SUPPLEMENTAL MATERIAL

Enhancer Transcription Reveals Subtype-Specific Gene Expression Programs Controlling Breast Cancer Pathogenesis

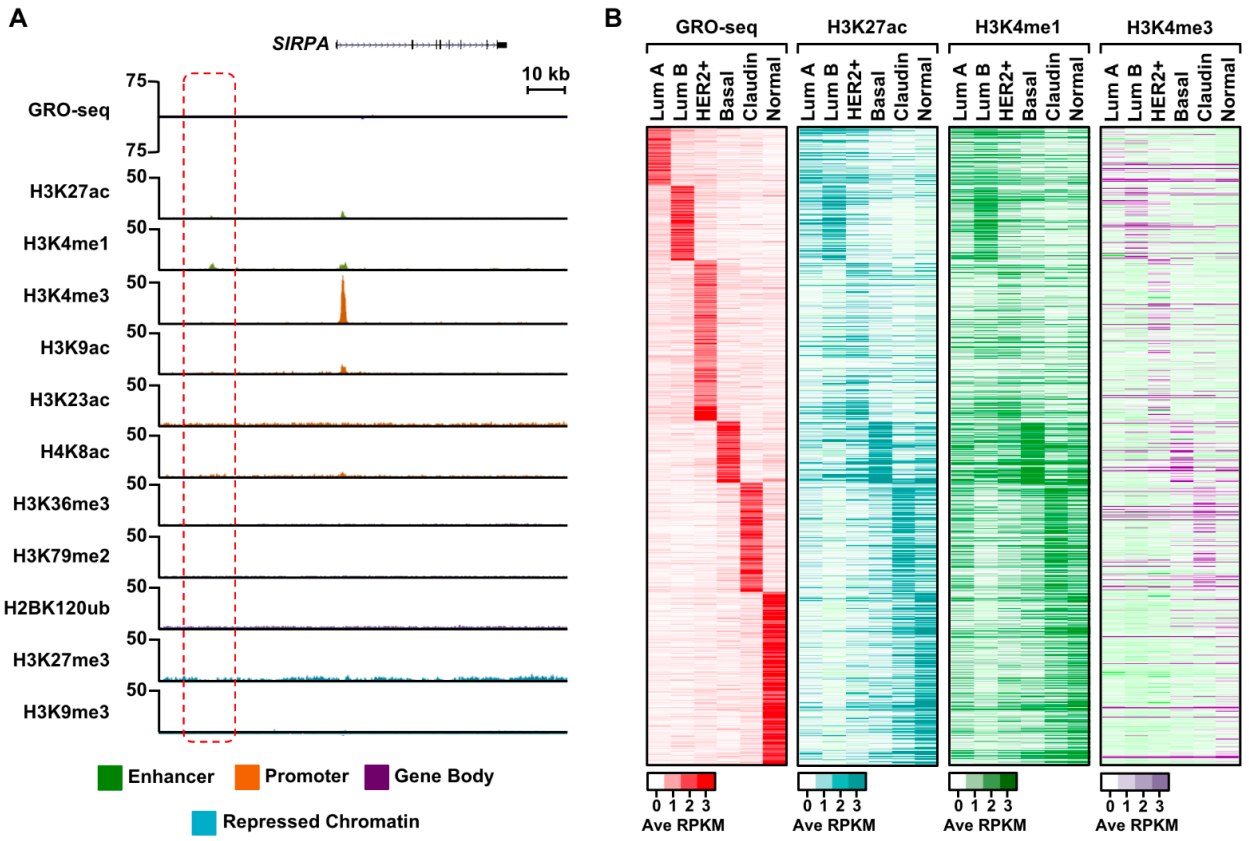
Franco *et al.* (Kraus)

Contents:

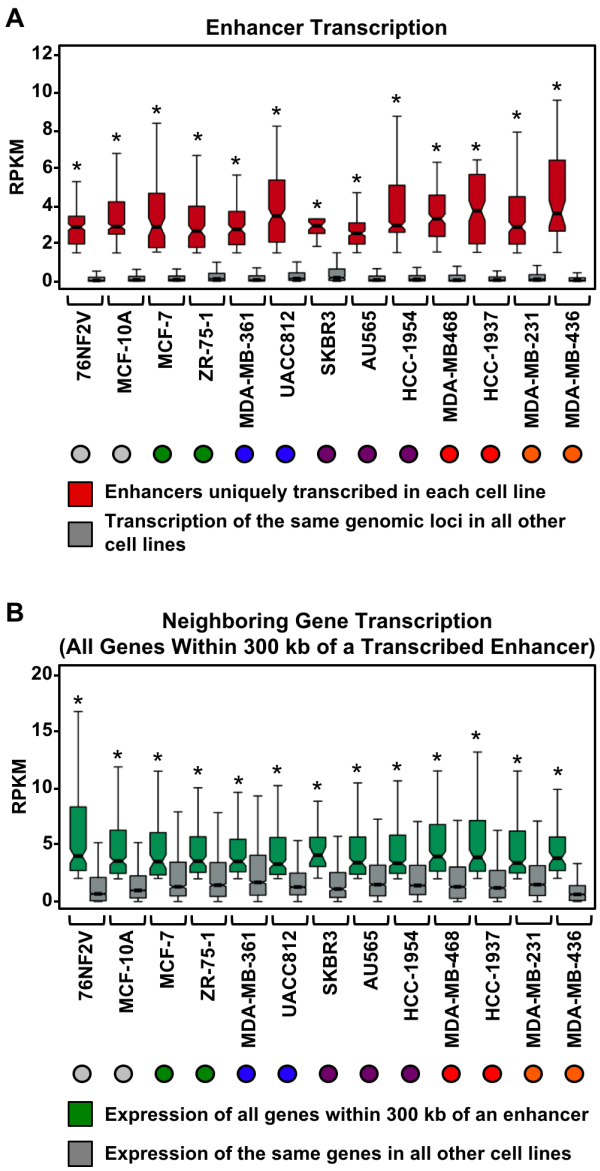
THE LONESTAR CONSORTIUM

The Lonestar Oncology Network for Epigenetics Therapy and Research (LONESTAR) Consortium comprises a group of basic science researchers at The University of Texas MD Anderson Cancer Center, Houston; The University of Texas MD Anderson Cancer Center Science Park, Smithville; The University of Texas Southwestern Medical Center, Dallas; and The Baylor College of Medicine, Houston. The LONESTAR Consortium was funded by the Cancer Prevention and Research Institute of Texas (CPRIT) from 2011 through 2016. The overarching goal of the LONESTAR Consortium is to define the epigenetic and transcriptional states that drive breast cancer formation. The participating principle investigators are: Xiaobing Shi, Michelle C. Barton, Khandan Keyomarsi (MDACC); Sharon Y.R. Dent, Mark T. Bedford (MDACC Science Park); W. Lee Kraus, Cheng-Ming Chiang (UT Southwestern); Wei Li, Orla M. Conneely, Ming-Jer Tsai, Daniel Medina (BCOM).

SUPPLEMENTAL FIGURES



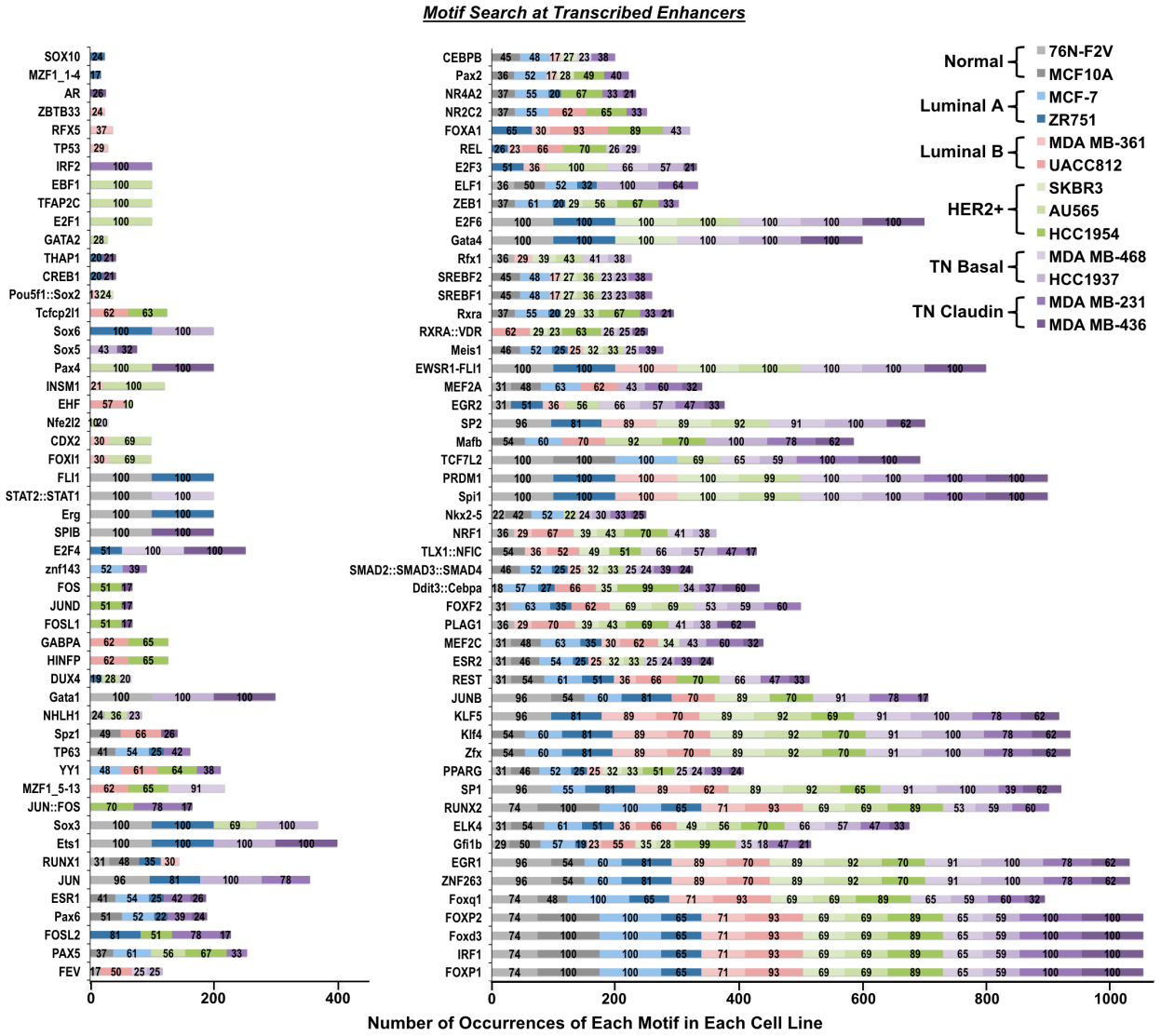
Supplemental Figure S1. Enhancer transcription is associated with other features of active chromatin and target gene expression.
(A) Genome browser views of GRO-seq and histone modification ChIP-seq data for the same genomic locus shown in Figure 1B (around the *SIRPA* gene) from a TN breast cancer cell line in which the enhancer is not transcribed (MDA-MB-231) *(red box with dashed line)*. The data include: transcription determined by GRO-seq *(red/blue)*, as well as histone modifications typically enriched at enhancers *(green)*, promoters *(brown)*, gene bodies *(purple)*, and repressed chromatin *(turquoise)* determined by ChIP-seq.
(B) Heatmap representations of genomic data for uniquely transcribed putative enhancers (short, bidirectional; see Figure 1A) compiled for cell lines representing the different molecular subtypes of breast cancer. The loci were defined by the production of enhancer transcripts and were centered on the center of the overlap between the bidirectionally transcribed eRNAs. The average RPKM value for each locus is shown (*far left panel*). Heatmaps of ChIP-seq data corresponding to the same loci shown for the GRO-seq analysis *(other panels)*. The average RPKM ChIP-seq signals for H3K27ac, H3K4me1, and H3K4me1 are shown.



