
MDA-MB-435S
MDA-MB-361
HCC202
HCC1008
MCF7
HCC2218
SUM-190PT
HCC1419
ZR-75-30
UACC-893-T
MDA-MB-453
CAMA-1
MDA-MB-415
Luminal (L)
D
Cancer-Associated lncRNAs
HOTAIR
PVT1
Log2 RNAseq FPKM
3
2
1
0
-1
-2
-3
T
B
T
B
LOC145837 (R152)
Log2 RNAseq FPKM
2
0
-2
-4
-6
T
B
L
B
CL
Breast Tumour
Breast Benign
Prostate Tumour
Prostate Benign
Gastric Tumour
Gastric Benign
Melanoma
Melanocyte Benign
Pancreas Tumour
Pancreas Benign
Bladder Tumour
Renal Tumour
Salivary Gland Tumour
Chronic Lymphocytic Leukemia
Myeloproliferative Neoplasms

the combined use of selected lncRNAs together with additional protein-coding genes, could potentially serve as new and improved prognostic biomarkers in predicting the treatment outcomes for breast cancer.

LncRNAs are required for the normal growth of MCF-7 human breast cancer cells

We have identified and annotated a comprehensive catalog of lncRNA genes in MCF-7 breast cancer cells, and close examination of these lncRNAs revealed many interesting features. Moreover, we want to characterize the molecular and functional roles of lncRNAs in human breast cancer cells, so as to fully appreciate their implications in breast cancer.

As mentioned earlier, we evaluated the differential expression of lncRNAs across a whole panel of cancer and normal tissue samples and cell lines, including 12 breast tumor samples and 11 benign breast samples. It is reasonable to hypothesize that lncRNAs with elevated expression in breast tumors are more likely implicated in breast cancer. Indeed, when comparing the expression of each lncRNA across breast tumor samples in comparison to that across benign samples, and selecting for ones with cancer-specific expression, *HOTAIR* and *PVT1*, two previously characterized lncRNAs that are associated with the development of breast cancer, are among the top hits (Fig. 3.3.11D, top). Therefore, we returned to the list of lncRNAs with breast cancer-specific expression, and focused on R152 (LOC145837), a previously uncharacterized lncRNA that shows luminal- and basal-specific expression in human breast cancer cells (Fig. 3.3.11D, bottom), and asked if it also mediate important functional roles in MCF-7 cells, a luminal breast cancer cell line. Indeed, the expression of R152 is required for the normal growth of MCF-7 cells, validated using two independent siRNA oligos (Fig. 3.3.12B).

Figure 3.3.12 LncRNAs are required for the normal growth of MCF-7 breast cancer cells.

(A) List of the top 10 REACTOME pathways from the "guilt-by-association" analysis performed on lncRNA R67, as described in materials and methods.

(B) siRNA-mediated knockdown of lncRNA R152, using two independent siRNA oligos, si152-1 (red), and si152-2 (green). *(Left)* Analysis of R152 expression by RT-qPCR. β-action was used as an internal control. Each bar represents mean + SEM, n=3. *(Right)* Analysis of basal cell growth in control (siCtrl) and R152 knockdown MCF-7 cells, over a course of 6 days post-transfection of siRNAs. Each point represents mean ± SEM, n=3.

(C) Similar to (B), siRNA-mediated knockdown of R67 by two independent oligos, si67-1 (orange), and si67-2 (blue), and its effects on basal cell growth in MCF-7 cells.

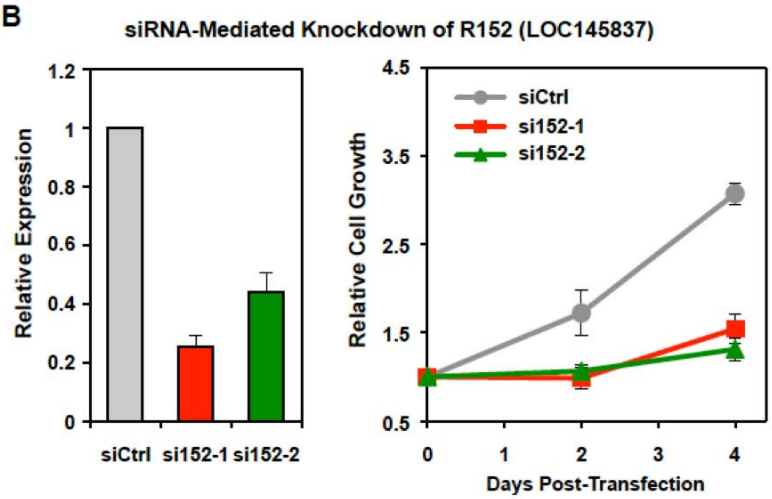
A

Guilt-by-Association Analysis on R67

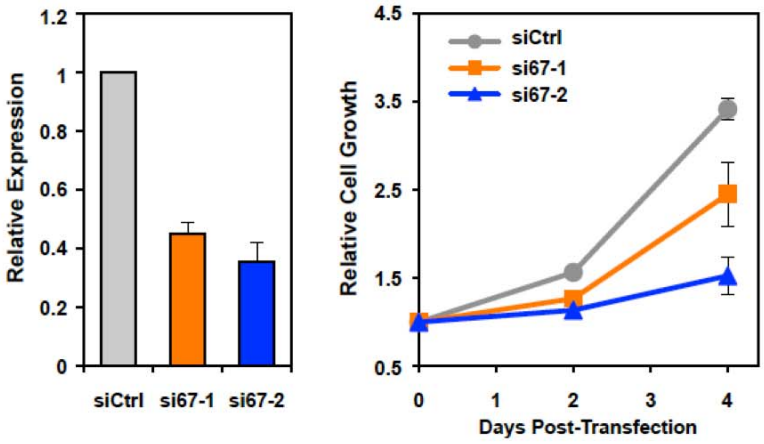
- Top 10 REACTOME Pathways

REACTOME_MITOTIC_M_M_G1_PHASES
REACTOME_CELL_CYCLE_MITOTIC
REACTOME_DNA_REPLICATION
REACTOME_CELL_CYCLE_CHECKPOINTS
REACTOME_MITOTIC_PROMETAPHASE
REACTOME_REGULATION_OF_MITOTIC_CELL_CYCLE
REACTOME_CELL_CYCLE
REACTOME_APC_C_CDC20_MEDIATED_DEGRADATION_OF_MITOTIC_PROTEINS
REACTOME_METABOLISM_OF_NON_CODING_RNA
REACTOME_MITOTIC_G1_G1_S_PHASES

