Supplemental Information

"Discovery, Annotation, and Functional Analysis of Long Noncoding RNAs Controlling Cell Cycle Gene Expression and Proliferation in Breast Cancer Cells"

Sun *et al.* (2015)

This document contains the following supplemental information: Page

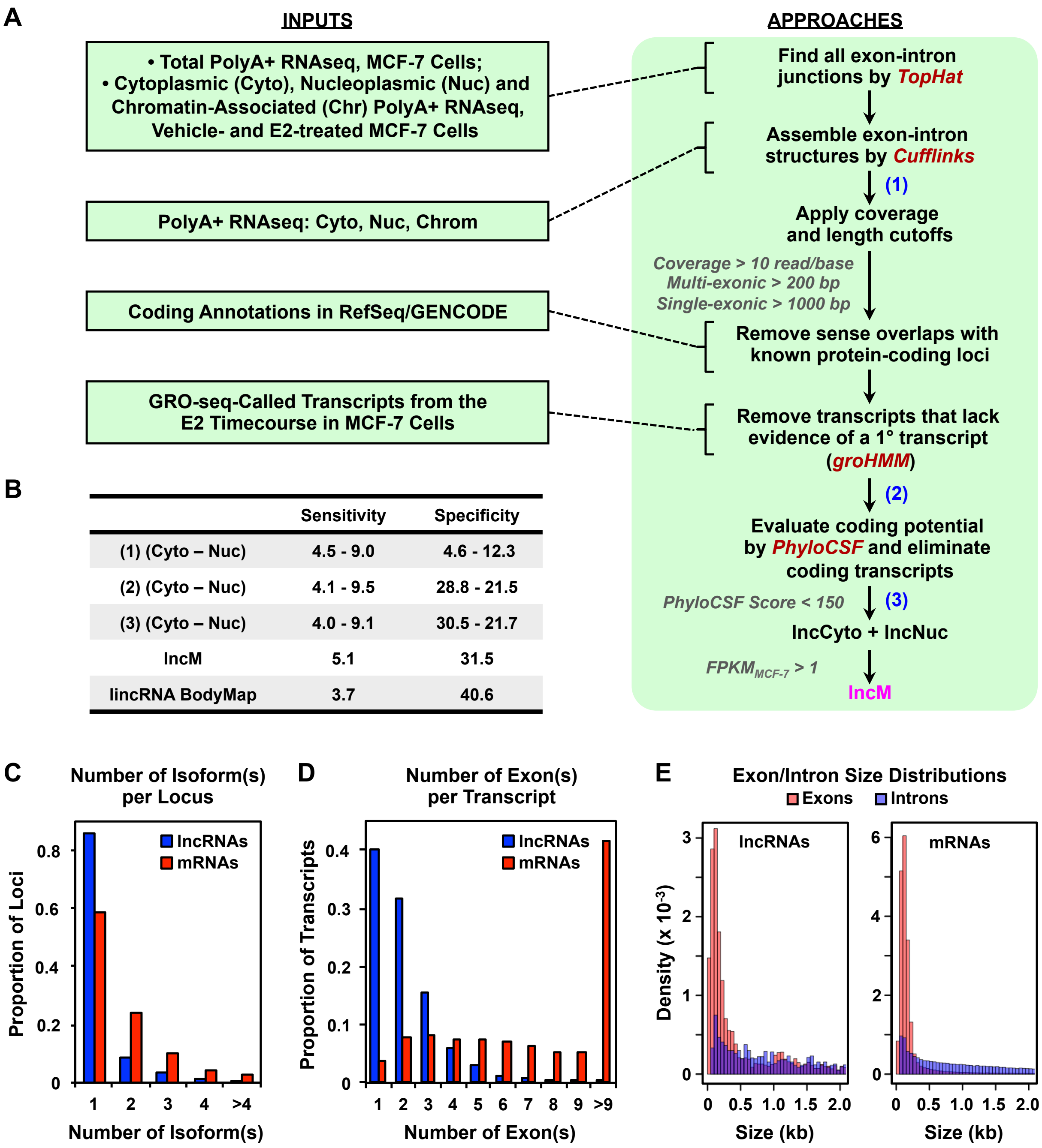
1) Supplemental Figures

[Figure S1 is on the next page]

Figure S1. Generation of a lncRNA catalog from MCF-7 cells.
(A) (*Left*) The pipeline takes as inputs: (1) polyA+ RNA-seq data (total and fractionated RNA) from MCF-7 cells under basal and E2-treated conditions, (2) existing coding annotations from RefSeq and GENCODE for humans, and (3) GRO-seq data from MCF-7 cells under basal and E2-treated conditions. (*Right*) RNA-seq data were aligned to the genome and assembled into transcripts by TopHat ver. 2.0.4 (Kim et al., 2013) and Cufflinks ver. 2.0.2 (Trapnell et al., 2010), respectively. RNA-seq data from the cytoplasmic RNA fraction (for lncCyto) was evaluated separately from the RNA-seq data from the nuclear RNA fraction (for lncNuc). RNA transcripts called from the RNA-seq data were filtered by coverage and size. Transcription units were called de novo from GRO-seq data using groHMM (Hah et al., 2011; Luo et al., 2014) and transcripts with no evidence of a primary transcript were removed. RNA transcripts with both RNA-seq and GRO-seq information were further filtered by coding capacity using physloCSF (Lin et al., 2011). The remaining transcripts were pass through an expression level filter (FPKM >1) to derive at a set of 1888 expressed lncRNA genes (lncM).
(B) Sensitivity and specificity at the base level measured by Cufflinks.cuffcompare, comparing transcripts annotations from lncCyto and lncNuc at the indicated points along the pipeline (1, 2, and 3; see panel A, right), lncM, and lincRNA Bodymap, with lncRNA annotations from GENCODE.
(C through E) Comparisons of the number of isoform(s) per gene locus (C), number of exon(s) per transcript (D), and exon and intron size distributions (E) for lncRNA and protein-coding genes assembled from cytoplasmic RNA-seq data.
[Related to Figure 1]

Figure S1

[see previous page for legend]



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

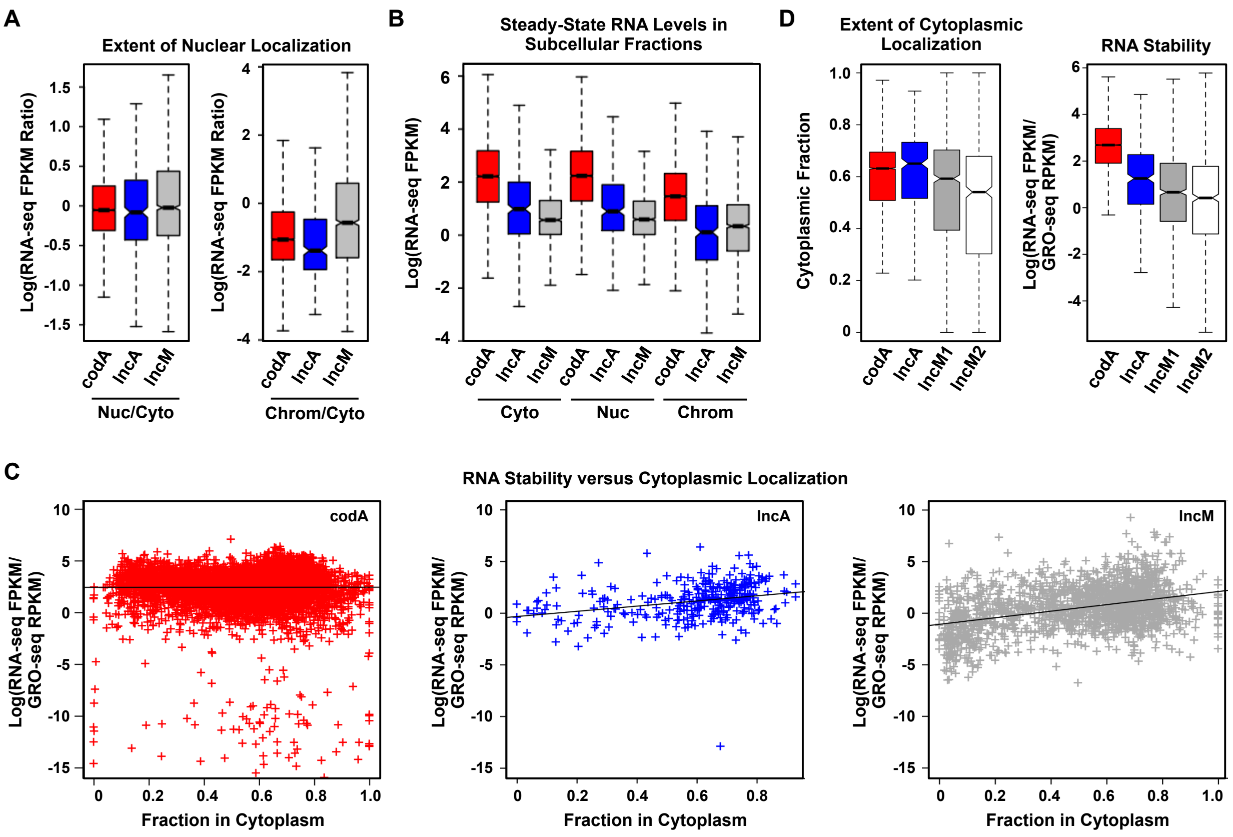


Figure S2. Subcellular localization and stability of lncRNAs and mRNAs.