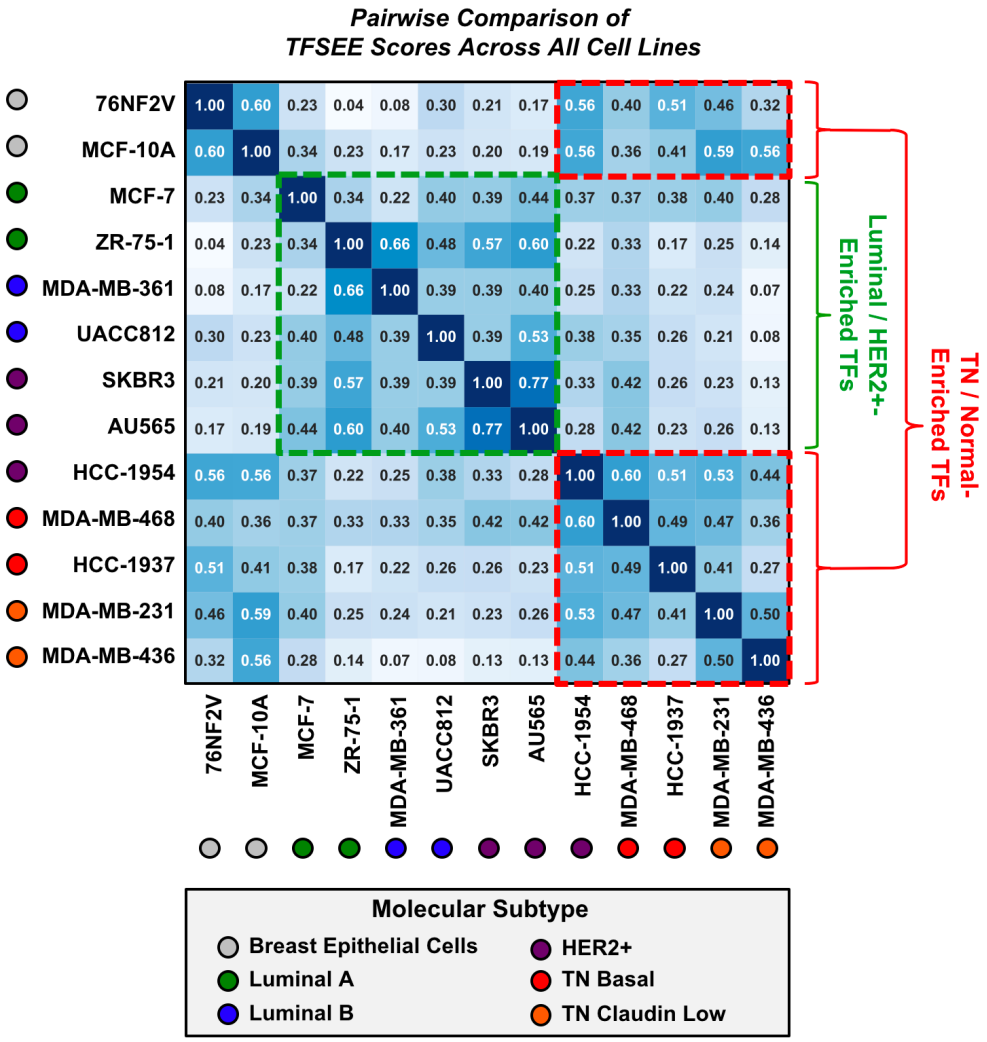
Supplemental Figure S2. Actively transcribed enhancers dictate subtype-specific transcriptional programs *(continued)*.

(A) Box plots of normalized GRO-seq read counts for enhancers uniquely transcribed in a single cell line compared to the transcription of the same genomic loci in all other cell lines. The set of enhancers in this analysis, which is distinct from those shown in Figure 3A, includes those enhancers that have one or more neighboring genes within 300 kb. Asterisks indicate significant differences between the two conditions tested for each cell line (Wilcoxon rank sum test, $p < 0.05$). Colored circles indicate the molecular subtype of each breast cancer cell line (refer to the key in Figure 2B for the color codes).

(B) Box plots of normalized GRO-seq read counts for all neighboring genes within 300 kb of a uniquely transcribed enhancer (from panel A) in a single cell line compared to the transcription of the same genes in all other cell lines. Asterisks indicate significant differences between the two conditions tested for each cell line (Wilcoxon rank sum test, $p < 0.05$).

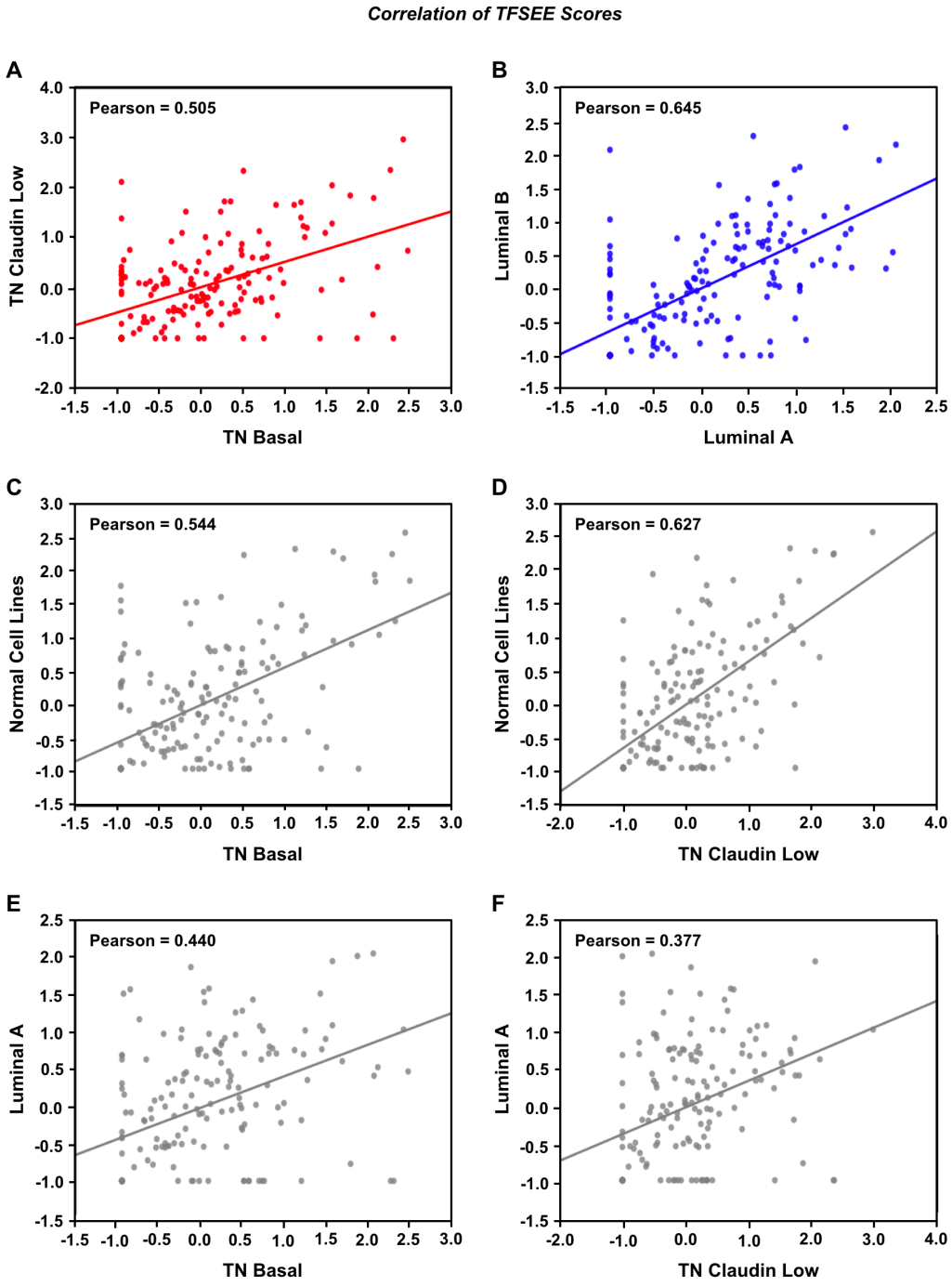


Supplemental Figure S3. Motif search at transcribed enhancers in breast cancer cells. Stacked bar chart showing enriched motifs identified at transcribed enhancers in breast cancer cell lines representing the distinct molecular subtypes in breast cancers. The analysis was performed using MEME and Tomtom/JASPAR according to the scheme shown in Figure 4A. The bars are colored by cell line. The values within each bar represent the percentage of the transcribed enhancers of a cell line that are enriched with that motif.



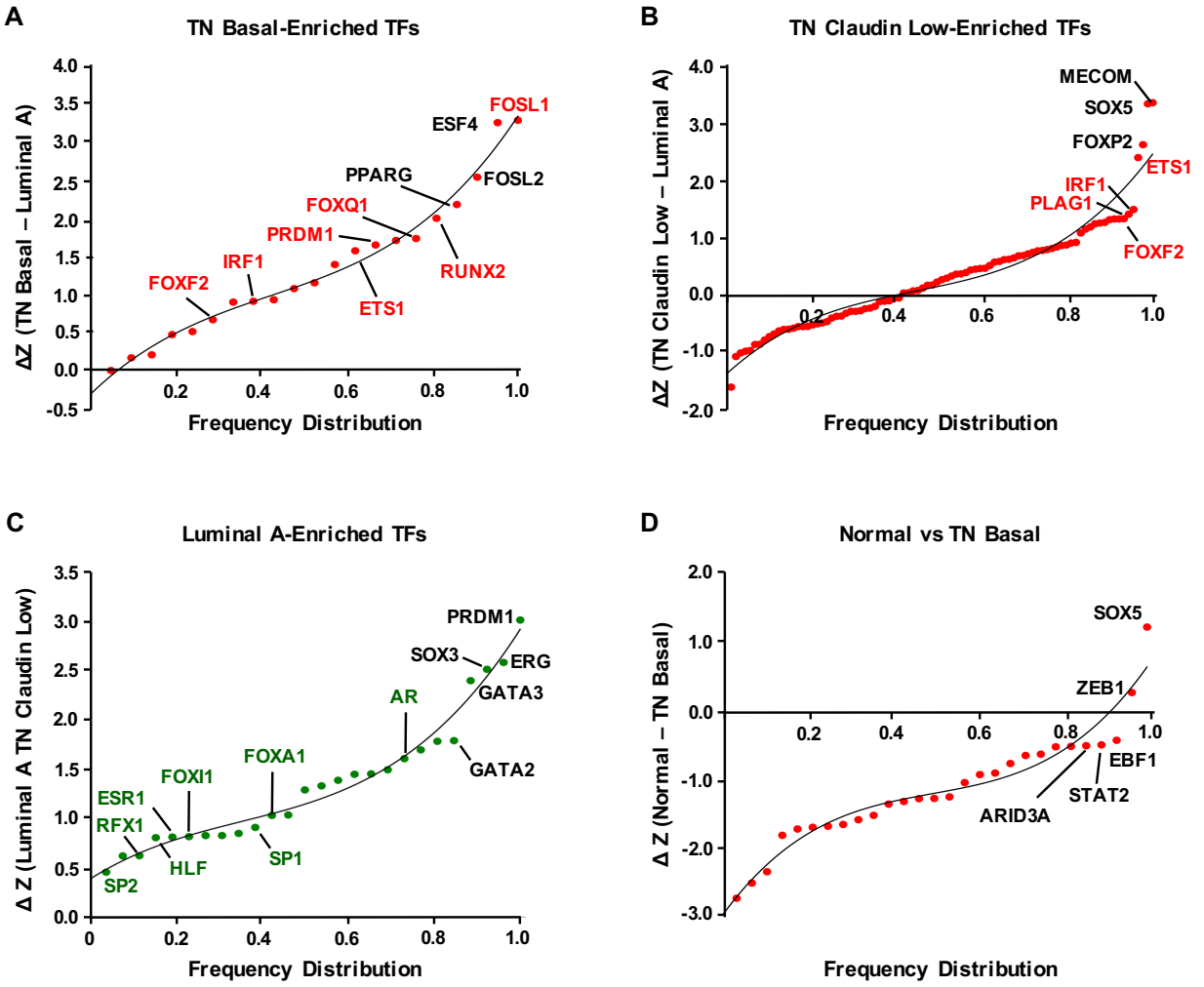
Supplemental Figure S4. Heat map of a pairwise Pearson's correlation matrix for the TFSEE scores from 13 mammary lines.

