Enhancer transcription reveals subtype-specific gene expression programs controlling breast cancer pathogenesis

Hector L. Franco, Anusha Nagari, Venkat Malladi, et al.

Genome Res. published online December 22, 2017
Access the most recent version at doi:10.1101/gr.226019.117

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

Running Title: Subtype-Specific Enhancers in Breast Cancers

Hector L. Franco[1, 6], Anusha Nagari[1], Venkat Malladi[1], Wenqian Li[2], Yuanxin Xi[3], Dana Richardson[4], Kendra L. Allton[5], Kaori Tanaka[5], Jing Li[5], Shino Murakami[1], Khandan Keyomarsi[4], Mark T. Bedford[2], Xiaobing Shi[5], Wei Li[3], Michelle C. Barton[5], Sharon Y. R. Dent[2], and W. Lee Kraus[1, 7]

[1] Laboratory of Signaling and Gene Regulation, Cecil H. and Ida Green Center for Reproductive Biology Sciences and Division of Basic Reproductive Biology Research, Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center, Dallas, TX, 75390.

[2] Department of Epigenetics and Molecular Carcinogenesis and The Center for Cancer Epigenetics, University of Texas M.D. Anderson Cancer Center, Smithville, Texas 78957, USA.

[3] Department of Molecular and Cellular Biology and Division of Biostatistics, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas 77030, USA.

[4] The Department of Experimental Radiation Oncology, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA.

[5] The Department of Epigenetics and Molecular Carcinogenesis, University of Texas Graduate School of Biomedical Sciences at Houston and The Center for Cancer Epigenetics, University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, USA.

[6] Current address: Department of Genetics and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

[7] Address correspondence to:
W. Lee Kraus, Ph.D.
Cecil H. and Ida Green Center for Reproductive Biology Sciences
The University of Texas Southwestern Medical Center at Dallas
5323 Harry Hines Boulevard
Dallas, TX 75390-8511

Phone: 214-648-2388
Fax: 214-648-0383
E-mail: LEE.KRAUS@utsouthwestern.edu

ABSTRACT

Non-coding transcription is a defining feature of active enhancers, linking transcription factor (TF) binding to the molecular mechanisms controlling gene expression. To determine the relationship between enhancer activity and biological outcomes in breast cancers, we profiled the transcriptomes (using GRO-seq and RNA-seq) and epigenomes (using ChIP-seq) of 11 different human breast cancer cell lines representing five major molecular subtypes of breast cancer, as well as two immortalized ('normal') human breast cell lines. In addition, we developed a robust and unbiased computational pipeline that simultaneously identifies putative subtype-specific enhancers and their cognate TFs by integrating the magnitude of enhancer transcription, TF mRNA expression levels, TF motif p-values, and enrichment of H3K4me1 and H3KK27. When applied across the 13 different cell lines noted above, the Total Functional Score of Enhancer Elements (TFSEE) identified key breast cancer subtype-specific TFs that act at transcribed enhancers to dictate gene expression patterns determining growth outcomes, including Forkhead TFs, FOSL1, and PLAG1. FOSL1, a Fos family TF, (1) is highly enriched at the enhancers of triple negative breast cancer (TNBC) cells, (2) acts as a key regulator of the proliferation and viability of TNBC cells, but not Luminal A cells, and (3) is associated with a poor prognosis in TNBC breast cancer patients. Taken together, our results validate our enhancer identification pipeline and reveal that enhancers transcribed in breast cancer cells direct critical gene regulatory networks that promote pathogenesis.

Keywords:

Breast cancer, Triple negative breast cancer, Gene regulation, GRO-seq, Epigenomics, Transcription, eRNA, Enhancer, Enhancer transcription, FOSL1, PLAG1.

INTRODUCTION

Gene expression profiling has shown that breast cancer is not a single disease with variable morphologic features and biomarkers, but rather a group of molecularly distinct neoplastic disorders (Perou et al. 2000; Sotiriou and Pusztai 2009; Ciriello et al. 2015). These analyses have identified at least five distinct molecular subtypes of breast cancer, which correlate with hormone responsiveness, patient prognosis, and response to therapy: Luminal A, Luminal B, HER2+, Triple negative (TN)/Basal-like, Triple negative/Claudin-low], which express unique sets of genes that persist from pre-neoplastic lesions through metastatic disease (Perou et al. 2000; Sotiriou and Pusztai 2009). Nonetheless, little is known about the initial, disease-driving transcriptional and epigenetic changes that promote the phenotypic outcomes.