(A) Box plots showing the extent of nuclear localization for codA, lncA, and lncM.

(B) Box plots showing the steady-state RNA levels for codA, lncA, and lncM in each of the subcellular fractions.

(C) Stability versus extent of cytoplasmic localization plots for codA, lncA, and lncM.

(D) Box blots showing the cytoplasmic fraction of codA, lncA, lncM1, and lncM2, as determined by fractionated RNA-seq (*right*), and the stability of codA, lncA, lncM1, and lncM2, as determined by taking the $\log_{10}$(RNA-seq FPKM/GRO-seq RPKM) (*right*). For the cytoplasmic fraction analysis, the lncM2 set is statistically different from all others (Wilcoxon rank sum test, $p < 1.7 \times 10^{-5}$). For the stability analysis, the lncM1 and lncM2 sets are statistically different from the others, but not from each other (Tukey's multiple comparison test, $p < 2 \times 10^{-16}$). The lncA lncRNAs (347 lncRNAs) are a subset of the lncM lncRNAs that match perfectly with the intron chain of an existing annotation in either RefSeq or GENCODE. The lncM1 lncRNAs (1088 lncRNAs) are a subset of the lncM lncRNAs that overlap to any extent with previous annotations in RefSeq, GENCODE, or the lincRNA Body Map. The lncM2 lncRNAs (1088 lncRNAs) are a subset of the lncM lncRNAs that do not match in any way with previous annotations in RefSeq, GENCODE, or the lincRNA Body Map. lncM = lncM1 + lncM2.

In all panels, Log = the natural log ($\log_e$).

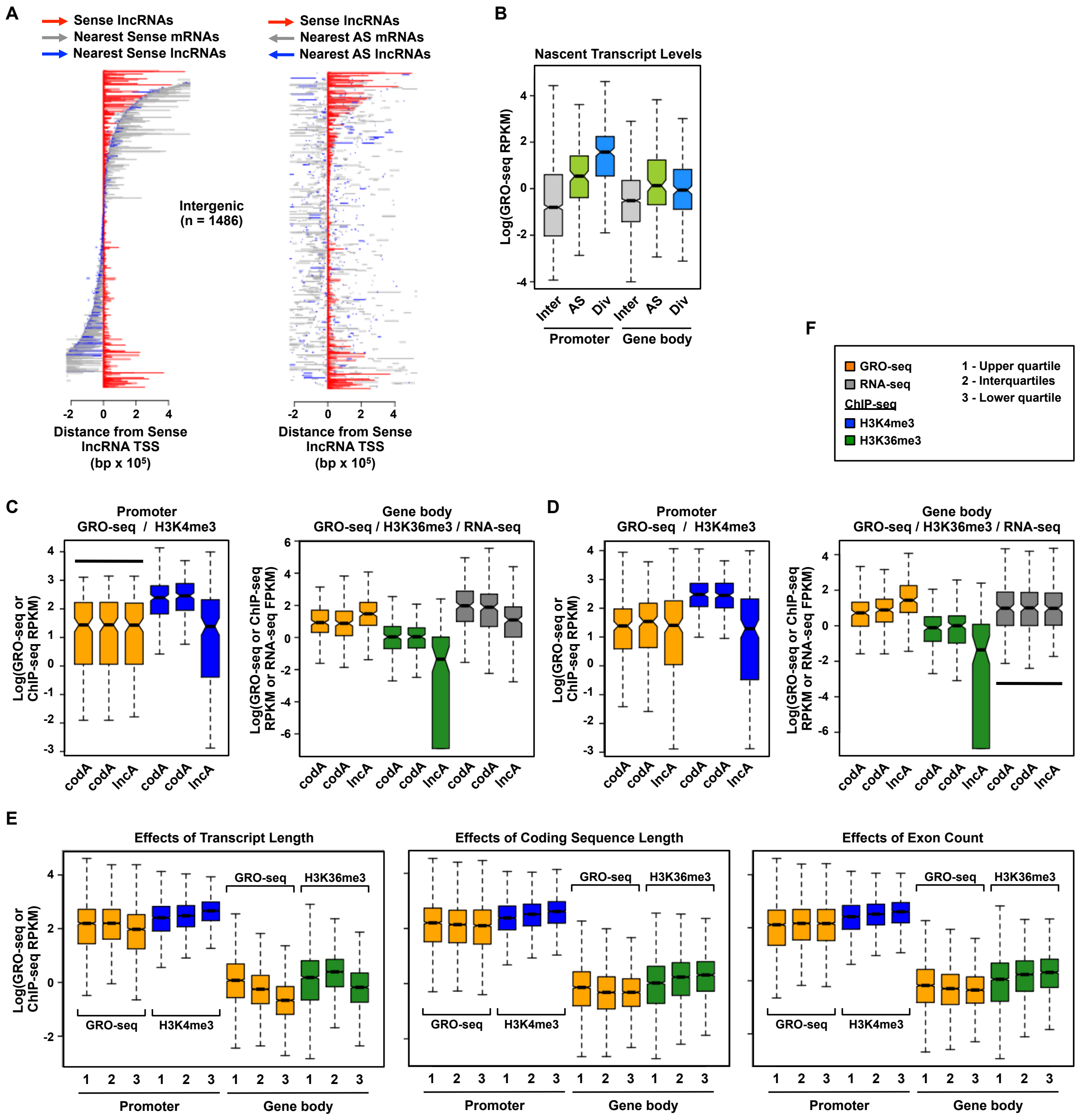
[Related to Figure 2]

[Figure S3 is on the next page]

Figure S3. Intergenic lncRNA genes are transcribed to lower levels than divergent and antisense lncRNA genes, and annotated lncRNAs have lower levels of promoter H3K4me3 and gene body H3K36me3 than equally expressed protein-coding genes.
(A) Graphical representation of the orientation, position, and length of intergenic (Inter) lncRNA genes relative to their nearest sense (*left*) or antisense (*right*) RNA genes.
(B) Box plot showing levels of transcription based on GRO-seq for intergenic, antisense, and divergent lncM genes.
(C and D) Box plots showing analyses similar to those in Fig. 3. The plots compare the levels of (1) active RNA Pol II (GRO-seq) and H3K4me3 (ChIP-seq) at the promoter (*left*), and (2) actively transcribing RNA Pol II (GRO-seq) in the gene body, H3K36me3 (ChIP-seq) in the gene body, and steady-state RNA levels (RNA-seq) (*right*) for selected codA genes and intergenic lncA genes. (C) Sampling of two non-overlapping sets of codA genes that have the same level of GRO-seq signal at the promoter as the intergenic lncA genes (box plots on the left highlighted by the solid bar above). (D) Sampling of two non-overlapping sets of codA genes that have the same level of steady-state RNA as the intergenic lncA genes (box plots on the right highlighted by the solid bar below).
(E) Box plots comparing the levels of (1) active RNA Pol II (GRO-seq) and H3K4me3 (ChIP-seq) at the promoter (*left side*), and (2) actively transcribing RNA Pol II (GRO-seq) and H3K36me3 (ChIP-seq) in the gene body (*right side*) for protein-coding genes separated into upper quartile (labeled “1”), interquartiles (labeled “2”), and lower quartile (labeled “3”) based on the length of the mRNA transcripts (left panel), length of the coding sequences (middle panel), and the number of exons of each mRNA transcript (right panel).
(F) Color and number key for (C) through (E).
In all panels (B), (C), (D), and (E), Log = the natural log ($\log_e$).
[Related to Figure 3]

Figure S3

[see previous page for legend]