Heatmap of a pairwise Pearson's correlation analysis of TFSEE scores for all cell lines shown in Figure 1A. The shade of blue in each box represents the magnitude of positive correlation for a given pairwise comparison (darker = more highly correlated). The numbers within the boxes are the Pearson's correlation values. Red and green dashed boxes highlight groups of cell lines that generally have stronger positive correlations between their TFSEE scores than cell lines not in the group. The red dashed box corresponds to the TN / Normal-enriched clade from Figure 4B, while the green dashed box corresponds to the Luminal / HER2+-enriched clade from Figure 4B. Colored circles indicate the molecular subtype of each breast cancer cell line, as indicated.



Supplemental Figure S5. Pairwise Pearson's correlation analyses of TFSEE scores by breast cancer molecular subtype.

(A-F) Scatterplots of pairwise Pearson's correlation analyses of TFSEE scores between different molecular subtypes of breast cancer. The highest correlations were observed between related subtypes (e.g., TN Basal-TN Claudin low, panel A; Luminal A-Luminal B, panel B; TN-Normal, panels C and D) and the lowest correlations were observed between unrelated subtypes (Luminal A-TN Basal, panel E; Luminal A-TN Claudin low, panel F). A subset of all possible comparisons is shown.



Supplemental Figure S6. Rank order frequency distribution of TFs enriched based on pairwise comparisons of TFSEE scores between different molecular subtypes of breast cancer.

Rank order frequency distribution plots for TFs enriched based on pairwise comparisons of TFSEE scores between different molecular subtypes of breast cancer. TFs that were also enriched in the original unsupervised hierarchical clustering analysis (Figure 4, C and D) are highlighted in red or green. A subset of all possible comparisons is shown.

(A) TN Basal-enriched TFs (versus Luminal A).

(B) TN Claudin low-enriched TFs (versus Luminal A).

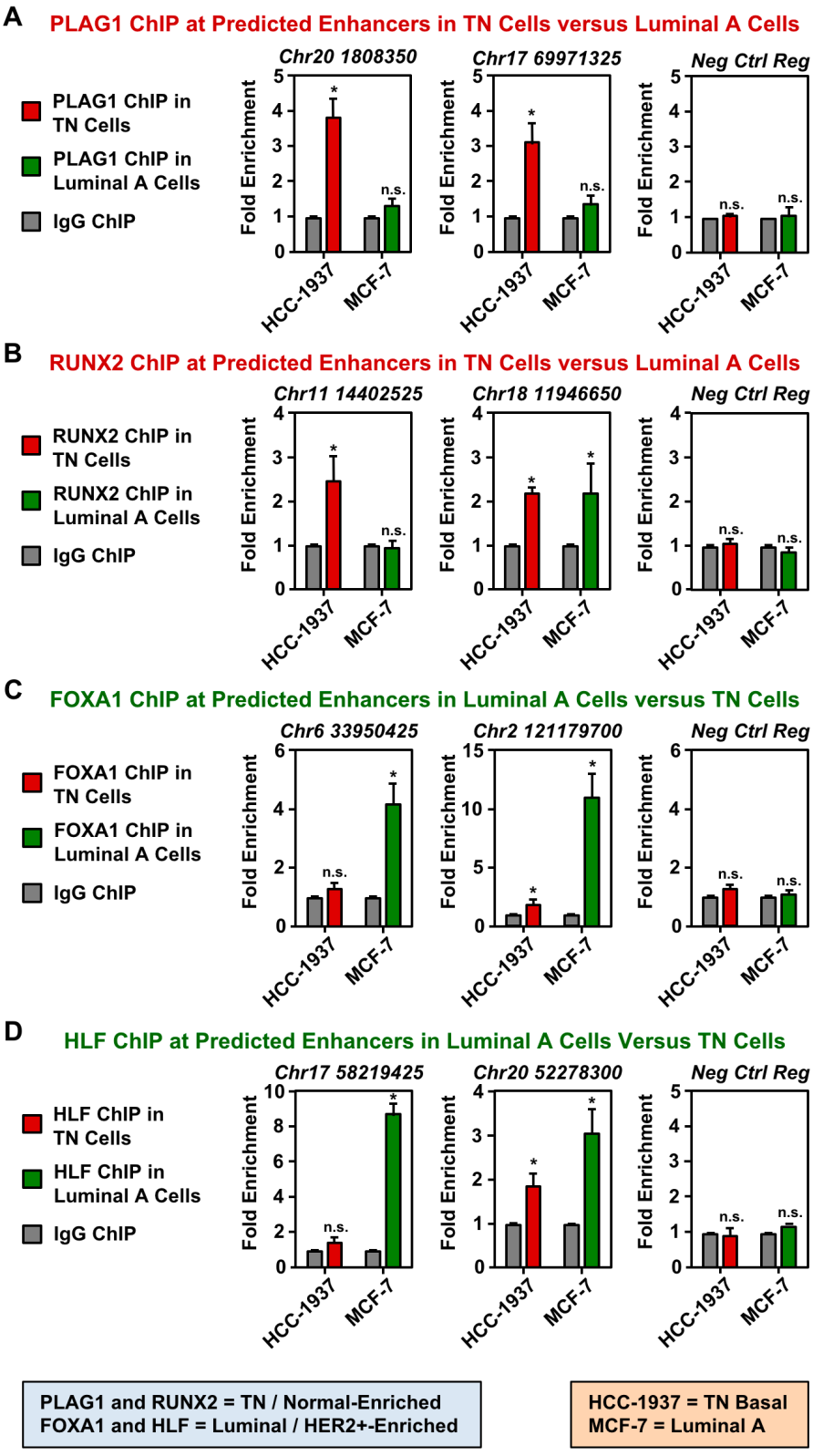
(C) Luminal A-enriched TFs (versus TN Claudin low).

(D) Normal versus TN Basal (no enrichment).

Supplemental Figure S7. TFSEE-predicted TFs are enriched at sites of enhancer transcription in distinct molecular subtypes of breast cancer.
ChIP-qPCR experiments confirming the cell type specificity of enrichment (fold over IgG control) for TN / normal-enriched TFs (PLAG1 and RUNX2) and luminal / HER2+-enriched TFs (FOXA1 and HLF) at transcribed enhancers in TN cells (HCC1937) and luminal A cells (MCF-7). The enhancers are designated by their genomic coordinates. A negative control region not predicted to be bound by the TF is shown for comparison (far right panel). Each bar represents the mean + SEM, n = 3. Asterisks indicate significant differences from the corresponding control (Student's *t*-test, p-value < 0.05). n.s., not significant (Student's *t*-test, p-value > 0.05).
(A and B) ChIP-qPCR experiments for TN / normal-enriched TFs PLAG1 (panel A) and RUNX2 (panel B), which are expected to bind in TN cells (HCC1937), but not in luminal A cells (MCF-7), as predicted by TFSEE.
(C and D) ChIP-qPCR experiments for luminal / HER2+-enriched TFs FOXA1 (panel C) and HLF (panel D), which are expected to bind in luminal A cells (MCF-7), but not in TN cells (HCC1937), as predicted by TFSEE.

[Supplemental Figure S7 is on the next page]

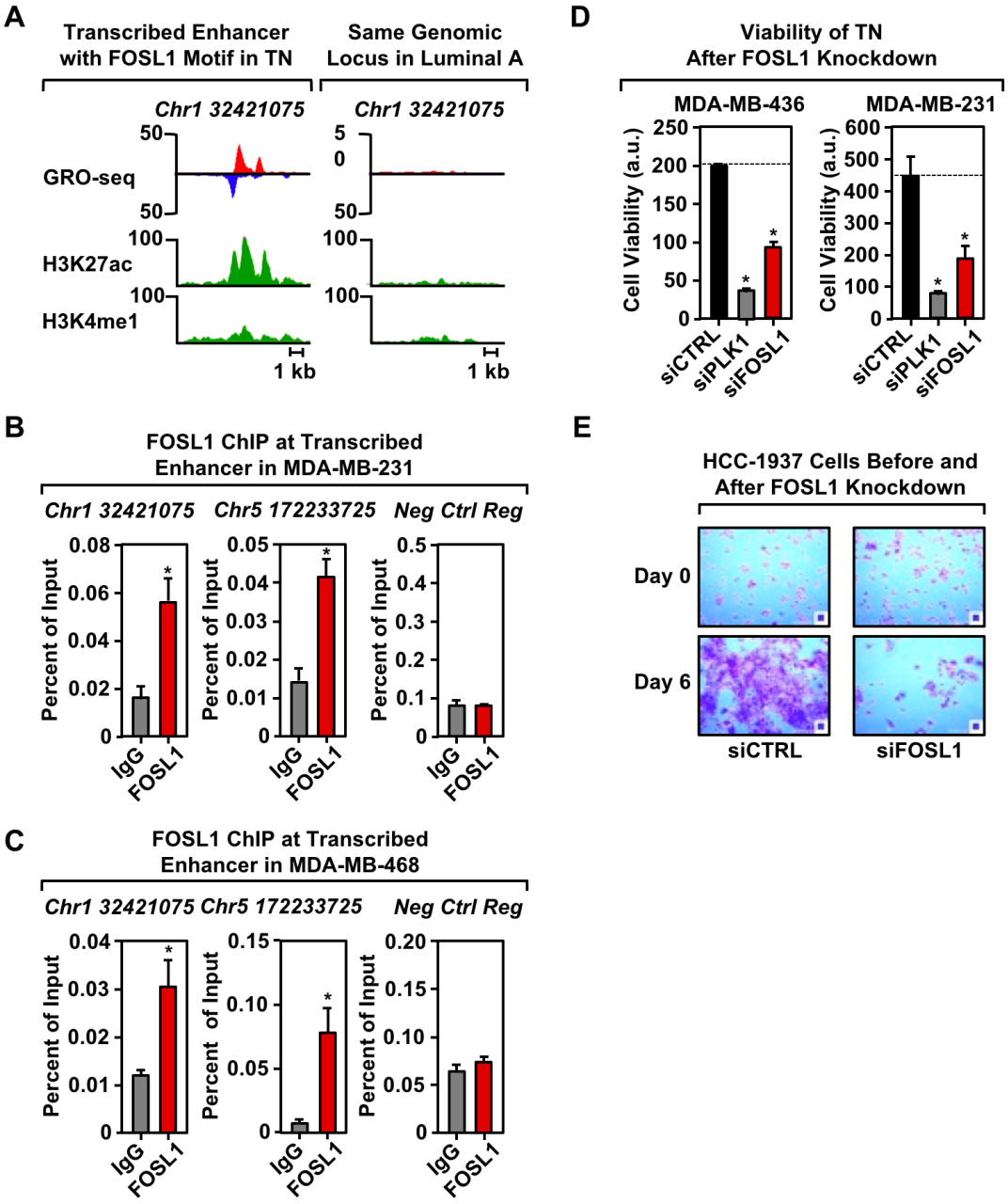
Supplemental Figure S7.