B



C



In addition, we searched for genomic methods that would provide some clues to the possible functions of more lncRNAs. One of the more commonly used ones is the “guilt-by-association” approach as described in materials and methods. In brief, it relies on the correlation relationships between uncharacterized lncRNAs and protein-coding genes of known function, hence linking the functional pathways associated with the most highly correlated proteins to the lncRNA of interest. Therefore, we performed the “guilt-by-association” analysis on lncRNAs and selected candidates that are associated with important cell viability pathways. For example, Fig. 3.3.12A lists the top 10 REACTOME pathways that are most enriched in the analysis performed on the lncRNA R67, most of these pathways are related to the control of cell cycle. Consistently, we observed a corresponding inhibition of cell growth as we knocked down the expression of R67 in MCF-7 cells using two independent siRNA oligos (Fig. 3.3.12C).

Therefore, we have shown two examples of lncRNAs, R152 and R67, that are required for the normal growth of MCF-7 cells. Future studies will reveal their functional interplay with the estrogen signaling pathway in E2-responsive breast cancer cells, whether they are also involved in cell growth responses in other breast cancer cell lines, and the molecular mechanisms underlying the observed inhibitory effects on cell growth, so as to ultimately establish a role for these lncRNAs in breast cancer biology and to raise the possibility of using lncRNAs as new therapeutic targets.

3.4. Discussion

In this study, we have generated the most comprehensive catalog of lncRNAs in MCF-7 human breast cancer cells, capturing lncRNAs that are expressed across an estrogen treatment time course, and localize to all three subcellular compartments. In addition to previously

annotated lncRNAs, we have identified ~800 novel lncRNA genes expressed in MCF-7 cells. Moreover, we annotated the exon-intron structure of these lncRNAs with the best-available knowledge, giving preference to transcript assembly performed on cytoplasmic RNAs. We then examined all lncRNAs (lncM) as a group, and revealed many interesting findings that are relevant for future efforts in the identification and annotation of lncRNAs.

Our results have raised concerns for the utility of H3K4me3-H3K36me3 domains in the discovery of lncRNAs, as we observed that majority of lncRNA genes are not associated with high levels of H3K4me3 and H3K36me3. The lower than expected level of chromatin signatures cannot be fully explained by the level of transcription, level of steady-state RNA, length of transcript and the number of exons. Therefore, using H3K4me3-H3K36me3 to define the genomic location of lncRNA genes will likely suffer from low sensitivity, and will be biased towards the discovery of divergent lncRNAs, which likely carry over some influences from the sense mRNAs. On the other hand, we have demonstrated the better performance of using GRO-seq results to capture the location and activity of the primary transcripts of lncRNA genes. In many cases, even with lower steady-state RNA level, and lower chromatin signatures, most lncRNAs show comparable nascent transcript level as measured by GRO-seq in comparison to their protein-coding counterparts.

In addition, contrary to common belief, lncRNAs are trafficked to all subcellular compartments, and there is only a slight enrichment in the nucleus. Nevertheless, others and we have independently demonstrated that lncRNAs in the nucleus tend to be less stable (Clark, Johnston et al. 2012), and are more likely to evade detection. Therefore, efforts should not be focused solely on nucleus-localized lncRNAs and their associated nuclear functions, as cytoplasmic lncRNAs form the other abundant class with a whole spectrum of new possibilities,

as in the case of lincRNA-p21 and TINCR (Yoon, Abdelmohsen et al. 2012; Kretz, Siprashvili et al. 2013). Nevertheless, it is reasonable to come up with more sensitive ways to detect lncRNAs in the nucleus, and the use of fractionated RNA-seq in the current study is one such example, as there may exist transiently expressed nuclear lncRNAs that are involved in the dynamic regulation of time-sensitive events, but have evaded earlier detection efforts due to their transient nature.

Furthermore, the observed subcellular localization of lncRNAs raises two interesting questions for future study. (1) Trafficking of proteins can be driven by peptide signals, such as a nuclear localization signal, or NLS, which is encoded in the amino acid sequence of the trafficked proteins. What drives the subcellular localization of lncRNAs? (2) The observed lower stability of lncRNAs and the even lower stability associated with the nuclear lncRNAs are properties that are distinct from protein-coding mRNAs. It suggests a different mechanism in the regulation of stability and degradation. Which factors are involved in this lncRNA-specific regulatory pathway? Answers to these questions will expand our knowledge in the biochemistry and molecular biology of lncRNAs, and in RNA biology in general.

Moreover, we have identified a significant fraction (28.1%) of lncRNA genes that are regulated by E2. The estrogen signaling pathway gets initiated, and triggers a profound and coordinated network of gene regulation involving almost all types of RNA transcripts (Hah, Danko et al. 2011), both transcriptionally and post-transcriptionally. Here, we are able to distinguish lncRNAs that are only regulated at the transcriptional level from those that are regulated by steady-state level. In a model that E2-regulated lncRNAs are integral players in the estrogen signaling pathway, we can speculate a scenario where some of them respond to E2 stimulation and result in altered levels of steady-state RNAs to directly modulate E2-dependent

cellular outcomes, while others are only transcriptionally regulated and may act as enhancers to coordinate the steady-state expression of other lncRNAs and mRNAs that may directly impact in the estrogen signaling pathway.