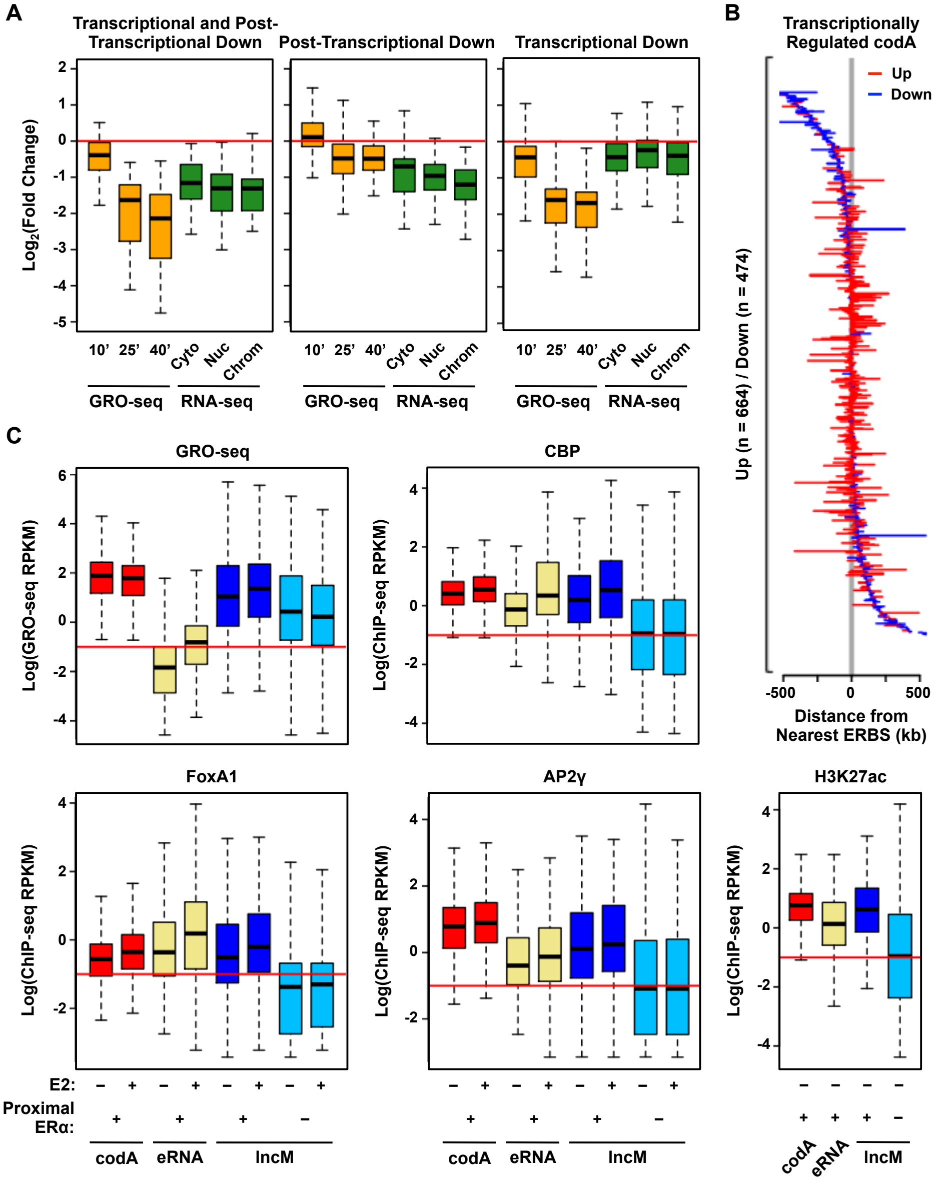


Figure S4. A subset of lncM genes is associated with ERα and elevated levels of enhancer features.

(A) Box plots showing the E2-induced fold changes in the expression (GRO-seq and RNA-seq) of lncM genes that are downregulated (1) both transcriptionally and post-transcriptionally, (2) post-transcriptionally only, and (3) transcriptionally only**.**

(B) Graphical representation of the gene length and distance from the nearest ERα-binding site (ERBS) for E2-responsive, transcriptionally regulated codA genes.

(C) Box plots comparing the levels of ERα, an enhancer-associated coregulator (i.e., CBP), pioneer factors (i.e., FoxA1 and AP2γ), and an enhancer-associated histone mark (i.e., H3K27ac) near the TSSs of (1) codA gene promoters, (2) enhancers that produce eRNAs, and (3) lncM gene promoters, with or without nearby (proximal) ERα binding as indicated. Log = the natural log ($\log_e$).

[Related to Figure 4]

[Figure S5 is on the next page]

Figure S5. Differential expression of lncRNAs informs tissue identity and predicts biological functions and outcomes.

(A) 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 (*left*) or 1888 codA genes that show a higher standard deviation among the tissue samples (*right*).

(B) List of REACTOME pathways from the "guilt-by-association" analysis demarcated by the purple bar in Fig. 5C.

(C) Properties of *lncRNA152* and *lncRNA67*.

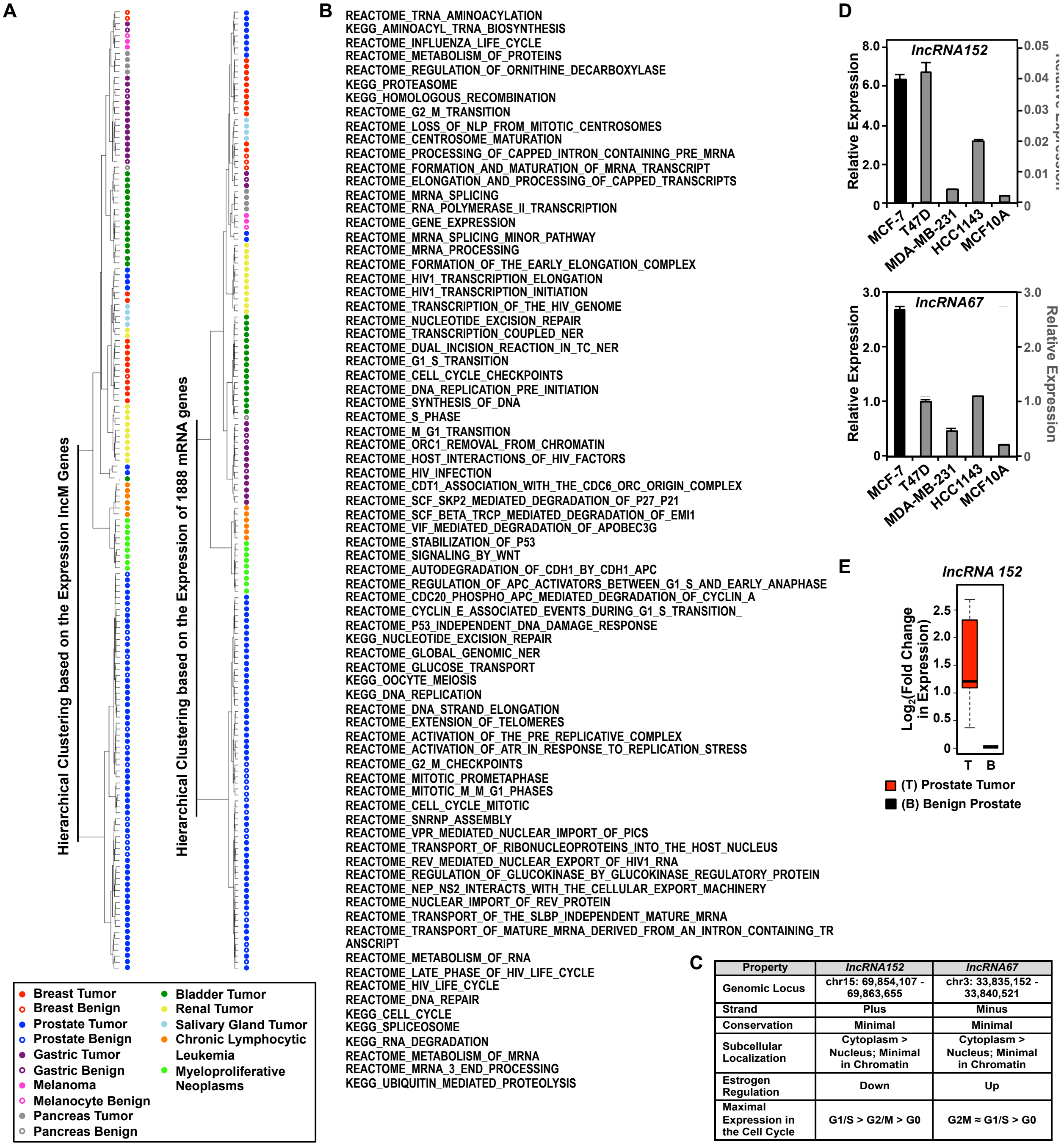
(D) Relative expression of *lncRNA152* (*top*) or *lncRNA67* (*bottom*) in MCF-7 (ERα+), T47D (ERα+), MDA-MB-231 (triple negative), and HCC1143 (triple negative) breast cancer cells, as well as MCF10A "normal" breast epithelial cells. β-actin mRNA was used as an internal control. Each bar represents the mean + SEM, n = 3.