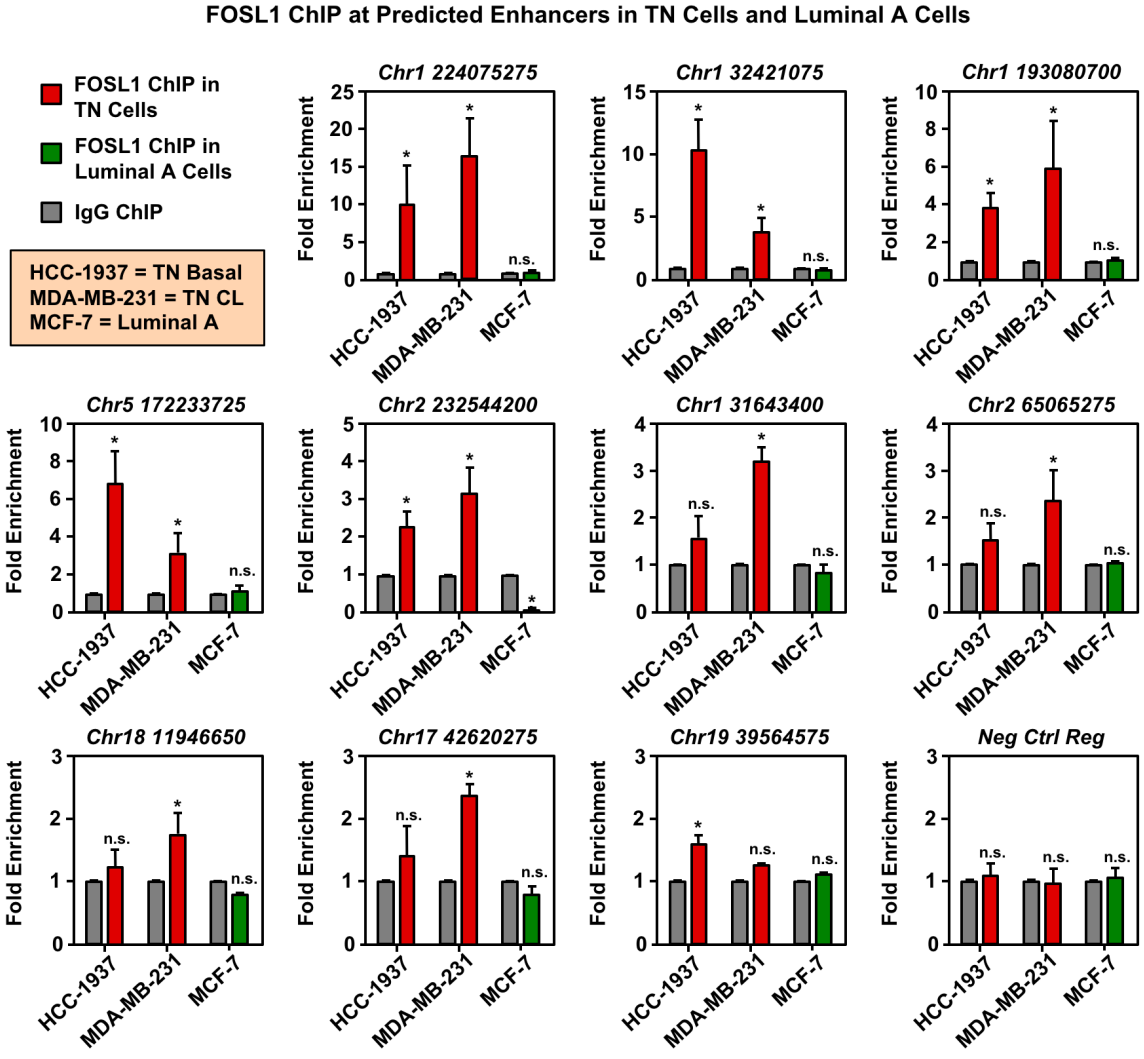


Supplemental Figure S8. FOSL1 is enriched at transcribed enhancers in TN cells, regulates cell proliferation, and correlates with breast cancer patient outcomes *(continued)*.
A similar set of experiments as those shown in Figure 6 performed in additional TN breast cancer cell lines.
(A) *(Left)* Genome browser views of an additional transcribed enhancer predicted to be bound by FOSL1, shown in a TN cell line (HCC-1937) (GRO-seq; H3K27ac and H3K4me1). *(Right)* Genome browser views of the same genomic locus in a Luminal A cell line (ZR-75-1), which is not transcribed or enriched for enhancer-related histone modifications in this cell type.
(B and C) ChIP-qPCR for FOSL1 at two transcribed enhancers predicted to be bound by FOSL1, shown in two additional TN cell lines (MDA-MB-231 and MDA-MB-468). The enhancers are designated by their genomic coordinates. A negative control region not bound by FOSL1 is shown for comparison. The enhancers are designated by their genomic coordinates. Genome browser views for the enhancer found on Chr 1 are shown in panel A. Each bar represents the mean + SEM, n = 3. Asterisks indicate significant differences from the corresponding control (Student's *t*-test, p-value < 0.05).
(D) siRNA-mediated knockdown of FOSL1 causes decreased viability in two additional TN cell lines (MDA-MB-436 and MDA-MB-231). siRNA-mediated knockdown of Polo-like Kinase 1 (PLK1) serves as a positive control. Each bar represents the mean + SEM, n = 3. Asterisks indicate significant differences from the corresponding control (Student's *t*-test, p-value < 0.05).
(E) Light microscopy images of crystal violet-stained HCC-1937 cells before and after 6 days of siRNA-mediated FOSL1 knockdown.

[Supplemental Figure S8 is on the next page]

Supplemental Figure S8.



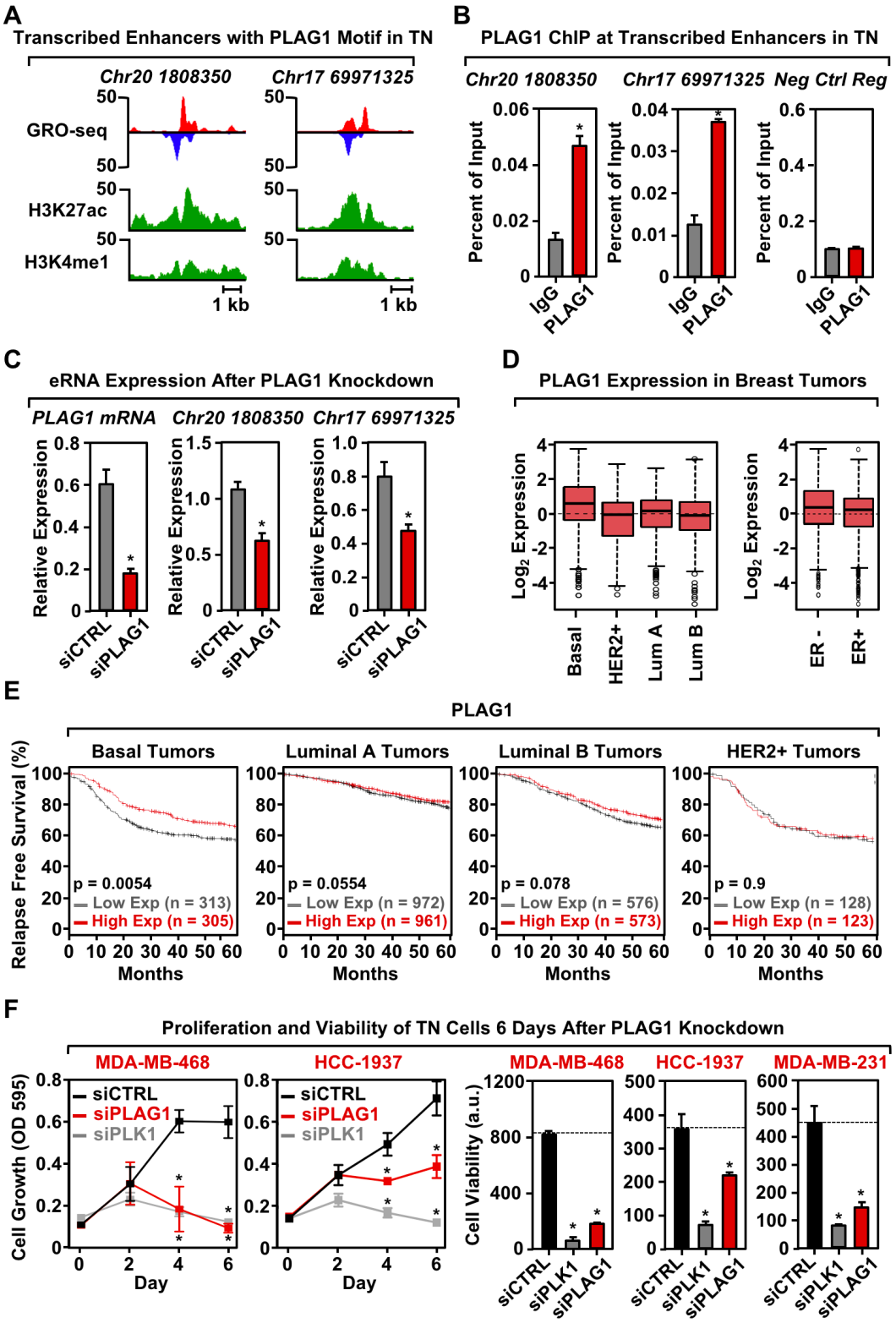


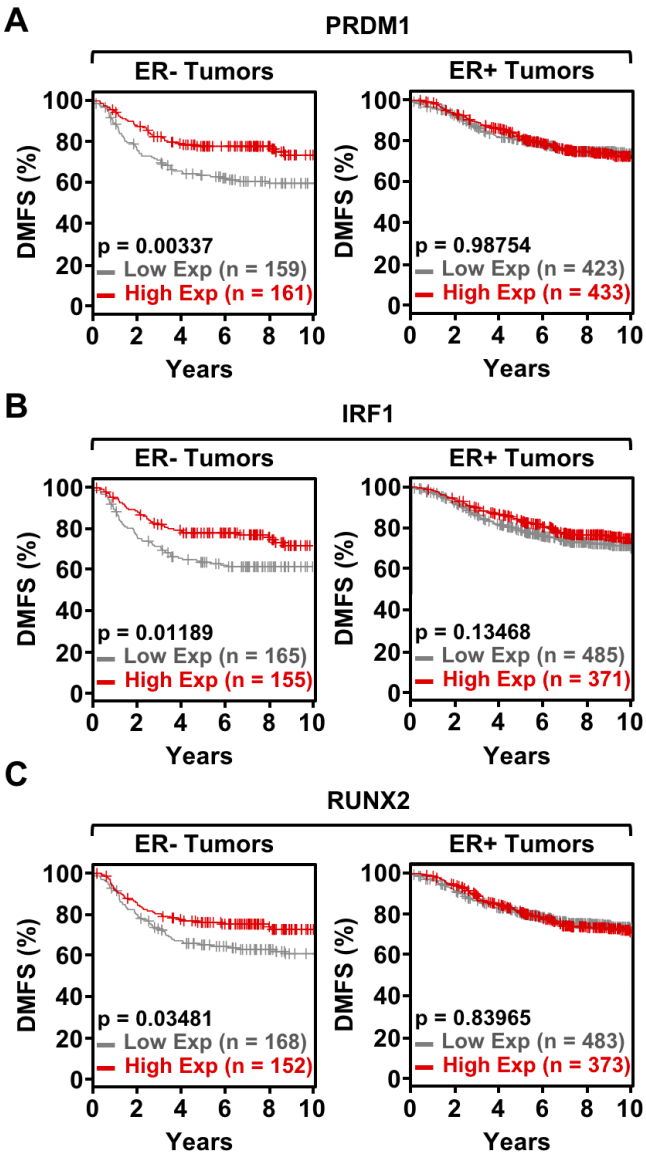
Supplemental Figure S9. FOSL1 is enriched at transcribed enhancers in TN cells.
ChIP-qPCR for FOSL1 at ten different transcribed enhancers predicted by TFSEE to be bound by FOSL1 in TN cells (HCC1937, MDA MB-231), but not in Luminal A cells (MCF-7). A negative control region not predicted to be bound by FOSL1 is shown for comparison (*bottom right panel*). Each bar represents the mean + SEM, n = 3. Asterisks indicate significant differences from the corresponding control (Student's *t*-test, p-value < 0.05). n.s., not significant (Student's *t*-test, p-value > 0.05).