Indeed, lncRNA genes with ERα binding at their promoters show an elevated level of enhancer mark, H3K4me1, which is comparable to the H3K4me1 profile of the recently characterized eRNAs (Hah and Murakami et al. 2013). Perhaps, a subset of these lncRNAs originate from a distinct class of enhancers, and form a distinct class of eRNAs, to perform enhancer functions as suggested by Lai and colleagues (Lai, Orom et al. 2013). Whether this model holds true will be the subject of future studies.

In agreement with previous findings, we have shown that lncRNAs demonstrate tissue- and cell type-specific expression in comparison to protein-coding genes. More importantly, the differential expression of lncRNAs carry useful information associated with the tissue and cell identity, and is able to classify breast cancer cells into their intrinsic molecular subtypes, suggesting its potential utility as prognostic biomarkers in real breast cancer patients. Future studies are required to select the optimal and minimal subset of lncRNA genes for this purpose, and to compare its efficacy with currently available methods.

Moreover, the exploration of the utility of lncRNAs does not stop at the prognostics. Ultimately, we wish to elucidate the molecular and functional roles of lncRNAs in human breast cancer cells so as to generate new targets for therapeutic interventions in treating breast cancer, and this has been an area with great challenges. In this study, we focused one lncRNA gene (R152) that shows elevated expression in breast tumors in comparison to benign tissues, and another one (R67) that is associated with important cell viability pathway from the “guilt-by-association” analysis, and demonstrated that they are required for the normal proliferation of

MCF-7 breast caner cells. While the molecular details associated with these two lncRNAs need to be further elucidated, the cancer-specific lncRNA R152 shows great promises as a new drug target. Clearly, a few lncRNAs, including the previously characterized HOTAIR, GAS5, SRA and PVT1, as well as the newly characterized R152 and R67, have been shown to play key roles in important cellular processes in breast cancer, the jury is still out regarding the implications of lncRNAs as a group. Just like many previous lncRNA studies, we have generated many answers, as well as many more questions that need to be addressed in future studies.

3.5. Materials and Methods

Cell culture and treatments. MCF-7 cells were maintained in MEM with Hank's salts (Sigma; M1018) supplemented with 5% calf serum. For experiments involving estrogen treatment, the cells were grown for at least 3 days in phenol red-free MEM Eagle medium with Earle's salts (Sigma; M3024) supplemented with 5% charcoal-dextran-treated calf serum, and treated with ethanol (vehicle) or E2 (100 nM) for the times specified in the figure legends.

Cell fractionation. Estrogen-starved MCF-7 cells were treated with ethanol or 100 nM E2 for 45 min. Two biological replicates of 10^7 cells were processed for each experimental condition. They were trypsinized and collected, and subsequently lysed in buffer A (0.01 M HEPES pH 7.6, 0.01 M KCl, 15 mM $MgCl_2$, 0.34 M sucrose, 10 % glycerol, 1mM DTT, 0.3 mg/ml digitonin) in the presence of proteinase and RNase inhibitors. The soluble extract represented the cytoplasmic fraction. The remainder was then washed twice with buffer A, each time accompanied by 10 strokes of douncing, to obtain clean and intact nuclei. The nuclei were then ruptured with the sequential addition of low salt buffer (0.02 M Tris-HCl pH 7.5, 0.02 M KCl, 15 mM $MgCl_2$, 0.2

mM EDTA, 25 % glycerol) and high salt buffer (low salt buffer with 1.2 M KCl) in the presence of proteinase and RNase inhibitors to extract the nucleoplasmic contents into the soluble fraction. The insoluble pellet was then resuspended in cell disruption buffer (PARIS kit, Ambion), digested with DNase I (Roche), and taken as the chromatin-associated fraction. Total RNAs were extracted from each of the fractions using the PARIS columns (Ambion). They were further processed for whole genome polyadenylated RNA sequencing (poly(A)+ RNA-seq) as described below. In addition, the fractionated extracts were subjected to immunoblotting using antibodies against the cytoplasmic marker β-tubulin (Abcam, ab6046), the nucleoplasmic marker SNRP70 (Abcam, ab83306), and the chromatin-associated marker histone H3K4me3 (Active Motif, pAb), for confirmation of the purity of the subcellular fractions.

Poly(A)+ RNA-seq. Total RNAs from subcellular fractions were isolated as described above. Total RNAs from unfractionated MCF-7 cells were isolated using the RNeasy kit (QIAgen). Poly(A)+ RNAs were purified from these samples using Dynabeads® Oligo(dT)^{25} (Invitrogen) as described previously (Zhong, Joung et al. 2011). Strand-specific libraries were prepared according to the "deoxyuridine triphosphate (dUTP)" method described in the same protocol. An Illumina HiSeq 2000 was used for sequencing with a single-endsequencing length of 50 nt for the unfractionated poly(A)+ RNA-seq sample and a paired-endsequencing length of 100 nt for the treated and fractionated poly(A)+ RNA-seq samples.

LncRNA annotation pipeline.

a. RNA-seq read mapping. All paired-end RNA-seq reads generated, one replicate from untreated MCF-7 cells and two biological replicates each from the fractionated and vehicle- or

E2-treated cells, were aligned to the human genome (NCBI 37, Hg19) using the spliced read aligner TopHat version V2.0.4 (Kim, Pertea et al. 2013). For this analysis, we used two iterations of Tophat alignments as previously suggested (Cabili, Trapnell et al. 2011), to maximize the use of splice site information derived across all samples. Reads from all samples were first aligned with the purpose of splice-junction discovery, by not supplying any annotation files and including the "min-anchor-length" and "microexon-search" parameters. We then pooled the predicted splice sites across all alignments, and used the pooled junction file to facilitate the re-alignment of each of the fractionated RNA samples with the "raw-juncs" and "no-novel-juncs" parameters.

b. Transcriptome assembly. The biological replicates were combined and the transcriptome for each subcellular fraction and each treatment condition was then assembled by Cufflinks version V2.0.2 (Trapnell, Williams et al. 2010). After obtaining six unique sets of assembled isoforms, a minimal read coverage threshold and size selection filters were applied.