(E) Box plot showing elevated expression of *lncRNA152* in prostate tumors (T, red) compared to benign breast tissues (B, grey).

[Related to Figure 5]

Figure S5

[see previous page for legend]



REACTOME_TRNA_AMINOACYLATION
KEGG_AMINOACYL_TRNA_BIOSYNTHESIS
REACTOME_INFLUENZA_LIFE_CYCLE
REACTOME_METABOLISM_OF_PROTEINS
REACTOME_REGULATION_OF_ORNITHINE_DECARBOXYLASE
KEGG_PROTEASOME
KEGG_HOMOLOGOUS_RECOMBINATION
REACTOME_G2_M_TRANSITION
REACTOME_LOSS_OF_NLP_FROM_MITOTIC_CENTROSOMES
REACTOME_CENTROSOME_MATURATION
REACTOME_PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_MRNA
REACTOME_FORMATION_AND_MATURATION_OF_MRNA_TRANSCRIPT
REACTOME_ELONGATION_AND_PROCESSING_OF_CAPPED_TRANSCRIPTS
REACTOME_MRNA_SPLICING
REACTOME_RNA_POLYMERASE_II_TRANSCRIPTION
REACTOME_GENE_EXPRESSION
REACTOME_MRNA_SPLICING_MINOR_PATHWAY
REACTOME_MRNA_PROCESSING
REACTOME_FORMATION_OF_THE_EARLY_ELONGATION_COMPLEX
REACTOME_HIV1_TRANSCRIPTION_ELONGATION
REACTOME_HIV1_TRANSCRIPTION_INITIATION
REACTOME_TRANSCRIPTION_OF_THE_HIV_GENOME
REACTOME_NUCLEOTIDE_EXCISION_REPAIR
REACTOME_TRANSCRIPTION_COUPLED_NER
REACTOME_DUAL_INCISION_REACTION_IN_TC_NER
REACTOME_G1_S_TRANSITION
REACTOME_CELL_CYCLE_CHECKPOINTS
REACTOME_DNA_REPLICATION_PRE_INITIATION
REACTOME_SYNTHESIS_OF_DNA
REACTOME_S_PHASE
REACTOME_M_G1_TRANSITION
REACTOME_ORC1_REMOVAL_FROM_CHROMATIN
REACTOME_HOST_INTERACTIONS_OF_HIV_FACTORS
REACTOME_HIV_INFECTION
REACTOME_CDT1_ASSOCIATION_WITH_THE_CDC6_ORC_ORIGIN_COMPLEX
REACTOME_SCF_SKP2_MEDIATED_DEGRADATION_OF_P27_P21
REACTOME_SCF_BETA_TRCP_MEDIATED_DEGRADATION_OF_EMI1
REACTOME_VIF_MEDIATED_DEGRADATION_OF_APOBEC3G
REACTOME_STABILIZATION_OF_P53
REACTOME_SIGNALING_BY_WNT
REACTOME_AUTODEGRADATION_OF_CDH1_BY_CDH1_APC
REACTOME_REGULATION_OF_APC_ACTIVATORS_BETWEEN_G1_S_AND_EARLY_ANAPHASE
REACTOME_CDC20_PHOSPHO_APC_MEDIATED_DEGRADATION_OF_CYCLIN_A
REACTOME_CYCLIN_E_ASSOCIATED_EVENTS_DURING_G1_S_TRANSITION_
REACTOME_P53_INDEPENDENT_DNA_DAMAGE_RESPONSE
KEGG_NUCLEOTIDE_EXCISION_REPAIR
REACTOME_GLOBAL_GENOMIC_NER
REACTOME_GLUCOSE_TRANSPORT
KEGG_OOCYTE_MEIOSIS
KEGG_DNA_REPLICATION
REACTOME_DNA_STRAND_ELONGATION
REACTOME_EXTENSION_OF_TELOMERES
REACTOME_ACTIVATION_OF_THE_PRE_REPLICATIVE_COMPLEX
REACTOME_ACTIVATION_OF_ATR_IN_RESPONSE_TO_REPLICATION_STRESS
REACTOME_G2_M_CHECKPOINTS
REACTOME_MITOTIC_PROMETAPHASE
REACTOME_MITOTIC_M_M_G1_PHASES
REACTOME_CELL_CYCLE_MITOTIC
REACTOME_SNRNP_ASSEMBLY
REACTOME_VPR_MEDIATED_NUCLEAR_IMPORT_OF_PICS
REACTOME_TRANSPORT_OF_RIBONUCLEOPROTEINS_INTO_THE_HOST_NUCLEUS
REACTOME_REV_MEDIATED_NUCLEAR_EXPORT_OF_HIV1_RNA
REACTOME_REGULATION_OF_GLUCOKINASE_BY_GLUCOKINASE_REGULATORY_PROTEIN
REACTOME_NEP_NS2_INTERACTS_WITH_THE_CELLULAR_EXPORT_MACHINERY
REACTOME_NUCLEAR_IMPORT_OF_REV_PROTEIN
REACTOME_TRANSPORT_OF_THE_SLBP_INDEPENDENT_MATURE_MRNA
REACTOME_TRANSPORT_OF_MATURE_MRNA_DERIVED_FROM_AN_INTRON_CONTAINING_TRANSCRIPT
REACTOME_METABOLISM_OF_RNA
REACTOME_LATE_PHASE_OF_HIV_LIFE_CYCLE
REACTOME_HIV_LIFE_CYCLE
REACTOME_DNA_REPAIR
KEGG_CELL_CYCLE
KEGG_SPLICEOSOME
KEGG_RNA_DEGRADATION
REACTOME_METABOLISM_OF_MRNA
REACTOME_MRNA_3_END_PROCESSING
KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS

Property	*lncRNA152*	*lncRNA67*
Genomic Locus	chr15: 69,854,107 - 69,863,655	chr3: 33,835,152 - 33,840,521
Strand	Plus	Minus
Conservation	Minimal	Minimal
Subcellular Localization	Cytoplasm > Nucleus; Minimal in Chromatin	Cytoplasm > Nucleus; Minimal in Chromatin
Estrogen Regulation	Down	Up
Maximal Expression in the Cell Cycle	G1/S > G2/M > G0	G2M ≈ G1/S > G0

[Figure S6 is on the next page]

Figure S6. ***LncRNA152* and *lncRNA67* are required for the growth of breast cancer cells and the expression of cell cycle-related genes.**

(A) Analysis of the growth of MCF-7 cells after control (si-Ctrl, si-Luc, si-GFP) or siRNA-mediated knockdown of *lncRNA152* and *lncRNA67* (si-lncRNA152 and si-lncRNA67, respectively) over a six day time course post-transfection. Each point represents the mean ± SEM, n = 3.

(B) siRNA-mediated knockdown of *lncRNA152* (*left*) or *lncRNA67* (*right*) in T47D or MDA-MB-231 cells using two independent siRNA oligos. The levels of *lncRNA152* and *lncRNA67* after knockdown were monitored by RT-qPCR. β-actin mRNA was used as an internal control. Each bar represents the mean + SEM, n = 3.

(C) Analysis of the growth of T47D (ERα+) or MDA-MB-231 (triple negative) breast cancer cells after control or siRNA-mediated knockdown of *lncRNA152* (*top*) or *lncRNA67* (*bottom*) over a 6 day time course post-transfection. Each point represents the mean ± SEM, n = 3.

(D) Ectopic expression of *lncRNA152* or *lncRNA67* enhances the growth of MCF-7 cells (*left*) and rescues the growth of MCF-7 cells following siRNA-mediated knockdown of the lncRNAs (*right*). Stably transfected cells were treated with Dox to induce the expression of either GFP or the lncRNAs after the cells were transfected with siRNAs (si-Ctrl, si-lncRNA152 and si-lncRNA67). The cells were grown for three days after treatment with Dox. Each point represents the mean ± SEM, n = 3. Asterisks, p-value < 0.05 (**, relative to GFP control; *, relative to no ectopic expression).