Supplemental Figure S10. PLAG1 is enriched at transcribed enhancers in TN cells and regulates cell proliferation.
(A) Genome browser views of two transcribed enhancers predicted to be bound by PLAG1 in TN cells (GRO-seq; H3K27ac and H3K4me1). The data shown are from TN basal breast cancer cells (HCC-1937).
(B) ChIP-qPCR for PLAG1 at the two transcribed enhancers shown in panel A predicted to be bound by PLAG, shown in TN basal breast cancer cells (HCC-1937). A negative control region not bound by PLAG1 is shown for comparison. The enhancers are designated by their genomic coordinates. Each bar represents the mean + SEM, n = 3. Asterisks indicate significant differences from the corresponding control (Student's *t*-test, p-value < 0.05).
(C) siRNA-mediated knockdown of PLAG1 in a TN basal breast cancer cell line (HCC-1937) decreases the transcription of cognate enhancers as determined by RT-qPCR. The enhancers are designated by their genomic coordinates. Each bar represents the mean + SEM, n = 3. Asterisks indicate significant differences from the corresponding control (Student's *t*-test, p-value < 0.05).
(D) Box plots of *PLAG1* mRNA expression levels in patient tumor samples confirm enrichment of PLAG1 in Basal-like and in ER-negative (ER-) breast tumor samples, as predicted by the TFSEE analysis in breast cancer cell lines. Observed differences are significant as determined by an ANOVA comparison of the means (p-value < 0.00001).
(E) PLAG1 expression is predictive of clinical outcomes in basal breast cancers, but not in luminal or HER2+ breast cancers. The breast cancer outcome-linked gene expression data were accessed and graphed using the Kaplan Meier Plotter tool (KM plotter).
(F) siRNA-mediated knockdown of PLAG1 in TN cell lines (MDA-MB-468, HCC-1937, and MDA-MB-231) causes decreased proliferation and viability, as observed in proliferation assays (*left panels*) and cell viability assays (*right panels*). siRNA-mediated knockdown of Polo-like Kinase 1 (PLK1) serves as a positive control. Each point or bar represents the mean + SEM, n = 3. Asterisks indicate significant differences from the corresponding control (Student's *t*-test, p-value < 0.05).

[Supplemental Figure S10 is on the next page]

Supplemental Figure S10.





Supplemental Figure S11. TFs enriched in the TN / Normal clade are predictive of clinical outcomes in breast cancer.

The mRNA expression levels of three transcription factors identified by TFSEE specifically in the TN / Normal clade, PRDM1, RUNX2 and IRF1, are predictive of clinical outcomes in patients with ER-negative (ER-) breast tumors, but not patients with ER-positive (ER+) breast tumors. Kaplan-Meier survival analyses of patients expressing high levels of the TF mRNA (*maroon line*) exhibit a better outcome compared to patients expressing low levels of the TF mRNA (*grey line*). The breast cancer outcome-linked gene expression data were accessed and graphed using the Gene Expression-Based Outcome for Breast Cancer Online (GOBO) tool.

SUPPLEMENTAL TABLE

Supplemental Table S1. Transcription factors enriched in different clades identified through the total functional score of enhancer elements (TFSEE) matrix analysis.
Table of transcription factors enriched in two different clades as determined by the total functional score of enhancer elements (TFSEE) matric analysis. The Z-score used to rank the TFs integrates the following parameters: TF mRNA expression levels, TF motif p-value, magnitude of eRNA transcription, and enrichment of H3K4me1 and H3KK27.
(A) TFs enriched in the TN/normal clade ranked by Z-score.
(B) TFs enriched in the luminal/HER2+ clade ranked by Z-score.

(A) TFs enriched in the TN/normal clade

Rank	TF	ΔZ Score TN - Luminal
1	FOXF2	3.194304504
2	FOXQ1	2.301805006
3	PLAG1	2.228580118
4	RUNX2	2.028638798
5	IRF1	1.973366828
6	ETS1	1.633965203
7	FOSL1	1.443260697
8	RUNX1	1.249395006
9	PRDM1	1.240059877
10	KLF5	1.213311945
11	TP63	1.097269851
12	JUNB	1.024384922
13	FOSL2	1.020370898
14	FOXC1	0.927078187
15	POU2F2	0.899445091
16	ZEB1	0.847480853
17	MAFF	0.843349756
18	ZNF354C	0.778202441
19	NR2F1	0.696080551
20	MEF2A	0.660335267
21	NFIL3	0.533566105
22	MECOM	0.346235798
23	YY1	0.31783797
24	E2F4	0.307921934
25	TP53	0.282961028
26	NFE2L1	0.188050336
27	EHF	-0.648448309
28	TFAP2A	-0.983532911

(B) TFs enriched in the luminal/ HER2+ clade

Rank	TF	ΔZ Score Luminal - TN
1	HLF	2.104521633
2	FOXI1	1.797234192
3	ESR1	1.760643159
4	RFX1	1.656601753
5	SP2	1.488408307
6	FOXA1	1.385535823
7	SP1	1.332423935
8	AR	1.294678916
9	TFAP2C	1.195956912
10	RREB1	1.162140157
11	SOX3	1.147069415
12	ERG	1.108864723
13	ELF5	1.094882978
14	ZBTB33	1.075067081
15	E2F1	0.997912819
16	ARID3A	0.991582255
17	HINFP	0.941532838
18	GATA2	0.90307598
19	FOXP2	0.899246705
20	RFX5	0.876441689
21	REL	0.867674596
22	MAFK	0.859284193
23	CTCF	0.841827006
24	E2F6	0.777217824
25	NR1H2	0.762485571
26	HOXC9	0.741885088
27	GATA3	0.72544663
28	CEBPA	0.717536683
29	NFATC2	0.68866411
30	MZF1	0.672701223

(B) ***Continued***

Rank	TF	ΔZ Score Luminal - TN
31	RXRA	0.670080759
32	SPIB	0.669803793
33	ZNF263	0.58262325
34	MAFG	0.555525195
35	VDR	0.546786962
36	CREB1	0.542579323
37	MAFB	0.45894365
38	E2F3	0.417206583
39	ZFX	0.414207359
40	USF1	0.413209233
41	FOXP1	0.385355134
42	SREBF2	0.320347371
43	ELK4	0.298115985
44	ZNF143	0.255001184
45	ESRRA	0.24109391
46	REST	0.23857392
47	GATA4	0.226858214
48	NRF1	0.196081429
49	ARNT	0.171626383
50	NR4A2	0.163038348
51	SMAD4	0.147835226
52	STAT2	0.14239847
53	TCF12	0.118102147
54	KLF4	0.117551739
55	TCF3	0.071687777
56	GABPA	0.023786031
57	RFX2	0.023765755
58	FOXD1	0.00473939
59	JUND	5.75438E-05
60	MEIS1	-0.000198187
61	DDIT3	-0.01145404
62	ELK1	-0.026273209
63	USF2	-0.051624087
64	NFIC	-0.065309816
65	ELF1	-0.077384524
66	FOS	-0.121434803
67	FOXO3	-0.187877451
68	NR2C2	-0.187886632
69	TEAD1	-0.189190316
70	STAT5A	-0.19778881
71	FOXO1	-0.237103593
72	NR3C1	-0.256216792
73	BACH1	-0.277033541
74	NFE2L2	-0.321668546
75	JUN	-0.358374864
76	TCF7L2	-0.374987101
77	EGR1	-0.439552283
78	PPARG	-0.460925055

SUPPLEMENTAL METHODS

Cell Culture

All cell lines were purchased from the American Type Culture Collection (ATCC) and were maintained, propagated, and plated for experiments in the laboratory of Dr. Khandan Keyomarsi at the MD Anderson Cancer Center. The use of a centralized cell culture core facility facilitated consistency and reproducibility among all of the labs in the LONESTAR consortium conducting assays for this work. All collections of RNA, protein, chromatin, and nuclei were performed in the cell culture core facility and distributed to the different labs for use in the various assays described herein. The two immortalized breast epithelial cell lines, MCF-10A and 76N-F2V, were grown in D medium (described below). All other cell lines were grown in Alpha-MEM medium (Sigma, M8042). All cells were grown as adherent cultures at 37^{o}C with 6.5% CO_2.

The D medium comprised a 1:1 mixture of Alpha-MEM medium (Sigma, M8042) and Ham's F12 base medium (Fisher, MT10080CV) containing the following additives: 0.1 M HEPES, 2 mM L-glutamine, 1% fetal bovine serum (Sigma, F4135), 0.035 mg/ml of bovine pituitary extract (Hammond Cell Tech, 1078NZ), 0.01 mM ascorbic acid, 2 nM β-estradiol, 2.5 ng/mL sodium selenite, 10 nM triiodothyronine, ethanolamine, 1 μg/mL insulin, 1 ng/mL hydrocortisone, 0.1 mM phosphoethanolamine, 0.01 mg/mL transferrin, 12.5 ng/mL epidermal growth factor, and 1% Penicillin/Streptomycin. The Alpha-MEM contained the following additives: 0.1 M HEPES, 10% fetal bovine serum (Sigma, F4135), 1% non-essential amino acids, 2 mM L-glutamine, 1% sodium pyruvate, 1 μg/mL insulin, 1 ng/mL hydrocortisone, 12.5 ng/mL epidermal growth factor, and 1% Penicillin/Streptomycin.

Cell Proliferation Assays

Cell proliferation was assessed using a crystal violet staining assay. HCC1937, MDA-MB-468, and MCF-7 cells were plated at two densities, 2 x 10^4 and 4 x 10^4 cells per well, in six well plates. The cells were grown to ~50% confluence (approximately 1 to 2 days of growth) and transfected with 20 nM of siRNA (siGenome SMART Pool, Dharmacon) using 3 μL of RNAiMAX Transfection reagent from Qiagen according to the manufacturer's instructions. The cells were collected every 2 days after transfection; the cells were washed with PBS, fixed for 10 minutes with 10% formaldehyde at room temperature, and stored in PBS at 4°C until all time points had been collection. The collected cells were stained with a 0.1% crystal violet in 20% methanol solution containing 200 mM phosphoric acid. After washing to remove unincorporated stain, the crystal violet was extracted using 10% glacial acetic acid and the absorbance was read at 595 nm. All growth assays were performed a minimum of three times using independent platings of cells to ensure reproducibility.