Minimal read coverage threshold. We ran Cufflinks with its transcript abundance calculation mode to estimate the read coverage of each transcript, and we removed transcripts with a maximal coverage below 10 reads per base.

Size selection. lncRNAs are mostly defined to be longer than 200 bp, therefore, we excluded multi-exonic transcripts smaller than 200 bp. We also considered the limitation of cufflinks in resolving the start and stop site of each transcript, and applied a more stringent size threshold of 1 kb to single exon transcripts.

The filtered transcripts were then merged into two sets using Cuffmerge, the cytoplasmic set and the nucleus-localized set containing both the nucleoplasmic and chromatin-associated fractions.

In addition, we removed any transcripts from the nucleus-localized set that overlaps with transcripts from the cytoplasmic set, to obtain two distinct sets of transcripts.

c. Filter of known non-lncRNA annotations. In both sets of transcripts, we eliminated those that had an exon overlapping a transcript from either coding genes annotated in RefSeq or in GENCODE V12. Single-exonic transcripts that trail behind and are within 2000 bp of an annotated coding gene were also removed, as they may represent polymerase run-on fragments of the coding gene. In addition, we classified each transcript based on its relative location with respect to its nearby coding gene, into divergent, antisense or intergenic.

d. Filter of transcripts lacking primary transcript evidence. GRO-seq data sets in MCF-7 cells, following 0 to 40 min. of E2 treatment, were obtained from earlier work in the Kraus Lab. They were re-analyzed to provide evidence of primary transcripts for potential lncRNA genes. As previously described, GRO-seq data were aligned to hg19 using SOAP2, and uniquely mappable reads were converted into bigWig files for visualization in UCSC genome browser, and into R data files for subsequent analysis. We called transcripts de novo based on a two-state Hidden Markov model, using the GRO-seq data analysis package (Hah, Danko et al. 2011). We then compared the filtered transcripts assembled from RNA-seq experiments to transcripts being called from the GRO-seq data, and retained only those that present evidence at both RNA-seq and GRO-seq levels.

e. Positive coding potential threshold. We estimated for each transcript the coding potential, as measured by codon substitution frequency (CSF), or the degree of evolutionary pressure in

maintaining the signature of an open reading frame against random substitutions. We ran PhyloCSF using a multiple sequence alignment of 29 mammalian genomes, (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz46way/) to obtain the best scoring open reading frame across all three reading frames. We excluded from our lncRNA catalog all transcripts with a PhyloCSF score greater than 150. This PhyloCSF threshold is determined by optimizing the sensitivity and specificity in correctly classifying RefSeq annotated protein-coding and noncoding transcripts, and it corresponds to a false discovery rate of 9 % for coding genes and false positive rate of 12 % for noncoding genes.

f. Transcript abundance threshold. We obtained transcript abundance, in terms of FPKM, for all transcripts by running Cufflinks. A FPKM threshold of 1 was applied to all lncRNA transcripts for most of the bioinformatics analysis in this study to only characterize those transcripts that were reasonably expressed in MCF-7 cells under our experimental conditions.

Sequence conservation estimation. For each lncRNA or protein-coding transcript, sequence conservation levels of the exons, introns and promoters were measured by the phastCons scores, which were extracted from the vertebrate phastCons elements 46-way table (UCSC table browser, comparative genomics, conservation tracks). We set the regions 1 kb upstream the transcription start site (TSS) as promoters.

Chromatin signatures. ChIP-seq data sets for H3K4me3, H3K36me3, H3K4me1, H3K27ac, ERα, CBP, SRC2, FOXA1 and AP2γ in either untreated, vehicle- or E2- treated conditions were obtained from various sources as listed in Supplementary Table 1. They were aligned to hg19

using Bowtie (Langmead, Trapnell et al. 2009), and uniquely mappable reads were converted into R data files for subsequent analysis. We then took processed data from both GRO-seq and ChIP-seq to delineate chromatin signatures around specified regions. Metagenes were used to illustrate the distribution of GRO-seq and ChIP-seq reads around the specified regions, using the metagene function in the GRO-seq data analysis package (Hah, Danko et al. 2011). Boxplots representations were used to minimize the bias caused by outliers in the data, which can lead to inaccurate interpretation of metagene representations. The read distribution in a given region was calculated and plotted using the boxplot function in R.

Estimation of subcellular distribution. To estimate the contribution of each subcellular fraction to the total population of poly(A)+ RNA, we utilize the relationship $a \times Cyto + b \times Nuc + c \times Chr = Total$, where *Cyto*, *Nuc*, *Chr* and *Total* refer to the FPKM values of each transcript in cytoplasmic, nucleoplasmic, chromatin-associated and total poly(A)+ RNA-seq samples respectively, and *a*, *b* and *c* indicate their corresponding contributions. We sampled the values of *a*, *b* and *c* from 0.01 to 0.99, and calculated the corresponding values of estimated total FPKM, $a \times Cyto + b \times Nuc + c \times Chr$, and the observed total FPKM, *Total,* for each of the Cufflinks-assembled transcript. We then performed a Kolmogorov–Smirnov (KS) test and calculated the Pearson correlation coefficient between the estimated and the observed total FPKM. The set of *a*, *b* and *c* that yield significant KS p-values and highest correlation coefficients represent the estimated contribution of subcellular fractions.

Estimation of transcript stability. To obtain a simple and convenient measure of transcript stability, we calculated the ratio of RNA-seq FPKM over GRO-seq RPKM for each lncRNA and

mRNA transcript, as it reflects the relative abundance of the mature RNA transcript over its corresponding primary transcript.