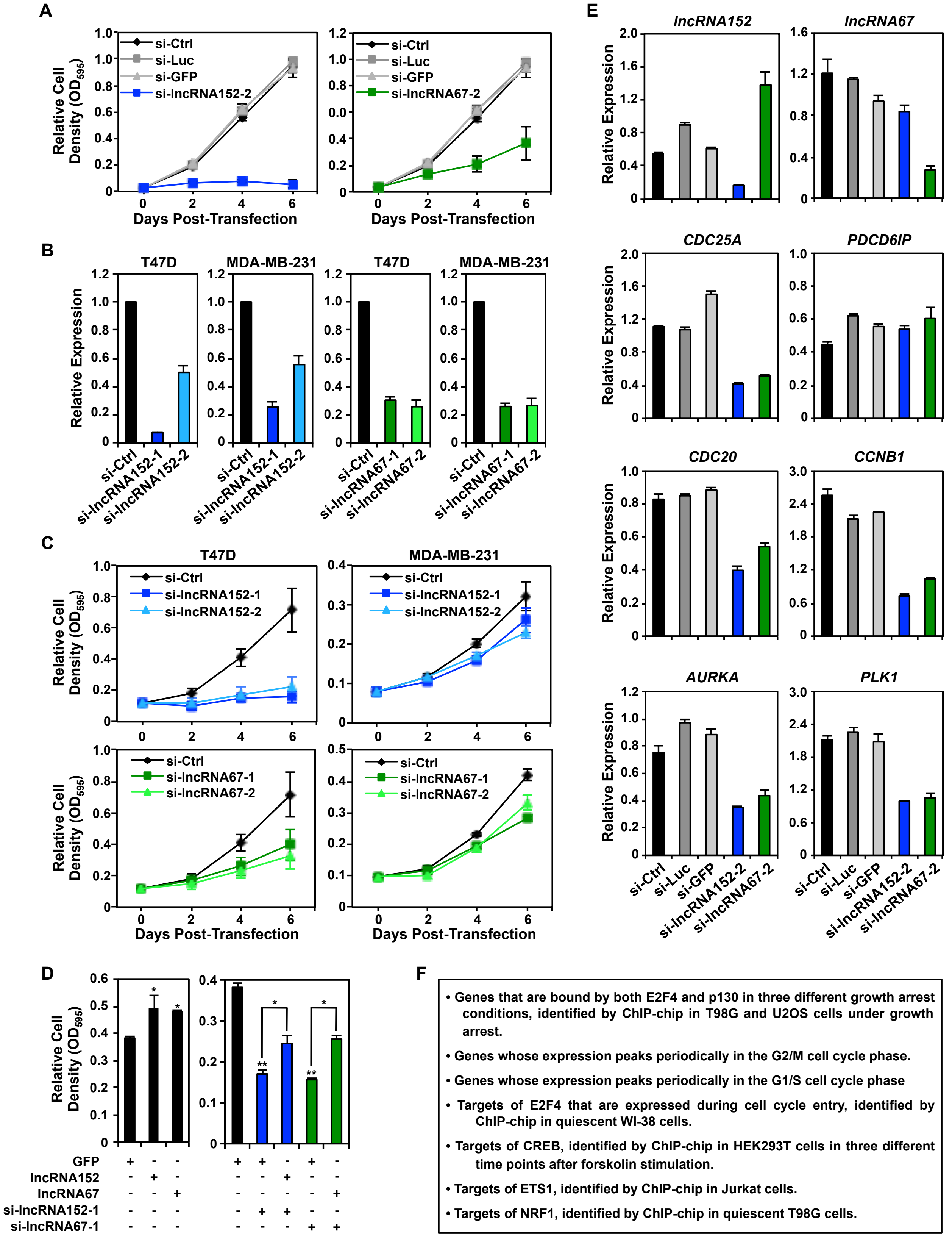
(E) Analysis of the expression of cell cycle-related genes in MCF-7 cells after control (si-Ctrl, si-Luc, si-GFP) or siRNA-mediated knockdown of *lncRNA152* and *lncRNA67* (si-lncRNA152 and si-lncRNA67, respectively; 20 nM). The expression of each gene after knockdown was monitored by RT-qPCR. β-actin mRNA was used as an internal control. Each bar represents the mean + SEM, n = 3.

(F) The full set of gene descriptors from the GREAT analysis shown in Fig. 6D.

[Related to Figure 6]

Figure S6

[see previous page for legend]



- Genes that are bound by both E2F4 and p130 in three different growth arrest conditions, identified by ChIP-chip in T98G and U2OS cells under growth arrest.
- Genes whose expression peaks periodically in the G2/M cell cycle phase.
- Genes whose expression peaks periodically in the G1/S cell cycle phase
- Targets of E2F4 that are expressed during cell cycle entry, identified by ChIP-chip in quiescent WI-38 cells.
- Targets of CREB, identified by ChIP-chip in HEK293T cells in three different time points after forskolin stimulation.
- Targets of ETS1, identified by ChIP-chip in Jurkat cells.
- Targets of NRF1, identified by ChIP-chip in quiescent T98G cells.

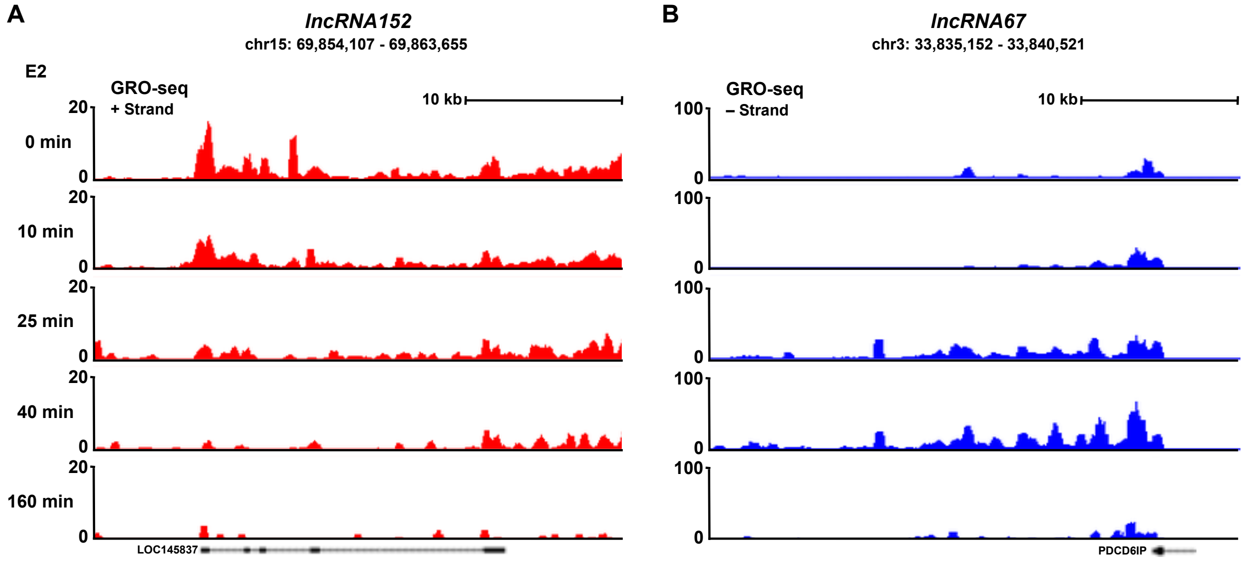


Figure S7. The transcription of the genes encoding *lncRNA152 and lncRNA67* is regulated by estrogen.
Genome browser tracks of GRO-seq data showing the regulation of transcription of the genes encoding (A) *lncRNA152* (LOC145837) and (B) *lncRNA67* over a time course of E2 treatment. *[Related to Figure 7]*

2) Supplemental Experimental Procedures

Cell culture and treatments

MCF-7 cells, kindly provided by Benita Katzenellenbogen (University of Illinois, Urbana-Champaign), were maintained in MEM medium with Hank's salts (Sigma; M1018) supplemented with 5% calf serum (Hyclone). 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 then treated with ethanol (vehicle) or 17β-estradiol (E2; 100 nM) for the times specified in the figures and legends. T47D and MDA-MB-231 cells, kindly provided by Khandan Keyomarsi, (MD Anderson Cancer Center, Houston), were grown in RPMI 1640 supplemented with 10% fetal bovine serum (Atlanta Biologicals).