Cell Viability Assays

Cell viability assays were performed in a 96 well plate format with cells that were transfected with siRNAs using reverse transfection methodology. For the reverse transfections, 20 nM of siRNA were mixed (in a batch master mix) with 0.1 μL of RNAiMAX transfection reagent (Qiagen) according to the manufacturer's instructions and added to the wells of a 96 well plate at before adding the cells. HCC-1937 cells were plated at 500 cells per well; MDA-MB-231 cells were plated at 1000 cells per well; and MDA-MB-468, MDA-MB-436, and MCF-7 cells were plated at 2500 cells per well. Viability was measured 6 days after transfection using

Cell Titer Glo reagent (Promega) according to the manufacturer's instructions. All cell viability assays were performed a minimum of three times using independent platings of cells to ensure reproducibility.

Kaplan-Meier and Gene Expression Analyses in Patient Tumor Samples

Kaplan-Meier estimators (Kaplan and Meier 1958; Dinse and Lagakos 1982) were generated using the Gene Expression-Based Outcome for Breast Cancer Online (GOBO) tool (http://co.bmc.lu.se/gobo/) (Ringner et al. 2011) and the KM Plotter Tool (Szasz et al. 2016). Gene expression levels in patient tumor samples were also obtained using the GOBO tool.

Genomic Data Sequencing Alignments

Genomic reads from the deep sequencing experiments described below (GRO-seq, RNA-seq, and ChIP-seq) were mapped to the human reference genome hg19 using the read aligners indicated below. Realigning the reads to the more recent build of the human genome, GRCh38, would not significantly affect the analyses and the conclusions presented herein because the sequence content and coverage is largely unchanged between hg19 and GRCh38. Although the GRCh38 build provides alternate sequences, which may help in capturing variations that are not represented in hg19 (a single representation of multiple genomes), this additional information is not relevant to our studies since our conclusions and the focus of our paper are not about genetic variations and SNPs. Furthermore, the comprehensive lncRNA annotations from LNCipedia that we used in our paper were not available until May 2016, well after many of our analyses were completed. Thus, hg19 was best suited for the hypotheses, conclusions, and focus of our paper, and our results would remain largely unchanged by using GRCh38.

Preparation and Sequencing of GRO-seq Libraries

Nuclei Preparation. All cells were grown using the growth conditions listed above at the LONESTAR Consortium cell culture core facility. The cells were collected at ~70-80% confluence. Nuclei were isolated from fresh (not frozen) cells, as described previously (Luo et al. 2014). Briefly, the cells were washed three times with ice-cold PBS and swollen in Hypotonic Lysis Buffer [10 mM Tris•HCl pH 7.4, 0.5% NP-40, 10% glycerol, 3 mM $CaCl_2$, 2 mM $MgCl_2$, 1 mM DTT, 1x protease inhibitor cocktail (Sigma-Aldrich), and SUPERase-In (Ambion)]. The swollen cells were collected by centrifugation at 500 x *g* for 10 min at 4°C. The cell pellets were resuspended in 1.5 ml of Hypotonic Lysis Buffer and pipetted up and down through a narrow opening 30 to 50 times to release the nuclei. The nuclei were collected by centrifugation and washed once with 1 mL of ice cold Lysis Buffer. The nuclei were collected by again centrifugation and washed once with 1 mL of Hypotonic Lysis Buffer. After a final collection by centrifugation, the resulting pelleted nuclei were resuspended in 500 µL of Freezing Buffer (50 mM Tris•HCl pH 8.3, 40% glycerol, 5 mM $MgCl_2$, 0.1 mM EDTA, and 4 units/mL of SUPERase-In per mL), counted, frozen in liquid nitrogen in 100 µL aliquots containing 5 x 10^6 nuclei, and stored at -80°C until use.

Nuclear Run-On, GRO-seq Library Preparation, and Sequencing. Nuclear run-on and GRO-seq library preparation were performed as previously described (Hah et al. 2011), with modifications (Danko et al. 2013; Luo et al. 2014). Libraries were prepared from two biological replicates using a circularized ligation-based protocol for adaptor addition for improved efficiency of library preparation, reduced sequence bias, and ease for barcoding. Briefly, nuclear run-ons were extended for ~100 bases in the presence of Sarkosyl (to prevent reengagement of

RNA polymerases), rNTPs, α^{32}P-CTP, and 5-bromo-UTP. The nascent RNAs were isolated, hydrolyzed to ~150 bases, and enriched using α-BrdUTP antibody-conjugated agarose beads (Santa Cruz). Enriched run-on RNA was subjected to poly(A) tailing with *E. coli* PolyA Polymerase and subsequently converted to cDNA by reverse transcription using SuperScript III reverse transcriptase (Invitrogen) and an RT primer (pGATCGTCGGACTGTAGAACTCT/idSp/CCTTGGCACCCGAGAATTCCATTTTTTTTTT TTTTTTTTTTVN), where p indicates phosphorylation, idSp indicates dSpacer Furan, and VN indicates degenerate nucleotides. Reverse transcribed cDNAs were size selected (120 to 400 bp) by denaturing PAGE and circularized with CircLigase (Epicentre) to position Illumina adapter sequences relative to the 5' and 3' ends of reverse-transcribed run-on cDNA. The circularized cDNAs were re-linearized by with ApeI (NEB), purified by phenol-chloroform extraction, and PCR amplified using unique Illumina TrueSeq small RNA sample barcoded primers with Phusion high fidelity DNA polymerase (NEB). The PCR products (175 bp to 300 bp) were separated by native PAGE, eluted, and purified. After library quality control assessment using a Bioanalyzer (Agilent), the samples were subjected to 50 bp single-end sequencing using an Illumina HiSeq 2000 Sequencing System. At least two biological replicates were sequenced for each cell line. The replicates were re-sequenced to achieve a minimum of ~115 M raw reads per cell line. Thus, this experiment has 2 biological and 2 technical replicates.

Analysis of GRO-seq Data

The GRO-seq data were analyzed using the groHMM package as described previously (Hah et al. 2011; Danko et al. 2014; Luo et al. 2014; Chae et al. 2015) and the approaches described below. Additional information about the analyses can be obtained by contacting the corresponding author (W.L.K.).

Quality Control and Trimming. Quality control for the GRO-seq data was performed using the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). GRO-seq reads were trimmed to remove adapter contamination and poly(A) tails using the default parameters of Cutadapt software (Martin 2011).

Read Alignment and Gene Annotation. Reads >32 bp long were retained for alignment to the human reference genome, including autosomes, X chromosome, one complete copy of an rDNA repeat (GenBank ID: U13369.1), and the coding and lncRNA gene annotations described below using the BWA aligner v 0.6.1 (Li and Durbin 2010). A collection of coding gene annotations was built by combining the RefSeq, UCSC, and Gencode v. 14 databases. Overlaps and redundancies were removed from the combined gene lists to eliminate the possibility of double counting. A collection of >30,000 lncRNA gene annotations was obtained from the LNCipedia 2.0 database. All of the biological and technical replicates for a particular cell line were merged after confirming a strong positive correlation.

Transcript calling. Transcript calling was performed using groHMM, a two-state hidden Markov model-based algorithm as described previously (Hah et al. 2011; Danko et al. 2014; Chae et al. 2015) on each individual cell line. The shape setting parameters and -log transition probabilities used to predict the transcription units for each cell lines were as follows:

Cell Line	Shape Setting	-Log Transition Probability
76N-F2V	200	15
MCF-10A	300	25
MCF-7	300	25

ZR-75-1	200	20
MDA-MB-361	250	20
UACC812	300	20
SKBR3	250	25
AU565	250	25
HCC-1954	250	15
MDA-MB-468	200	20
HCC-1937	250	25
MDA-MB-231	300	25
MDA-MB-436	250	25

We then built a universe of transcripts by merging the groHMM-called transcripts from individual cell lines and stratifying the boundaries to remove overlaps/redundancies occurring from the union of all transcript calls.

Enhancer Transcript Calling. We filtered and collected a subset of short intergenic transcripts <9 kb in length and >5 kb away from the 5' or 3' ends of annotated genes. These were further classified into (1) short paired eRNAs and (2) short unpaired eRNAs as described previously (Hah et al. 2013). For the short paired eRNAs, the sum of the GRO-seq RPKM values for both strands of DNA was used to call an enhancer transcript pair as expressed using a criterion of RPKM ≥ 2. For the short unpaired eRNAs, an RPKM cutoff of ≥ 3 was used call an enhancer transcript as expressed. The universe of expressed eRNAs (short paired and short unpaired) was assembled using the cutoffs noted above for each cell line and was used for further analyses.

Nearest Neighboring Gene Analyses and Box Plots. The universe of expressed genes in each cell line was determined from the GRO-seq data using an RPKM cutoff ≥ 2. The set of nearest neighboring expressed genes for each enhancer defined by an expressed eRNA was determined for each cell line. Box plot representations were used to express the levels of transcription for each called enhancer and their nearest neighboring expressed genes. The read distribution (RPKM) for each eRNA or gene was calculated and plotted using the box plot function in R. Wilcoxon rank sum tests were performed to determine the statistical significance of all comparisons.

Motif Analyses. De novo motif analyses were performed on a 1 kb region (± 500 bp) surrounding the peak summit or the transcription start site for short paired and short unpaired eRNAs, respectively, using the command-line version of MEME (Bailey et al. 2009). The following parameters were used for motif prediction: (1) zero or one occurrence per sequence (-mod zoops); (2) number of motifs (-nmotifs 15); (3) minimum, maximum width of the motif (-minw 8, -maxw 15); and (4) search for motif in given strand and reverse complement strand (-revcomp). The predicted motifs from MEME were matched to known motifs using Tomtom (Gupta et al. 2007).

RNA Isolation and Gene Expression Analyses

All cells were grown using the growth conditions listed above at the LONESTAR Consortium cell culture core facility. The cells were collected at ~70-80% confluence. RNA isolation for RT-qPCR and RNA-seq was performed using the RNeasy Mini Kit (Qiagen). Changes in the expression of eRNAs and mRNAs were analyzed by RT-qPCR, as previously described (Franco et al. 2015) with a few modifications. Two micrograms of total RNA were

reverse-transcribed using annealed random hexamer primers (Sigma-Aldrich) using 600 units of MMLV reverse transcriptase (Promega) to generate cDNA. The cDNA was treated with 3 units of RNase H (Ambion) for 30 min at 37°C and then analyzed by qPCR using the primer sets listed below and a LightCycler 480 real-time PCR thermocycler (Roche) for 45 cycles. Expression changes were normalized to the levels of β-actin mRNA as an internal standard. All experiments were conducted a minimum of three times with independent RNA isolations to ensure reproducibility.

Preparation and Sequencing of RNA-seq Libraries

Total RNA was isolated as described above. The integrity of the RNA was assessed and verified using an Experion Automated Electrophoresis System (Bio-Rad) before mRNA-seq libraries were prepared using methods described previously (Zhong et al. 2011). Briefly, poly(A)+ RNA was enriched using Dynabeads oligo(dT)25 (Invitrogen), heat fragmented, and reverse transcribed using random hexamers in the presence of dNTPs. Second strand cDNA synthesis was performed with dNTPs, but replacing dTTP with dUTP. After end-repair, dA-tailing, ligation to adaptors containing barcode sequences, and size selection using AMPure beads (Agencourt), the synthesized second-strand was digested using uracil DNA glycosylase (Enzymatics). A final PCR reaction was performed using Phusion high-fidelity DNA polymerase (NEB). After library quality control assessment using a Bioanalyzer (Agilent), the samples were subjected to 50 bp single-end sequencing using an Illumina HiSeq 2000 Sequencing System. At least two biological replicates were sequenced for each cell line. The replicates were re-sequenced to achieve a minimum of ~65 M raw reads per cell line. Thus, this experiment has 2 biological and 2 technical replicates.