Determination of regulation at transcriptional level and steady-state RNA level. Transcriptional regulation was determined from GRO-seq reads using the bioconductor package edgeR as previously described, and we applied a 5% false discovery rate (FDR) to the analysis. Regulation at steady state RNA level was determined from RNA-seq reads using Cuffdiff, also with a 5% FDR.

Breadth and specificity of lncRNA and mRNA expression. Unstranded poly(A)+ RNA-seq datasets from 135 tumour tissues, 27 benign tissues, 109 tumour cell lines and 22 benign cell lines of the breast, prostate, stomach, malenocytes, pancreas, bladder, kidney, salivary gland, lymphoid and myeloid tissue were obtained from the Michigan Center for Translational Pathology (Kalyana-Sundaram, Kumar-Sinha et al. 2012). RNA-seq data from 3 additional breast cancer cell lines and 8 benign breast tissue samples were obtained from the Mayo Clinic (Asmann, Hossain et al. 2011). These RNA-seq results were mapped using Tophat and assembled in a similar manner as the newly generated RNA-seq datasets in MCF-7 cells using Cufflinks, which also generates the expression of all lncRNAs and mRNAs in terms of FPKM. A FPKM cutoff of 1 was applied to determine if a given lncRNA or mRNA is expressed in any particular tissue sample or cell line. Hierarchical clustering was performed on the differential expression of lncRNAs across all tissue samples and all breast cancer cell lines to evaluate its ability in predicting tissue identity and the intrinsic molecular subtype of breast cancer cells respectively.

Guilt-by-Association Analysis. The expression level of each lncRNA across the panel of 304 tissues and cell lines was correlated with all protein-coding genes. Each lncRNA was then associated with the entire list of protein-coding genes, ranked by their correlation with the lncRNA. We ran gene set enrichment analysis (GSEA), using the curated gene sets of canonical pathways and oncogenic signatures, on the ranked list of protein-coding genes to indentify pathways and signatures that are significantly enriched for the particular lncRNA.

Knockdown of lncRNAs in MCF-7 cells. Transient RNAi-mediated knockdown of lncRNAs was performed using transient transfection of siRNAs targeting lncRNAs and a control siRNA purchased from Sigma. The siRNA oligos was transfected into MCF-7 cells using the Lipofectamine RNAiMAX reagent following the manufactures' protocol. Forty-eight hours post transfections, the cells were collected for the evaluation of the efficiency of knockdown using RT-qPCR.

Proliferation Assay. After indicated days post-transfection, MCF-7 cells were fixed with 10% formaldehyde, stained with 0.1% of Crystal Violet, and subsequently washed and detained with 10% acetic acid. The acetic acid was then collected and red at absorbance 595nm to record the relative cell growth.

Analysis of lncRNAs and mRNAs by RT-qPCR. RT-qPCR detection of lncRNAs and mRNAs were performed as described previously. Total RNA was isolated from the cells using Trizol (Invitrogen) followed by isopropanol precipitation. It was then subjected to RT using oligoDT

(dT_{22}) and MMLV Reverse transcriptase (Promega). The cDNA was then subjected to qPCR analysis using a Roche LightCycler 480 system with SYBR Green detection and gene-specific primers. Gene-specific primers used in this study were listed in Supplementary Table 3. Each experiment was performed a minimum of three times with independent biological samples to ensure reproducibility.

REFERENCES

Alizadeh, A. A., M. B. Eisen, et al. (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." Nature 403(6769): 503-511.

Amaral, P. P., M. B. Clark, et al. (2011). "lncRNAdb: a reference database for long noncoding RNAs." Nucleic Acids Res 39(Database issue): D146-151.

Asmann, Y. W., A. Hossain, et al. (2011). "A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines." Nucleic Acids Res 39(15): e100.

Badger, J. H. and G. J. Olsen (1999). "CRITICA: coding region identification tool invoking comparative analysis." Mol Biol Evol 16(4): 512-524.

Beltran, M., I. Puig, et al. (2008). "A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition." Genes Dev 22(6): 756-769.

Bernard, D., K. V. Prasanth, et al. (2010). "A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression." EMBO J 29(18): 3082-3093.

Borsani, G., R. Tonlorenzi, et al. (1991). "Characterization of a murine gene expressed from the inactive X chromosome." Nature 351(6324): 325-329.

Brockdorff, N., A. Ashworth, et al. (1991). "Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome." Nature 351(6324): 329-331.

Brockdorff, N., A. Ashworth, et al. (1992). "The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus." Cell 71(3): 515-526.

Brown, C. J., B. D. Hendrich, et al. (1992). "The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus." Cell 71(3): 527-542.

Bu, D., K. Yu, et al. (2012). "NONCODE v3.0: integrative annotation of long noncoding RNAs." Nucleic Acids Res 40(Database issue): D210-215.

Cabili, M. N., C. Trapnell, et al. (2011). "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses." Genes Dev 25(18): 1915-1927.

Callis, T. E., K. Pandya, et al. (2009). "MicroRNA-208a is a regulator of cardiac hypertrophy and conduction in mice." J Clin Invest 119(9): 2772-2786.

Carninci, P., T. Kasukawa, et al. (2005). "The transcriptional landscape of the mammalian genome." Science 309(5740): 1559-1563.

Carrieri, C., L. Cimatti, et al. (2012). "Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat." Nature 491(7424): 454-457.

Castrignano, T., A. Canali, et al. (2004). "CSTminer: a web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison." Nucleic Acids Res 32(Web Server issue): W624-627.

Cesana, M., D. Cacchiarelli, et al. (2011). "A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA." Cell 147(2): 358-369.

Chooniedass-Kothari, S., E. Emberley, et al. (2004). "The steroid receptor RNA activator is the first functional RNA encoding a protein." FEBS Lett 566(1-3): 43-47.

Chooniedass-Kothari, S., M. K. Hamedani, et al. (2006). "The steroid receptor RNA activator protein is expressed in breast tumor tissues." Int J Cancer 118(4): 1054-1059.

Chu, C., K. Qu, et al. (2011). "Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions." Mol Cell 44(4): 667-678.