Cell fractionation and RNA isolation

Estrogen-withdrawn MCF-7 cells were treated with ethanol or 100 nM E2 for 3 hours. Two biological replicates of 10^7 cells were processed for each experimental condition (except for the unfractionated MCF-7 cells, which had one replicate; see below). Adherent cells were trypsinized, 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 protease and RNase inhibitors. The nuclei were collected by gentle centrifugation at 4°C, and the resulting supernatant was collected as the cytoplasmic fraction. The nuclear pellet was washed twice with buffer A, each time accompanied by 10 strokes of douncing, to obtain clean and intact nuclei. The nuclei were then extracted 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 protease and RNase inhibitors (Gadad et al., 2009). The soluble extract was collected as the nuclear fraction. The remaining pellet was then resuspended in cell disruption buffer (PARIS kit, Ambion; AM1921), digested with DNase I (Roche), and centrifuged to remove any remaining insoluble material. The resulting supernatant was collected as the chromatin fraction.

After removal of a small aliquot from each fraction for Western blotting, total RNA was extracted using the PARIS kit (Ambion) according to the manufacturer's instructions. In addition, total RNA was also isolated from unfractionated MCF-7 cells using the RNeasy kit (QIAgen) according to the manufacturer's instructions. The RNA collected from each subcellular fraction, as well as the unfractionated MCF-7 cells, was processed for whole genome polyadenylated RNA sequencing (polyA+ RNA-seq) as described below. In addition, samples of the fractionated extracts were subjected to Western blotting 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; 39159), for confirmation of the subcellular fractionation procedure.

PolyA+ RNA-seq

The RNA collected from each subcellular fraction, as well as the unfractionated MCF-7 cells, was subjected to enrichment of polyA+ RNA using Dynabeads Oligo(dT)25 (Invitrogen) as described previously (Zhong et al., 2011). Strand-specific RNA-seq libraries were prepared from the polyA+ RNA according to the "deoxyuridine triphosphate (dUTP)" method as described previously (Zhong et al., 2011). The RNA-seq libraries were sequenced using an Illumina HiSeq 2000 as follows: (1) the polyA+ RNA-seq libraries from E2- or control-treated fractionated RNA were sequenced using paired-end methodology with a length of 100 nt (PE100) (two replicates each) and (2) the polyA+ RNA-seq library from untreated unfractionated RNA was sequenced using single-end methodology with a length of 50 nt (SE50) (one replicate).

Computational pipeline for annotation of lncRNAs

We used the following computational pipeline to annotate lncRNAs using RNA-seq and GRO-seq data.

RNA-seq read mapping. The RNA-seq reads were aligned to the human genome (NCBI 37, hg19) using the spliced read aligner TopHat version 2.0.4 (Kim et al., 2013). For this analysis, we used two iterations of TopHat alignments as previously suggested to maximize the use of splice site information derived across all samples (Cabili et al., 2011). We first aligned the reads from all samples with the purpose of splice-junction discovery, which we achieved by not supplying 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.

Transcriptome assembly. The biological replicates were combined, and the transcriptome for each subcellular fraction and each treatment condition was then assembled by Cufflinks ver. 2.0.2 (Trapnell 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 using its transcript abundance calculation mode to estimate the read coverage of each transcript. We removed transcripts with a maximal coverage below 10 reads per base.

Size selection. Since lncRNAs are defined as RNAs >200 bp, we excluded multi-exonic transcripts smaller than 200 bp. We also considered the limitations 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: a cytoplasmic set and a nuclear set, with the latter containing lncRNAs from both the nucleoplasmic and chromatin-associated fractions. In addition, we removed any transcripts from the nuclear set that overlap with transcripts from the cytoplasmic set, to obtain two distinct sets of transcripts.

Filtering transcripts versus known annotations and classifying based on gene location and orientation. We eliminated any transcripts from the cytoplasmic and nuclear sets that have an exon overlapping protein-coding transcripts annotated in RefSeq or in GENCODE ver. 12. Single exon RNAs that are transcribed from genes located within 2 kb of the 3' end of an annotated protein-coding gene were also removed, as they may represent polymerase run-on products from the coding gene. In addition, we classified each transcript based on the relative location of its gene in the genome in relation to the nearest neighboring protein-coding gene, into divergent, antisense, or intergenic.

Filtering transcripts lacking evidence of a primary transcript using GRO-seq. Published GRO-seq data sets from control and E2-treated MCF-7 cells (Hah et al., 2011) were used for this analysis. The data sets were re-analyzed to provide evidence of primary transcripts for potential lncRNA genes. The GRO-seq data were aligned to hg19 using SOAP2, and the uniquely mappable reads were converted into (1) bigWig files for visualization in a genome browser and (2) R data files for subsequent analyses. We called transcripts de novo based on a two-state Hidden Markov model using the GRO-seq data analysis package groHMM (Hah et al., 2011; Luo et al., 2014). We then compared the filtered transcripts assembled from the RNA-seq analyses (see above) to the primary transcripts called from the GRO-seq data. We retained only those transcripts with evidence of a steady-state transcript and a primary transcript, based on RNA-seq and GRO-seq, respectively.

Filtering transcripts based on a coding potential threshold. For each transcript, we estimated the coding potential based on 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 (Lin et al., 2011) 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, which was determined by optimizing the sensitivity and specificity in correctly classifying RefSeq annotated protein-coding and noncoding transcripts, corresponds to a false discovery rate of 9 % for coding genes and false positive rate of 12 % for noncoding genes.

Filtering transcripts based on a transcript abundance threshold. We obtained transcript abundance, in terms of FPKM, for all transcripts from Cufflinks. A FPKM threshold of 1 was applied to all lncRNAs for most of the bioinformatics analysis described herein to limit our focus to those transcripts that are reasonably expressed in MCF-7 cells under the experimental conditions used.

Additional analyses of lncRNAs

After annotating the lncRNAs, we performed a variety of additional analyses to characterize the lncRNAs as a class of RNAs.

Estimation of sequence conservation. For each lncRNA or mRNA, the sequence conservation levels of the exons, introns, and promoters were determined using phastCons scores (Siepel et al., 2005), which were extracted from the vertebrate phastCons 46-way alignment (UCSC Genome Browser). We set the region from -1000 bp to -1 bp relative to the TSS as the promoters.

Estimation of subcellular distribution. To estimate the contribution of each subcellular fraction to the total population of polyA+ RNAs, we used the relationship *a* x Cyto + *b* x Nuc + *c* x Chrom = Total, where (1) Cyto, Nuc, Chrom are the FPKM values of each transcript in the cytoplasmic (Cyto), nucleoplasmic (Nuc), and chromatin-associated (Chrom) polyA+ RNA-seq samples, respectively, (2) *a*, *b*, and *c* indicate their corresponding contributions, and (3) Total is the estimated total FPKM. We sampled values of *a*, *b*, and *c* ranging from 0.01 to 0.99, and then calculated the corresponding values of estimated total FPKM, *a* x Cyto + *b* x Nuc + *c* x Chr, and the observed total FPKM for each 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 values for *a*, *b*, and *c* that yield significant KS p-values and the highest correlation coefficients represent the estimated contribution to each of the 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. This value reflects the relative abundance of the mature RNA transcript over its corresponding primary transcript.

Determination of regulation at the transcriptional level versus the steady-state RNA level. Transcriptional regulation was determined from GRO-seq reads using the Bioconductor package edgeR as previously described (Hah et al., 2011; Robinson et al., 2010), using a 5% false discovery rate (FDR) for the analysis. Regulation at the steady state RNA level was determined from RNA-seq reads using Cuffdiff (Trapnell et al., 2013), using a 5% FDR.

Determination of the breadth and specificity of lncRNA and mRNA expression. Non-strand-specific polyA+ RNA-seq datasets from 135 tumor tissues, 27 benign tissues, 109 tumor cell lines, and 22 benign cell lines of the breast, prostate, stomach, melanocytes, pancreas, bladder, kidney, salivary gland, lymphoid and myeloid tissue were obtained from the Michigan Center for Translational Pathology (Kalyana-Sundaram et al., 2012). RNA-seq data from three additional breast cancer cell lines and eight benign breast tissue samples were obtained from the Mayo Clinic (Asmann et al., 2011). These RNA-seq reads from these samples were mapped using Tophat and assembled using Cufflinks in a manner similar to that described above for our newly generated RNA-seq datasets from MCF-7 cells. These analyses also generate 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 was expressed in each tissue sample or cell line. Hierarchical clustering was performed on the differential expression of lncRNAs across all samples and breast cancer cell lines to evaluate its ability to predict tissue identity and the intrinsic molecular subtype of breast cancer cells, respectively.