Analysis of RNA-seq Data

The raw data were subjected to QC analyses using the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The reads were then mapped to the human reference genome using the default parameters in TopHat (v2.0.12) (Trapnell et al. 2009). For expression analyses, we created a collection of annotations by merging Gencode (v.19) and a set of previously unannotated lncRNAs identified in MCF-7 cells (Sun et al. 2015). FPKM values were calculated per gene using Cufflinks (v.2.1.2) (Trapnell et al. 2010) using the merged annotation GTF file.

Chromatin Immunoprecipitation (ChIP) and Enrichment Analyses

Antibodies for ChIP. The following antibodies were used qPCR- and sequencing-based ChIP assays in the amounts specified:

Modification or Factor	Amount per IP	Company	Catalog No.
H3K4me1	5 µg	Abcam	ab8895
H3K4me3	5 µg	Abcam	ab8580
H3K9ac,	5 µg	EMD Millipore	07-352
H3K9me3	5 µg	Abcam	ab8898
H3K27ac	5 µg	Abcam	ab4729
H3K27me3	5 µg	Millipore	07-449
H3K36me3	5 µg	Abcam	ab9050
H3K79me2	5 µg	Abcam	ab3594

H2BK120ub1	5 µg	Millipore	05-1312
H3K23ac	5 µg	Millipore	07-355
H4K8ac	5 µg	Millipore	07-328
PLAG1	10 µg	Thermo Scientific	PA5-32188
RUNX2	10 µg	Santa Cruz Biotech	(M-70) sc-10758
HLF	10 µg	Santa Cruz Biotech	(H-71) sc-367607
FOSL1 (Fra-1)	10 µg	Santa Cruz Biotech	(N-17) sc-183
FOXA1	10 µg	Abcam	ab23738

ChIP and qPCR. ChIP was performed as previously described (Franco et al. 2015) with a few modifications. The cells were grown to ~70-80% confluence, cross-linked with 1% formaldehyde for 10 min at 37°C, and quenched in 125 mM glycine for 5 min at 4°C. The cells were then collected and lysed in Farnham Lysis Buffer [5 mM PIPES pH 8.0, 85 mM KCl, 0.5% NP-40, 1 mM DTT, and 1x protease inhibitor cocktail (Sigma-Aldrich)]. The crude nuclear pellet was collected by centrifugation, resuspended in lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris•HCl pH 7.9, 1 mM DTT, and 1x protease inhibitor cocktail), and incubated on ice for 10 minutes. The chromatin was sheared by sonication at 4°C using a Bioruptor 300 at the highest setting for fifteen 1-minute cycles of 30 seconds on and 30 seconds off to generate chromatin fragments of ~200-400 bp in length. The soluble chromatin was diluted 1:10 with dilution buffer (20 mM Tris•HCl, pH 7.9, 0.5% Triton X-100, 2 mM EDTA, 150 mM NaCl, 1 mM DTT and 1x protease inhibitor cocktail) and pre-cleared with protein A agarose beads. Five percent of the material was removed and saved as input, and the rest of the pre-cleared supernatant was incubated overnight at 4°C with the antibody of interest and a non-specific IgG control antibody (each 15 cm dish yielded three immunoprecipitations).

The following day, the immune complexes were collected by adding protein A agarose beads and incubating for 2 hours at 4°C. The immunoprecipitated material was washed once with low salt wash buffer [20 mM Tris•HCl pH 7.9, 2 mM EDTA, 125 mM NaCl, 0.05% SDS, 1% Triton X-100, and 1x protease inhibitor cocktail], once with high-salt wash buffer (20 mM Tris•HCl pH 7.9, 2 mM EDTA, 500 mM NaCl, 0.05% SDS, 1% Triton X-100, and 1x protease inhibitor cocktail), once with LiCl wash buffer (10 mM Tris•HCl pH 7.9, 1 mM EDTA, 250 mM LiCl, 1% NP-40, 1% sodium deoxycholate, and 1x protease inhibitor cocktail), and twice with Tris-EDTA (TE) containing 1x protease inhibitor cocktail. The immunoprecipitated material was eluted at room temperature in elution buffer (100 mM $NaHCO_3$, 1% SDS), and the crosslinks were reversed by adding 100 mM NaCl with incubation at 65°C overnight. The eluted material was then digested with proteinase K and RNase H to remove protein and RNA, respectively, and the enriched genomic DNA was extracted with phenol:chloroform:isoamyl alcohol followed by ethanol precipitation. The ChIPed DNA was dissolved in water and analyzed by qPCR using the enhancer- or gene-specific primers listed below, or used for ChIP-seq as described below. All ChIP-qPCR experiments were conducted a minimum of three times with independent cell platings to ensure reproducibility.

Preparation and Sequencing of ChIP-seq Libraries

ChIP-seq libraries were prepared using a modified Kapa LTP Library Preparation kit (KAPA Biosystems, cat# KK8232) for Illumina Platforms (Xi et al. 2017). Ten ng of sheared DNA was used to repair the ends of the damaged fragments using a proprietary master mix. The resulted blunted fragments were 3' A-tailed using a proprietary mixture of enzymes to allow

ligation to the specific NexTflex adaptors from Bioo Scientific (Bioo Scientific, cat# 514102). Each of the steps (i.e., end repair, 3'A tailing, and adaptor ligation) was followed by column clean up (Qiagen, cat# 28204). After adapter ligation, DNA enrichment was performed using Kapa HiFi Hot Start Ready PCR mix, and a cocktail of primers (1 cycle at 98°C for 45 seconds; 4 cycles at 98°C for 15 seconds, 60°C for 30 seconds, and 72°C for 30 seconds; and 1 cycle at 72°C for 1 minute), and purified with AmpureXP beads (Beckman Coulter, cat# A63881). The quality of the final libraries was assessed using a 2200 TapeStation (Agilent Technologies). The libraries were quantified using a Kapa Library Quantification Kit (KAPA Biosystems, KK4933) and loaded in a flow cell for cluster generation using the Illumina cBOT (Illumina) at final concentration of 10 pM. After clustering generation, the samples were sequenced using a HiSeq 2500 sequencer (Illumina; Single-end reads, 36 bp for all samples). At least two biological replicates were sequenced for each cell line for a minimum of ~100 M raw reads per cell line.

Analysis of ChIP-seq Data

The raw reads were aligned to the human reference genome using default parameters in Bowtie (ver. 1.0.0) (Langmead et al. 2009). The aligned reads were subsequently filtered for quality and uniquely mappable reads using SAMtools (ver. 0.1.19) (Li et al. 2009) and Picard (ver. 1.127; http://broadinstitute.github.io/picard/). Library complexity was measured using BEDTools (v 2.17.0) (Quinlan and Hall 2010) and met minimum ENCODE data quality standards (Landt et al. 2012). Relaxed peaks were called using MACS (v2.1.0) (Feng et al. 2012) with a p-value = 1 x 10^{-2} for each replicate, pooled replicates, and pseudoreplicates. Called peaks from the pooled replicate that were observed in both replicates or in both pseudoreplicates were used for subsequent analyses.

Genomic Data Sets

We generated libraries from at least 2 distinct biological replicates for each of the three assays (GRO-seq, RNA-seq, and ChIP-seq), with a total minimum sequencing depth of ~115 M raw reads per cell line for GRO-seq, ~65 M raw reads per cell line for RNA-seq, and ~100 M raw reads per cell line for ChIP-seq.

The genomic data sets generated for this study can be accessed from the NCBI's Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/) using the following accession numbers:

- GRO-seq: GSE96859
- RNA-seq: GSE96860
- ChIP-seq: GSE85158

Predicting Breast Cancer Subtype-Specific TFs from Transcribed Enhancers and Other Genomic Data

We developed a pipeline in Python (Total Functional Score of Enhancer Elements or TFSEE) that combines GRO-seq, RNA-seq, and ChIP-seq data with TF motif information to predict TF driving the formation of active enhancers in each breast cancer cell line, as well as the locations of the cognate enhancers. The TFSEE scripts are provided in a separate compressed folder.

Normalization of Enhancer Transcription using GRO-seq. For each cell line, C, we quantified the GRO-seq reads, RPKM, that fall within a 1 kb region around the center of the overlap for paired enhancer transcripts or from the 5′ end of unpaired enhancer transcripts, E.

We represented this profile as a matrix G_{CxE} of RPKM values. Before log_2-scaling the data, a small value was added to the RPKM to avoid taking the log of zero. For each cell line, enhancer signals were first scaled to the median and interquartile range using sklearn.preprocessing.RobustScaler and then transformed to be a number between 0 and 1 using sklearn.preprocessing.MinMaxScaler.

Normalization of Enhancer Activity using ChIP-seq. For each cell line, C, we quantified the ChIP-seq reads, RPKM, from H3K4me1, H3K27ac, and input for each enhancer within the universe of GRO-seq-defined enhancers, E. We represented this profile as a matrix M_{CxE} or H_{CxE} of RPKM values for H3K4me1 and H3K27ac, respectively. Enrichment was calculated by dividing the normalized read count from histone modification ChIP (i.e., H3K4me1 or H3K27ac) by the normalized read counts from the input. For each cell line, enrichment signals were transformed to be a number between 0 and 1 using sklearn.preprocessing.MinMaxScaler.

Total Enhancer Activity. The enhancer activity for any set of E enhancers for a set of C cell lines was represented here as a matrix A_{CxE}. We assumed that the enhancer activity of each cell type is linearly correlated to the amount enhancer transcription defined by GRO-seq (G_{CxE}), and to the enrichment of H3K4me1 and H3K27ac ChIP-seq signal (M_{CxE} and H_{CxE}) at each enhancer location. A_{CxE} was calculated using the following formula:

$$A = G + M + H$$

Normalizing Motif Predictions. De novo motif analyses were performed on a 1 kb region of expressed enhancers for each cell line using the command-line version of MEME (Bailey et al. 2009). The predicted motifs from MEME were matched to known motifs using Tomtom (Gupta et al. 2007) and JASPAR (Mathelier et al. 2014). For each cell line, we calculated the probability of a given motif, F, for a given set of E enhancers represented here as a matrix $T_{(ExF)}$. All heterodimer motifs were split into their constituent targets and assigned the same p-value. If a given motif target was represented multiple times for a given enhancer location, we represented the motif target as a single p-value using scipy.stats.combine_pvalues and Stouffer's method. For each motif, the p-values were scaled across all enhancer locations to be a number between 0 and 1. The probability of any given motif for the universe of enhancers was calculated using the following formula and scaled across all enhancer locations to be a number between 0 and 1:

$$T_{(ExF)} = \sum_{i=1}^{i=C} T_{(ExF)i}$$

Normalizing Transcription Factor Expression using RNA-seq. For each cell line, C, we quantified the RNA-seq reads, FPKM, for each F transcription factor that is a binding target for the motifs. We represented this profile as a matrix R_{CxF} of FPKM values. For any motifs representing that are binding targets of fused transcription factors (e.g., EWSR1-FLI1), we used the lowest FPKM between the two factors. Before log_2 scaling the data, a small value was added to the FPKM to avoid taking the log of zero. For each cell line, enhancer signals were first scaled to the median and interquartile range using sklearn.preprocessing.RobustScaler and then transformed to be a number between 0 and 1 using sklearn.preprocessing.MinMaxScaler. Expression values of transcription factors with FPKM values less than 0.4 were set to 0.