Churchman, L. S. and J. S. Weissman (2011). "Nascent transcript sequencing visualizes transcription at nucleotide resolution." Nature 469(7330): 368-373.

Clamp, M., B. Fry, et al. (2007). "Distinguishing protein-coding and noncoding genes in the human genome." Proc Natl Acad Sci U S A 104(49): 19428-19433.

Clark, M. B., R. L. Johnston, et al. (2012). "Genome-wide analysis of long noncoding RNA stability." Genome Res 22(5): 885-898.

Cooper, C., J. Guo, et al. (2009). "Increasing the relative expression of endogenous non-coding Steroid Receptor RNA Activator (SRA) in human breast cancer cells using modified oligonucleotides." Nucleic Acids Res 37(13): 4518-4531.

Core, L. J., J. J. Waterfall, et al. (2008). "Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters." Science 322(5909): 1845-1848.

Danko, C. G., N. Hah, et al. (2013). "Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells." Mol Cell 50(2): 212-222.

Derrien, T., R. Johnson, et al. (2012). "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression." Genome Res 22(9): 1775-1789.

Dinger, M. E., K. C. Pang, et al. (2009). "NRED: a database of long noncoding RNA expression." Nucleic Acids Res 37(Database issue): D122-126.

Finn, R. D., J. Mistry, et al. (2006). "Pfam: clans, web tools and services." Nucleic Acids Res 34(Database issue): D247-251.

Frohman, M. A., M. K. Dush, et al. (1988). "Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer." Proc Natl Acad Sci U S A 85(23): 8998-9002.

Geng, Y. J., S. L. Xie, et al. (2011). "Large intervening non-coding RNA HOTAIR is associated with hepatocellular carcinoma progression." J Int Med Res 39(6): 2119-2128.

Gong, C. and L. E. Maquat (2011). "lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements." Nature 470(7333): 284-288.

Gupta, R. A., N. Shah, et al. (2010). "Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis." Nature 464(7291): 1071-1076.

Gutschner, T., M. Hammerle, et al. (2013). "The Noncoding RNA MALAT1 Is a Critical Regulator of the Metastasis Phenotype of Lung Cancer Cells." Cancer Res.

Guttman, M., I. Amit, et al. (2009). "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals." Nature 458(7235): 223-227.

Guttman, M., J. Donaghey, et al. (2011). "lincRNAs act in the circuitry controlling pluripotency and differentiation." Nature 477(7364): 295-300.

Guttman, M., M. Garber, et al. (2010). "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs." Nat Biotechnol 28(5): 503-510.

Hah, N., C. G. Danko, et al. (2011). "A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells." Cell 145(4): 622-634.

Hah, N., S. Murakami, et al. (2013). "Enhancer transcripts mark active estrogen receptor binding sites." Genome Res.

Harrow, J., F. Denoeud, et al. (2006). "GENCODE: producing a reference annotation for ENCODE." Genome Biol 7 Suppl 1: S4 1-9.

Hatchell, E. C., S. M. Colley, et al. (2006). "SLIRP, a small SRA binding protein, is a nuclear receptor corepressor." Mol Cell 22(5): 657-668.

Hawkins, P. G. and K. V. Morris (2010). "Transcriptional regulation of Oct4 by a long non-coding RNA antisense to Oct4-pseudogene 5." Transcription 1(3): 165-175.

Hsu, F., W. J. Kent, et al. (2006). "The UCSC Known Genes." Bioinformatics 22(9): 1036-1046.

Hube, F., G. Velasco, et al. (2011). "Steroid receptor RNA activator protein binds to and counteracts SRA RNA-mediated activation of MyoD and muscle differentiation." Nucleic Acids Res 39(2): 513-525.

Hulo, N., A. Bairoch, et al. (2006). "The PROSITE database." Nucleic Acids Res 34(Database issue): D227-230.

Hutchinson, J. N., A. W. Ensminger, et al. (2007). "A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains." BMC Genomics 8: 39.

Ingolia, N. T., S. Ghaemmaghami, et al. (2009). "Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling." Science 324(5924): 218-223.

Ingolia, N. T., L. F. Lareau, et al. (2011). "Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes." Cell 147(4): 789-802.

Jan, C. H., R. C. Friedman, et al. (2011). "Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs." Nature 469(7328): 97-101.

Janowski, B. A. and D. R. Corey (2010). "Minireview: Switching on progesterone receptor expression with duplex RNA." Mol Endocrinol 24(12): 2243-2252.

Janowski, B. A., S. T. Younger, et al. (2007). "Activating gene expression in mammalian cells with promoter-targeted duplex RNAs." Nat Chem Biol 3(3): 166-173.

Jeon, Y. and J. T. Lee (2011). "YY1 tethers Xist RNA to the inactive X nucleation center." Cell 146(1): 119-133.

Ji, P., S. Diederichs, et al. (2003). "MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer." Oncogene 22(39): 8031-8041.

Jia, H., M. Osak, et al. (2010). "Genome-wide computational identification and manual annotation of human long noncoding RNA genes." RNA 16(8): 1478-1487.

Jin, Z. B., G. Hirokawa, et al. (2009). "Targeted deletion of miR-182, an abundant retinal microRNA." Mol Vis 15: 523-533.

Kalyana-Sundaram, S., C. Kumar-Sinha, et al. (2012). "Expressed pseudogenes in the transcriptional landscape of human cancers." Cell 149(7): 1622-1634.

Kapranov, P., J. Cheng, et al. (2007). "RNA maps reveal new RNA classes and a possible function for pervasive transcription." Science 316(5830): 1484-1488.

Karreth, F. A., Y. Tay, et al. (2011). "In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma." Cell 147(2): 382-395.

Khalil, A. M., M. Guttman, et al. (2009). "Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression." Proc Natl Acad Sci U S A 106(28): 11667-11672.

Kim, D., G. Pertea, et al. (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." Genome Biol 14(4): R36.