Analysis of histone modification, coregulator, and transcription factor signatures at lncRNA genes from published ChIP-seq and GRO-seq data

Published ChIP-seq data sets for H3K4me3, H3K36me3, H3K4me1, H3K27ac, ERα, CBP, FOXA1, and AP2γ from untreated or E2-treated MCF-7 cells were obtained from public repositories, as listed below. The data were aligned to hg19 using Bowtie (Langmead et al., 2009)), and uniquely mappable reads were converted into R data files for subsequent analyses. Processed ChIP-seq and GRO-seq data were then used to explore the enrichment of histone modifications, coregulators, and transcription factors around specified regions. Metagene plots 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 groHMM (Hah et al., 2011). Boxplot 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.

Guilt-by-association analyses

To explore the possible functions of the lncRNAs from MCF-7 cells, we used guilt-by-association analyses, as described previously (Guttman et al., 2009; Hung et al., 2011; Ravasi et al., 2006). For these analyses, the expression of each lncRNA in the lncM set (based on RNA-seq) across a panel of 304 tissues and cell lines (see above) was correlated with the expression of each mRNA. Each lncRNA was then associated with the entire list of mRNAs, ranked by their correlation with the lncRNA. We then performed gene set enrichment analysis (GSEA) (Subramanian et al., 2005), using curated gene sets of canonical pathways and oncogenic signatures, on the ranked list of mRNAs and back-associated the GSEA results to the lncRNAs to identify pathways and signatures that were significantly enriched for the particular lncRNA.

Functional analyses of lncRNAs in MCF-7 cells

To assess the functions of lncRNAs of selected lncRNAs, we performed a set of gene-specific and global assays with and without RNAi-mediated knockdown of the selected lncRNAs.

Knockdown of lncRNAs in MCF-7 cells. Transient RNAi-mediated knockdown of lncRNAs in MCF-7 cells was performed by transfection of (1) siRNAs targeting selected lncRNAs (designed using SciTools RNAi design software from Integrated DNA Technologies) and (2) a commercially available control siRNA (Sigma, MISSION siRNA universal negative control; SIC001). MCF-7 cells were plated at a density of 2 x 10^5 cells per well in six well dishes. The siRNA oligos (5 nM) were transfected into MCF-7 cells using Lipofectamine RNAiMAX reagent (2.5 µl per well) following the manufacturer's protocol. Forty-eight hours post transfection, the cells were collected to evaluate of the efficiency of lncRNA knockdown using RT-qPCR, and for the analysis of global changes in gene expression using RNA-seq.

Analysis of lncRNA and mRNA expression by RT-qPCR. RT-qPCR detection of lncRNAs and mRNAs was performed as described previously (Hah et al., 2013; Sun et al., 2012). Total RNA was isolated from siRNA-treated MCF-7 cells using the EZ-10 DNAaway RNA Mini-Prep Kit (Biobasic) and then subjected to reverse transcription using oligoDT (dT22) and MMLV reverse transcriptase (Promega). The resulting cDNA was then subjected to qPCR analysis using a Roche LightCycler 480 system with SYBR Green detection and gene-specific primers (see the list of primer sequences used). Each experiment was performed a minimum of three times with independent biological samples to ensure reproducibility.

Analysis of lncRNA-regulated genes by RNA-seq. PolyA+ RNA-seq libraries were prepared from control and lncRNA knockdown MCF-7 cells as described above using the dUTP method (Zhong et al., 2011). Two biological replicated were generated for each sample. The RNA-seq results were mapped to the hg19 human reference genome by TopHat (Kim et al., 2013), using RefSeq gene annotations as the reference for alignment. Differentially regulated RefSeq mRNAs were called by Cuffdiff, using a 5% FDR, comparing the control samples to the lncRNA knockdown samples targeting each lncRNA of interest. We derived a high-confidence regulated mRNA set by filtering the Cuffdiff-called regulated mRNA lists with a fold cutoff of either 2^(0.8) or 2^(-0.8) for each siRNA-treated condition relative to the control. The resulting mRNAs with their corresponding fold changes were represented in heatmaps using Java Treeview. We also performed a transcription factor target analysis using Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al., 2010) on the regulated mRNA set to draw inferences about transcription factors that may contribute to the observed gene regulation.

Analysis of RNA stability. The stability of lncRNAs and mRNAs was determined experimentally by treating MCF-7 cells by with 2.5 µg/mL actinomycin D (Sigma) for four hours and then monitoring RNA levels by RT-qPCR

Cell proliferation assays

MCF-7 cells were plated at a density of 1 x 10^5 cells per well in six well dishes. The siRNA oligos (2.5 nM) were transfected into MCF-7 cells using the Lipofectamine RNAiMAX reagent (2.5 µl per well) following the manufacturer's protocol. After transfection, the cells were grown for the number of days indicated in the figures. The cells were then fixed with 10% formaldehyde, stained with 0.1% crystal violet in 200 mM phosphoric acid, and washed and destained with 10% acetic acid. The acetic acid destain was collected and read at absorbance 595 nm. The results were expressed as relative cell growth.

Inducible ectopic expression of lncRNAs in MCF-7 cells. We generated a doxycycline (Dox)-inducible lentiviral vector for the inducible expression of *lncRNA152* and *lncRNA67*. We mapped the 5' and 3' ends of *lncRNA152* and *lncRNA67* and cloned the cDNAs. The cDNAs, or a GFP cDNA control, were inserted into pInducer20 (Meerbrey et al., 2011) and stably introduced into MCF-7 cells under neomycin/G418 selection. The stably transfected cells were plated at a density of 5 x 10^4 cells per well in six well dishes. Three days after induction with Dox (50 ng/mL), the cells numbers were quantified using crystal violet staining as described above.

Cell cycle analyses

Expression of lncRNAs throughout the cell cycle was determined for G0 (serum withdrawn), G1/S (double thymidine block/hydroxyurea), and G2/M (nocodazole) synchronized MCF-7 cells. Total RNA was isolated and subjected to RT-qPCR using gene-specific primers as described above. FACS analysis was performed on siRNA-transfected cells. Seventy-two hours post transfection, the cells were harvested, fixed in 85% ethanol, and stained with propidium iodide. Each experiment was performed a minimum of three times with independent biological samples to ensure reproducibility.

Data sets

All of the RNA-seq, GRO-seq, and ChIP-seq data sets from MCF-7 cells used in this study can be accessed through the NCBI's Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/) or the DDBJ Sequence Read Archive (DRA) (http://trace.ddbj.nig.ac.jp/dra/index_e.html) using the accession numbers listed below. In addition, a list of all of the lncRNAs with their start, end, and exon positions can be accessed from GEO using the series accession number GSE63189.

The new RNA-seq data sets generated specifically for this study, which can be accessed from GEO using the series accession number GSE63189, are as follows:

New MCF-7 cell data sets	Accession number(s)
• RNA-seq: polyA+ RNA from total RNA [(1)]	GSM1543640
• RNA-seq: polyA+ RNA from the cytoplasmic fraction ± E2 [(2)]	GSM1543653, GSM1543654, GSM1543655, GSM1543656
• RNA-seq: polyA+ RNA from the nuclear fraction ± E2 [(2)]	GSM1543657, GSM1543658, GSM1543659, GSM1543660
• RNA-seq: polyA+ RNA from the chromatin fraction ± E2 [(2)]	GSM1543661, GSM1543662, GSM1543663, GSM1543664
• RNA-seq: polyA+ RNA - siRNA control for *lncRNA152* [(1)]	GSM1543640, GSM1543642
• RNA-seq: polyA+ RNA - *lncRNA152* knockdown [(1)]	GSM1543643, GSM1543644, GSM1543645, GSM1543646
• RNA-seq: polyA+ RNA - siRNA control for *lncRNA672* [(1)]	GSM1543647, GSM1543648
• RNA-seq: polyA+ RNA - *lncRNA67* knockdown [(1)]	GSM1543649, GSM1543650, GSM1543651, GSM1543652

[(1)] single-end, 50 nt
[(2)] paired-end, 100 nt

RNA-seq data set information: Reads, mappability, and correlation between replicates:

Experimental Conditions		**Replicates**	**No. of Aligned Reads**	**Sequencing Depth**	**Percent aligned**
RNA-seq for Annotation (polyA+ and total RNA)					
Cyto	Vehicle	U1	64979693	99694130	65
		U2	50101035	69995584	72
	E2	E1	37139176	53278738	70
		E2	47261933	66486194	71
Nuc	Vehicle	U1	22696324	34211200	66
		U2	36219975	53525538	68
	E2	E1	29412077	43994948	67
		E2	42167861	62733056	67
Chrom	Vehicle	U1	68423201	89373162	77
		U2	36503483	47688604	77
	E2	E1	37999191	50017200	76
		E2	40745228	53027136	77
Total RNA		-	138942229	162393686	86
RNA-seq with lncRNA knockdown					
Control		1	9025253	11110357	81
		2	9475717	11275933	84
lncRNA152 knockdown #1		1	9634085	10948131	88
		2	9321999	10952335	85
lncRNA152 knockdown #2		1	8809903	10548269	84
		2	10871700	12631122	86
Control		1	9696258	10901186	89
		2	10033370	11557133	87
lncRNA67 knockdown #1		1	15098342	17358689	87
		2	9807485	11032132	89
*lncRNA67*knockdown #2		1	12825357	14358733	89
		2	9995430	11712380	85