Determining the Total Functional Score of Enhancer Elements (TFSEE) and Generating Heatmap. To identify the transcription factors driving enhancer formation in each breast cancer cell line, we calculated a matrix I_{CxF} using the following formula:

$$I_{CxF} = (A_{CxE} \times T_{FxE}) \circ R_{CxF}$$

For each cell line, the functional scores were Z-score normalized. To identify cognate transcription factors by subtype, we performed hierarchical clustering by calculating the Euclidean distance. The rank order of the transcription factors that were enriched between clades, B, was calculated using the following formula:

$$\Delta Z = \overline{B1} - \overline{B2}$$

Pairwise Pearson's Correlation Analyses. To determine the correlations between the TFSEE scores of each breast cancer cell line, we performed pairwise Pearson's correlation analyses. To determine the Pearson's correlations between molecular subtypes, we calculated the average TFSEE score for all the cell lines within a given subtype.

Oligonucleotide Sequences for RT-qPCR, ChIP-qPCR, and siRNA-mediated Knockdown

The following oligonucleotide sequences were used for RT-qPCR, ChIP-qPCR, and siRNA-mediated knockdown, as indicated.

• ***Primers for* RT-qPCR** *(listed 5' to 3')*

For mRNAs:

PLAG1 Fwd:	AAATGGGAAGGATTGGATTC
PLAG1 Rev:	CATGTGCCTGATTACTGATG
FOSL1 Fwd:	CTTGTGAACAGATCAGCC
FOSL1 Rev:	CCAGATTTCTCATCTTCCAG

For eRNAs (listed 5' to 3'; the enhancers are named by transcription factor motif and genomic coordinates:

PLAG1 chr20 180835-1809350 FW:	CACGGCTGGAGGTGAACTAT
PLAG1 chr20 180835-1809350 RV:	GTGGCCTTTGGAGATGAAAA
PLAG1 chr17 69971325-69972325 FW:	CCTTTTCCTCCTCGGAGACT
PLAG1 chr17 69971325-69972325 RV:	GTATAGGGGAGCCAGGAAGC
FOSL1 chr5 172233725-172234725 FW:	CTCCCAAAGTGCTGGGATTA
FOSL1 chr5 172233725-172234725 RV:	AGTCATCTCGCTTCCTCCAA
FOSL1 chr1 32421075-32422075 FW:	AGGCTTGGAGAGCCATTTTT
FOSL1 chr1 32421075-32422075 RV:	GGGGAAGTTGGATTCCTTTC

• ***Primers for ChIP-qPCR*** *(listed 5' to 3'; the enhancers are named by transcription factor motif and genomic coordinates):*

PLAG1 chr20 180835-1809350 FW:	CACGGCTGGAGGTGAACTAT
PLAG1 chr20 180835-1809350 RV:	GTGGCCTTTGGAGATGAAAA
PLAG1 chr17 69971325-69972325 FW:	CCTTTTCCTCCTCGGAGACT
PLAG1 chr17 69971325-69972325 RV:	GTATAGGGGAGCCAGGAAGC
RUNX2 chr11 14402525-14403525 FW:	TGACCATGAGCAGGTCACAT
RUNX2 chr11 14402525-14403525 RV:	GACTCACGGCTACCTCTTGG
RUNX2 chr18 11946650-11947650 FW:	GCAGTGGCTCATGCTTGTAA
RUNX2 chr18 11946650-11947650 RV:	AGAGTGCAGTGGCTCAATCA
FOXA1 chr6 33950425-33951425 FW:	CGCCTGTAATCCCAACACTT
FOXA1 chr6 33950425-33951425 RV:	TTTGGTAGAGGCAGGGTGTC

FOXA1 chr2 121179700-121180700 FW:	TGGTTCAAAAGCAGATGCAC
FOXA1 chr2 121179700-121180700 RV:	ATACCAGCACCCTGGTCAAG
HLF chr17 58219425-58220425 FW:	GCCTGCCCCTAATCCTTTAC
HLF chr17 58219425-58220425 RV:	GGGAATGGCTTTTTCCTAGC
HLF chr20 52278300-52279300 FW:	CCACTGTGCCCAGCTAATTT
HLF chr20 52278300-52279300 RV:	TCACCTGAGATCGGGAGTTC
FOSL1 chr5 172233725-172234725 FW:	CTCCCAAAGTGCTGGGATTA
FOSL1 chr5 172233725-172234725 RV:	AGTCATCTCGCTTCCTCCAA
FOSL1 chr1 32421075-32422075 FW:	AGGCTTGGAGAGCCATTTTT
FOSL1 chr1 32421075-32422075 RV:	GGGGAAGTTGGATTCCTTTC
FOSL1 chr1 193080700-193081700 FW:	CTGGAGCCAAAGCCTATCTG
FOSL1 chr1 193080700-193081700 RV:	GCAGTGAGCTGTGATTGCAT
FOSL1 chr2 65065275-65066275 FW:	CCTCCAGCACACTTCCTCTC
FOSL1 chr2 65065275-65066275 RV:	CTAAAGCCTGCTCCAGGATG
FOSL1 chr18 11946650- 11947650 FW:	GCAGTGGCTCATGCTTGTAA
FOSL1 chr18 11946650- 11947650 RV:	AGAGTGCAGTGGCTCAATCA
FOSL1 chr2 232544200-232545200 FW:	TGCATGAGGCATGGTTTAGA
FOSL1 chr2 232544200-232545200 RV:	CTGGAGGCTCTGTTCCTCAC
FOSL1 chr17 42620275- 42621275 FW:	GGCTCAGCAGGCTCTCTCTA
FOSL1 chr17 42620275- 42621275 RV:	GGGGTCACTAGGAAGGGAAG
FOSL1 chr1 224075275-224076275 FW:	CCTTTCCGGTTTCATGCTTA
FOSL1 chr1 224075275-224076275 RV:	TGCCCTTAGGAGGAACATTG
FOSL1 chr1 31643400-31644400 FW:	TAGTTCCCCAAGGTCACAGG
FOSL1 chr1 31643400-31644400 RV:	CTCAGCAGTTCCCCTCAGTC
FOSL1 chr19 39564575-39565575 FW:	TAGCTCCAGGGGAGATGCTA
FOSL1 chr19 39564575-39565575 RV:	CCACCCGACTGTAGTTTGCT
Neg Ctrl Reg chr1 204765884-204767712 FW:	GAGTTGTGCCCAAATCCTGT
Neg Ctrl Reg chr1 204765884-204767712 RV:	CAGCCTTCTCTCACCCTCAC
Neg Ctrl Reg 1 [from Hah et. al. (2013)]	FW: CCTGCTTGCTGTCTGAGC
Neg Ctrl Reg 1 ([from Hah et. al. (2013)] RV:	TGTCGCCATCAGGATTTC

• ***siRNAs*** *(listed 5' to 3'; siRNAs were purchased from Dharmacon siGENOME and used as a pool of 4 oligos):*

FOSL1 #1	GCUCAUCGCAAGAGUAGCA [dT][dT]
FOSL1 #2	GGACACAGGCAGUACCAGU[dT][dT]
FOSL1 #3	AGCGAGAGAUUGAGGAGCU[dT][dT]
FOSL1 #4	GCAGGCGGAGACUGACAAA[dT][dT]
PLAG1 #1	GUACCACCCUCCCACGUUU[dT][dT]
PLAG1 #2	GGGAAUGACAUGCCCAAUA[dT][dT]
PLAG1 #3	AGUAAGAGAUACCCAGAAA[dT][dT]
PLAG1 #4	GUCCUUACCUUCCAGUGAA[dT][dT]
PLK1 #1	CAACCAAAGUCGAAUAUGA[dT][dT]
PLK1 #2	CAAGAAGAAUGAAUACAGU[dT][dT]
PLK1 #3	GAAGAUGUCCAUGGAAAUA[dT][dT]
PLK1 #4	CAACACGCCUCAUCCUCUA[dT][dT]

siGenome Non-Targeting siRNA pool #2, #1	UAAGGCUAUGAAGAGAUAC[dT][dT]
siGenome Non-Targeting siRNA pool #2, #2	AUGUAUUGGCCUGUAUUAG[dT][dT]
siGenome Non-Targeting siRNA pool #2, #3	AUGAACGUGAAUUGCUCAA[dT][dT]
siGenome Non-Targeting siRNA pool #2, #4	UGGUUUACAUGUCGACUAA[dT][dT]

TFSEE Files

- tfsee_analysis.py - Script used to compute TFSEE to identify cognate transcription factors.
- requirements.txt - Python package dependencies that need to be installed to run tfsee_analysis.py.
- README.pdf - PDF file describing the dependencies for identification of key breast cancer subtype-specific TFs that act at transcribed enhancers to dictate gene expression patterns determining growth outcomes, using TFSEE
- README.md - Markdown format of information present in the PDF file.

(These files are provided in a separate compressed folder)

SUPPLEMENTAL REFERENCES

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202-208.

Chae M, Danko CG, Kraus WL. 2015. groHMM: A computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics* **16**: 222.

Danko CG, Chae M, Martins A, Kraus WL. 2014. groHMM: GRO-seq Analysis Pipeline. In *Bioconductor*. Bioconductor, http://bioconductor.org/packages/release/bioc/html/groHMM.html.

Danko CG, Hah N, Luo X, Martins AL, Core L, Lis JT, Siepel A, Kraus WL. 2013. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol Cell* **50**: 212-222.

Dinse GE, Lagakos SW. 1982. Nonparametric estimation of lifetime and disease onset distributions from incomplete observations. *Biometrics* **38**: 921-932.

Feng J, Liu T, Qin B, Zhang Y, Liu XS. 2012. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**: 1728-1740.

Franco HL, Nagari A, Kraus WL. 2015. TNFalpha signaling exposes latent estrogen receptor binding sites to alter the breast cancer cell transcriptome. *Mol Cell* **58**: 21-34.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol* **8**: R24.

Hah N, Danko CG, Core L, Waterfall JJ, Siepel A, Lis JT, Kraus WL. 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 CG, Kraus WL. 2013. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res* **23**: 1210-1223.

Kaplan EL, Meier P. 1958. Nonparametric estimation from incomplete observations *Journal of American Statistical Association* **53**: 457-481.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**: 1813-1831.

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

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Luo X, Chae M, Krishnakumar R, Danko CG, Kraus WL. 2014. Dynamic reorganization of the AC16 cardiomyocyte transcriptome in response to TNFalpha signaling revealed by integrated genomic analyses. *BMC Genomics* **15**: 155.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17**.

Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H et al. 2014. JASPAR 2014: an extensively expanded and

updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* **42**: D142-147.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Ringner M, Fredlund E, Hakkinen J, Borg A, Staaf J. 2011. GOBO: gene expression-based outcome for breast cancer online. *PLoS One* **6**: e17911.

Sun M, Gadad SS, Kim DS, Kraus WL. 2015. Discovery, annotation, and functional analysis of long noncoding RNAs controlling cell-cycle gene expression and proliferation in breast cancer cells. *Mol Cell* **59**: 698-711.

Szasz AM, Lanczky A, Nagy A, Forster S, Hark K, Green JE, Boussioutas A, Busuttil R, Szabo A, Gyorffy B. 2016. Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. *Oncotarget* **7**: 49322-49333.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105-1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511-515.

Xi Y, Li W, Tanaka K, Allton KL, Richardson D, Li J, Franco HL, Nagari A, Malladi V, Coletta LD et al. 2017. Epigenetic landscapes define breast cancer subtypes. *(submitted)*.

Zhong S, Joung JG, Zheng Y, Chen YR, Liu B, Shao Y, Xiang JZ, Fei Z, Giovannoni JJ. 2011. High-throughput illumina strand-specific RNA sequencing library preparation. *Cold Spring Harb Protoc* **2011**: 940-949.