Kino, T., D. E. Hurt, et al. (2010). "Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor." Sci Signal 3(107): ra8.

Kogo, R., T. Shimamura, et al. (2011). "Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers." Cancer Res 71(20): 6320-6326.

Kong, L., Y. Zhang, et al. (2007). "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine." Nucleic Acids Res 35(Web Server issue): W345-349.

Kretz, M., Z. Siprashvili, et al. (2013). "Control of somatic tissue differentiation by the long non-coding RNA TINCR." Nature 493(7431): 231-235.

Krogh, A., B. Larsson, et al. (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." J Mol Biol 305(3): 567-580.

Krystal, G. W., B. C. Armstrong, et al. (1990). "N-myc mRNA forms an RNA-RNA duplex with endogenous antisense transcripts." Mol Cell Biol 10(8): 4180-4191.

Lai, F., U. A. Orom, et al. (2013). "Activating RNAs associate with Mediator to enhance chromatin architecture and transcription." Nature 494(7438): 497-501.

Langmead, B., C. Trapnell, et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol 10(3): R25.

Lanz, R. B., N. J. McKenna, et al. (1999). "A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex." Cell 97(1): 17-27.

Lee, G. L., A. Dobi, et al. (2011). "Prostate cancer: diagnostic performance of the PCA3 urine test." Nat Rev Urol 8(3): 123-124.

Lee, J. T., L. S. Davidow, et al. (1999). "Tsix, a gene antisense to Xist at the X-inactivation centre." Nat Genet 21(4): 400-404.

Lee, J. T. and N. Lu (1999). "Targeted mutagenesis of Tsix leads to nonrandom X inactivation." Cell 99(1): 47-57.

Leygue, E. (2007). "Steroid receptor RNA activator (SRA1): unusual bifaceted gene products with suspected relevance to breast cancer." Nucl Recept Signal 5: e006.

Li, J. T., Y. Zhang, et al. (2008). "Trans-natural antisense transcripts including noncoding RNAs in 10 species: implications for expression regulation." Nucleic Acids Res 36(15): 4833-4844.

Li, Y. M., G. Franklin, et al. (1998). "The H19 transcript is associated with polysomes and may regulate IGF2 expression in trans." J Biol Chem 273(43): 28247-28252.

Lin, M. F., I. Jungreis, et al. (2011). "PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions." Bioinformatics 27(13): i275-282.

Liu, J., J. Gough, et al. (2006). "Distinguishing protein-coding from non-coding RNAs through support vector machines." PLoS Genet 2(4): e29.

Liu, N., S. Bezprozvannaya, et al. (2008). "microRNA-133a regulates cardiomyocyte proliferation and suppresses smooth muscle gene expression in the heart." Genes Dev 22(23): 3242-3254.

Marahrens, Y., B. Panning, et al. (1997). "Xist-deficient mice are defective in dosage compensation but not spermatogenesis." Genes Dev 11(2): 156-166.

Mituyama, T., K. Yamada, et al. (2009). "The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs." Nucleic Acids Res 37(Database issue): D89-92.

Mondal, T., M. Rasmussen, et al. (2010). "Characterization of the RNA content of chromatin." Genome Res 20(7): 899-907.

Mourtada-Maarabouni, M., M. R. Pickard, et al. (2009). "GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer." Oncogene 28(2): 195-208.

Munroe, S. H. and M. A. Lazar (1991). "Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA." J Biol Chem 266(33): 22083-22086.

Mus, E., P. R. Hof, et al. (2007). "Dendritic BC200 RNA in aging and in Alzheimer's disease." Proc Natl Acad Sci U S A 104(25): 10679-10684.

Nam, J. W. and D. P. Bartel (2012). "Long noncoding RNAs in C. elegans." Genome Res 22(12): 2529-2540.

Nie, Y., X. Liu, et al. (2013). "Long non-coding RNA HOTAIR is an independent prognostic marker for nasopharyngeal carcinoma progression and survival." Cancer Sci.

Nielsen, H., J. Engelbrecht, et al. (1997). "A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites." Int J Neural Syst 8(5-6): 581-599.

Nielsen, T. O., R. B. West, et al. (2002). "Molecular characterisation of soft tissue tumours: a gene expression study." Lancet 359(9314): 1301-1307.

Ogawa, Y., B. K. Sun, et al. (2008). "Intersection of the RNA interference and X-inactivation pathways." Science 320(5881): 1336-1341.

Orom, U. A., T. Derrien, et al. (2010). "Long noncoding RNAs with enhancer-like function in human cells." Cell 143(1): 46-58.

Park, C. Y., L. T. Jeker, et al. (2012). "A resource for the conditional ablation of microRNAs in the mouse." Cell Rep 1(4): 385-391.

Parker, J. S., M. Mullins, et al. (2009). "Supervised risk predictor of breast cancer based on intrinsic subtypes." J Clin Oncol 27(8): 1160-1167.

Pauli, A., E. Valen, et al. (2012). "Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis." Genome Res 22(3): 577-591.

Penny, G. D., G. F. Kay, et al. (1996). "Requirement for Xist in X chromosome inactivation." Nature 379(6561): 131-137.

Perou, C. M., T. Sorlie, et al. (2000). "Molecular portraits of human breast tumours." Nature 406(6797): 747-752.

Pinter, S. F., R. I. Sadreyev, et al. (2012). "Spreading of X chromosome inactivation via a hierarchy of defined Polycomb stations." Genome Res 22(10): 1864-1876.

Preker, P., J. Nielsen, et al. (2008). "RNA exosome depletion reveals transcription upstream of active human promoters." Science 322(5909): 1851-1854.

Rinn, J. L., M. Kertesz, et al. (2007). "Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs." Cell 129(7): 1311-1323.

Rivas, E. and S. R. Eddy (2001). "Noncoding RNA gene detection using comparative sequence analysis." BMC Bioinformatics 2: 8.