The existing/published GRO-seq and ChIP-seq data sets used for this study are as follows:

Published MCF-7 cell data sets	Accession number(s)
• GRO-seq: E2 treatment time course	GSE27463, GSE41324
• ChIP-seq: H3K4me3 ± E2	GSM588571, GSM588570
• ChIP-seq: H3K4me1 ± E2	GSM588569, GSM588568
• ChIP-seq: H3K27ac	GSM946850
• ChIP-seq: H3K36me3	GSM916106
• ChIP-seq: ERα ± E2	GSM365925, GSM365926

• ChIP-seq: CBP ± E2 (3)	ERR045723, ERR045724
• ChIP-seq: FoxA1 ± E2	GSM588929, GSM588930
• ChIP-seq: AP2γ ± E2	GSM588928, GSM588927

(3) The CBP ChIP-seq data set is the only one from the DRA. All others arte from GEO.

Oligonucleotides used for siRNA-mediated knockdown and RT-qPCR

We used the following nucleic acid oligonucleotides for siRNA-mediated knockdown and RT-qPCR.

siRNAs for knockdown

• si-lncRNA152-1	5'-GCAGAAGUAUGAACAUAAU[dT][dT]-3'
• si-lncRNA152-2	5'-GGCAGAGAAACCUGGGUUU[dT][dT]-3'
• si-lncRNA67-1	5'-GGAAGAUUAAGGUGAUACU[dT][dT]-3'
• si-lncRNA67-2	5'-GACGAAAUCAGGAAAGCUA[dT][dT]-3'
• si-Luc	5'-GAUUUGUAUUCAGCCCAUA[dT][dT]-3'
• si-GFP	5'-ACAACAGCCACAACGUCUA[dT][dT]-3'

Primers for RT-qPCR

• LncRNA152 Fwd:	5'-AGAAATGCCACCGGACATAG-3'
• LncRNA152 Rev:	5'-CATACTTCTGCTGCGTCCAA-3'
• LncRNA67 Fwd:	5'-GTGGAGCCATGTGAAAGGTT-3'
• LncRNA67 Rev:	5'-CTCCAACCAGTGTCTGAGCA-3'
• LncRNA250 Fwd:	5'-TCAGTAGACACCTCCCGTCT-3'
• LncRNA250 Rev:	5'-CGACGGGCAACCAATGAAAC-3'
• LncRNA3 Fwd:	5'-TAACCTCCTCGGACTCCTGC-3'
• LncRNA3 Rev:	5'-TATAATCACTGCGCCCGCTC-3'
• LncRNA231 Fwd:	5'-AAATGCCAGTTCTGCGGGTA-3'
• LncRNA231 Rev:	5'-TGGAAATCCCAGGCCTACTTG-3'
• RPL19 Fwd:	5'-ACATCCACAAGCTGAAGGCA-3'
• RPL19 Rev:	5'-TGCGTGCTTCCTTGGTCTTA-3'
• FXYD3 Fwd:	5'-TCATCTGCGCTGGGGTTCT-3'
• FXYD3 Rev:	5'-CCTGGATGGTGACCGGACT-3'
• DBI Fwd:	5'-AATACCGTGGATGGTGGGAA-3'
• DBI Rev:	5'-CGTATGGTGAGCAGCCTTGA-3'
• COX8A Fwd:	5'-TGGGATCATGGAATTGGCCG-3'
• COX8A Rev:	5'-CTGTAGGTCTCCAGGTGTGAC-3'
• CDC25A Fwd:	5'-TTCCTACACGCGCATTGAGA-3'
• CDC25A Rev:	5'-AATCTGAAGGCCATCCCACC-3'
• PDCD6IP Fwd:	5'-CCATACCCCAGCTTGTTTGC-3'
• PDCD6IP Rev:	5'-GGGACACAAGGCTCTGTGAA-3'
• CDC20 Fwd:	5'-CATTCCCAGGTGTGCTCCAT-3'
• CDC20 Rev:	5'-CCGGGATGTGTGACCTTTGA-3'
• CCNB1 Fwd:	5'-TGCATGTAAGCCAAGTCATGGA-3'
• CCNB1 Rev:	5'-GGGACTAGGGATTCGGTGGT-3'
• AURKA Fwd:	5'-GGTAGGCCTGATTGGGTTTCT-3'
• AURKA Rev:	5'-GCCCTTAACAGCTCTGAGACA-3'

• PLK1 Fwd:	5’-GTTCTACAGCCTTGTCCCCC-3’
• PLK1 Rev:	5’-CCAAGGAAAGGACAGTTCCGA-3’
• SERPINB9 Fwd:	5’-GCTTCGTTTTTACATATGTCTTTGC-3’
• SERPINB9 Rev:	5’-AAGGCAATACAAGAATGAGAGAAAA-3’
• LHX4 Fwd:	5’-ACACAGCACAGGGGGTAATG-3’
• LHX4 Rev:	5’-CCAAGCCCCTAAGCAGAACA-3’
• MN1 Fwd:	5’-TAGCTCATGGTCCTGGCAAC-3’
• MN1 Rev:	5’-TTTCATTCTGGGGTCGTGGG-3’

3) Supplemental References

Asmann, Y.W., Hossain, A., Necela, B.M., Middha, S., Kalari, K.R., Sun, Z., Chai, H.S., Williamson, D.W., Radisky, D., Schroth, G.P., *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*, e100.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev *25*, 1915-1927.

Gadad, S.S., Shandilya, J., Swaminathan, V., and Kundu, T.K. (2009). Histone chaperone as coactivator of chromatin transcription: role of acetylation. Methods Mol Biol *523*, 263-278.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., *et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223-227.

Hah, N., Danko, C.G., Core, L., Waterfall, J.J., Siepel, A., Lis, J.T., and Kraus, W.L. (2011). A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. Cell *145*, 622-634.

Hah, N., Murakami, S., Nagari, A., Danko, C.G., and Kraus, W.L. (2013). Enhancer transcripts mark active estrogen receptor binding sites. Genome Res *23*, 1210-1223.

Hung, T., Wang, Y., Lin, M.F., Koegel, A.K., Kotake, Y., Grant, G.D., Horlings, H.M., Shah, N., Umbricht, C., Wang, P., *et al.* (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat Genet *43*, 621-629.

Kalyana-Sundaram, S., Kumar-Sinha, C., Shankar, S., Robinson, D.R., Wu, Y.M., Cao, X., Asangani, I.A., Kothari, V., Prensner, J.R., Lonigro, R.J., *et al.* (2012). Expressed pseudogenes in the transcriptional landscape of human cancers. Cell *149*, 1622-1634.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol *14*, R36.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

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

Luo, X., Chae, M., Krishnakumar, R., Danko, C.G., and Kraus, W.L. (2014). Dynamic reorganization of the AC16 cardiomyocyte transcriptome in response to TNFalpha signaling revealed by integrated genomic analyses. BMC Genomics *15*, 155.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol *28*, 495-501.

Meerbrey, K.L., Hu, G., Kessler, J.D., Roarty, K., Li, M.Z., Fang, J.E., Herschkowitz, J.I., Burrows, A.E., Ciccia, A., Sun, T., *et al.* (2011). The pINDUCER lentiviral toolkit for inducible RNA interference in vitro and in vivo. Proc Natl Acad Sci U S A *108*, 3665-3670.

Ravasi, T., Suzuki, H., Pang, K.C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M.C., Gongora, M.M.*, et al.* (2006). Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res *16*, 11-19.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S.*, et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*, 1034-1050.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Sun, M., Isaacs, G.D., Hah, N., Heldring, N., Fogarty, E.A., and Kraus, W.L. (2012). Estrogen regulates JNK1 genomic localization to control gene expression and cell growth in breast cancer cells. Mol Endocrinol *26*, 736-747.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol *31*, 46-53.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Zhong, S., Joung, J.G., Zheng, Y., Chen, Y.R., Liu, B., Shao, Y., Xiang, J.Z., Fei, Z., and Giovannoni, J.J. (2011). High-throughput illumina strand-specific RNA sequencing library preparation. Cold Spring Harb Protoc *2011*, 940-949.