Sado, T., Y. Hoki, et al. (2005). "Tsix silences Xist through modification of chromatin structure." Dev Cell 9(1): 159-165.

Salmena, L., L. Poliseno, et al. (2011). "A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?" Cell 146(3): 353-358.

Schwartz, J. C., S. T. Younger, et al. (2008). "Antisense transcripts are targets for activating small RNAs." Nat Struct Mol Biol 15(8): 842-848.

Shi, Y., M. Downes, et al. (2001). "Sharp, an inducible cofactor that integrates nuclear receptor repression and activation." Genes Dev 15(9): 1140-1151.

Shiraki, T., S. Kondo, et al. (2003). "Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage." Proc Natl Acad Sci U S A 100(26): 15776-15781.

Shore, A. N., E. B. Kabotyanski, et al. (2012). "Pregnancy-induced noncoding RNA (PINC) associates with polycomb repressive complex 2 and regulates mammary epithelial differentiation." PLoS Genet 8(7): e1002840.

Sigova, A. A., A. C. Mullen, et al. (2013). "Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells." Proc Natl Acad Sci U S A 110(8): 2876-2881.

Sun, L., L. A. Goff, et al. (2013). "Long noncoding RNAs regulate adipogenesis." Proc Natl Acad Sci U S A 110(9): 3387-3392.

Szymanski, M., V. A. Erdmann, et al. (2003). "Noncoding regulatory RNAs database." Nucleic Acids Res 31(1): 429-431.

Szymanski, M., V. A. Erdmann, et al. (2007). "Noncoding RNAs database (ncRNAdb)." Nucleic Acids Res 35(Database issue): D162-164.

Tani, H., R. Mizutani, et al. (2012). "Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals." Genome Res 22(5): 947-956.

Tay, Y., L. Kats, et al. (2011). "Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs." Cell 147(2): 344-357.

Tian, D., S. Sun, et al. (2010). "The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation." Cell 143(3): 390-403.

Trapnell, C., B. A. Williams, et al. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nat Biotechnol 28(5): 511-515.

Tripathi, V., J. D. Ellis, et al. (2010). "The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation." Mol Cell 39(6): 925-938.

Tsai, M. C., O. Manor, et al. (2010). "Long noncoding RNA as modular scaffold of histone modification complexes." Science 329(5992): 689-693.

Tsuchihara, K., Y. Suzuki, et al. (2009). "Massive transcriptional start site analysis of human genes in hypoxia cells." Nucleic Acids Res 37(7): 2249-2263.

Ulitsky, I., A. Shkumatava, et al. (2011). "Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution." Cell 147(7): 1537-1550.

van Rooij, E., L. B. Sutherland, et al. (2007). "Control of stress-dependent cardiac growth and gene expression by a microRNA." Science 316(5824): 575-579.

Wakaguri, H., R. Yamashita, et al. (2008). "DBTSS: database of transcription start sites, progress report 2008." Nucleic Acids Res 36(Database issue): D97-101.

Wang, K. C., Y. W. Yang, et al. (2011). "A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression." Nature 472(7341): 120-124.

Watanabe, M., J. Yanagisawa, et al. (2001). "A subfamily of RNA-binding DEAD-box proteins acts as an estrogen receptor alpha coactivator through the N-terminal activation domain (AF-1) with an RNA coactivator, SRA." EMBO J 20(6): 1341-1352.

Williams, A. H., G. Valdez, et al. (2009). "MicroRNA-206 delays ALS progression and promotes regeneration of neuromuscular synapses in mice." Science 326(5959): 1549-1554.

Williams, G. T., M. Mourtada-Maarabouni, et al. (2011). "A critical role for non-coding RNA GAS5 in growth arrest and rapamycin inhibition in human T-lymphocytes." Biochem Soc Trans 39(2): 482-486.

Willingham, A. T., A. P. Orth, et al. (2005). "A strategy for probing the function of noncoding RNAs finds a repressor of NFAT." Science 309(5740): 1570-1573.

Wilson, D., M. Madera, et al. (2007). "The SUPERFAMILY database in 2007: families and functions." Nucleic Acids Res 35(Database issue): D308-313.

Wilson, D., R. Pethica, et al. (2009). "SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny." Nucleic Acids Res 37(Database issue): D380-386.

Wutz, A. and R. Jaenisch (2000). "A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation." Mol Cell 5(4): 695-705.

Wutz, A., T. P. Rasmussen, et al. (2002). "Chromosomal silencing and localization are mediated by different domains of Xist RNA." Nat Genet 30(2): 167-174.

Yan, M. D., C. C. Hong, et al. (2005). "Identification and characterization of a novel gene Saf transcribed from the opposite strand of Fas." Hum Mol Genet 14(11): 1465-1474.

Yang, L., C. Lin, et al. (2011). "ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs." Cell 147(4): 773-788.

Yoon, J. H., K. Abdelmohsen, et al. (2012). "LincRNA-p21 suppresses target mRNA translation." Mol Cell 47(4): 648-655.

Zhang, B., G. Arun, et al. (2012). "The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult." Cell Rep 2(1): 111-123.

Zhang, Y., X. S. Liu, et al. (2006). "Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species." Nucleic Acids Res 34(12): 3465-3475.

Zhao, J., T. K. Ohsumi, et al. (2010). "Genome-wide identification of polycomb-associated RNAs by RIP-seq." Mol Cell 40(6): 939-953.

Zhao, J., B. K. Sun, et al. (2008). "Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome." Science 322(5902): 750-756.

Zhao, X., J. R. Patton, et al. (2007). "Pus3p- and Pus1p-dependent pseudouridylation of steroid receptor RNA activator controls a functional switch that regulates nuclear receptor signaling." Mol Endocrinol 21(3): 686-699.

Zhong, S., J. G. Joung, et al. (2011). "High-throughput illumina strand-specific RNA sequencing library preparation." Cold Spring Harb Protoc 2011(8): 940-949.