Deep sequencing technologies used to interrogate the genome have pointed to transcriptional enhancers, as well as the transcription factors (TFs) that promote their formation, as key regulatory elements controlling the cell type-specific biology of essentially all biological systems examined to date (Shlyueva et al. 2014; Heinz et al. 2015). In cancers, cell type-specific enhancers can become deregulated or 'hijacked,' allowing the activation of genes that promote tumor formation and metastasis (Chipumuro et al. 2014; Northcott et al. 2014; Franco and Kraus 2015). Key questions regarding enhancer biology in cancers and other biological systems include: (1) the location of enhancers throughout the genome, (2) the functionality or activity of the enhancers, (3) the TFs that nucleate enhancer formation, and (4) the target genes of the enhancers (Shlyueva et al. 2014; Heinz et al. 2015).

Although different enhancer-nucleating TFs may be expressed in different cell types, the enhancers that they form share several common features. For example, enhancers are typically found in open regions of chromatin (as assessed by DNase-seq) (Crawford et al. 2006; Sheffield

et al. 2013) and are enriched with a common set of histone modifications (as assessed by ChIP-seq), including histone H3 lysine 4 monomethylation (H3K4me1) and histone H3 lysine 27 acetylation (H3K27ac) (Heintzman et al. 2007; Heintzman et al. 2009). Recently, genomic assays have revealed that many enhancers are bound by RNA polymerase II (Pol II) and are actively transcribed, producing enhancer RNAs ('eRNAs') (De Santa et al. 2010; Kim et al. 2010; Hah et al. 2011), which serve as a robust mark of active enhancers that can be used to track enhancer activity (Wang et al. 2011; Hah et al. 2013; Li et al. 2013; Franco et al. 2015; Kim and Shiekhattar 2015). We and others have shown that global run-on sequencing (GRO-seq) and its derivatives (e.g., PRO-seq) are robust genomic methods to detect and measure ongoing transcription at enhancers, which can be used for enhancer prediction (Hah et al. 2011; Wang et al. 2011; Hah et al. 2013; Core et al. 2014; Chae et al. 2015; Heinz et al. 2015).

In the studies presented herein, we describe the development of a computational pipeline, called the Total Functional Score of Enhancer Elements (TFSEE), which we use to mine we a large number of transcriptomic and epigenomic data sets from human breast cancer cell lines, with the goal of identifying subtype-specific TFs that drive the subtype-specific biology of breast cancers.

RESULTS

Generating Transcriptional and Epigenomic Maps from Breast Cancer Cells

To better understand the TF-driven transcriptional programs that control subtype-specific gene expression programs in breast cancers, we generated transcriptional and epigenomic maps for 13 different breast cell lines, 11 of which represent five intrinsic molecular subtypes of breast cancer (Figure 1A), under basal growth conditions (i.e., without treatment). For these analyses, we used GRO-seq and RNA-seq, as well as ChIP-seq for 11 different histone modifications (Figure 1, A and B), the latter which is described in detail elsewhere (Xi et al. 2017). The histone modifications examined in this study are enriched at various functional elements and in a variety of chromatin environments, such as enhancers, promoters, gene bodies, and repressive regions of the genome, thus providing a broad survey of the epigenomic features associated with transcriptional outcomes (Figure 1, A and B; Supplemental Figure S1, A and B). In total, we generated data sets from 26 GRO-seq, 26 RNA-seq, and 286 ChIP-seq libraries (see Supplemental Materials for details about replicates and read depth).

Although previous studies have applied RNA-seq and ChIP-seq to transcriptomic and epigenomic analyses, respectively, in breast cancer cells, the inclusion of GRO-seq provides additional levels of unique information. GRO-seq is a direct measure of transcription, which yields the position and orientation of all actively transcribing RNA polymerases across the genome (Core et al. 2008), thus facilitating the comprehensive identification of transcribed regions and functional elements (Hah et al. 2011; Luo et al. 2014; Chae et al. 2015; Sun et al. 2015). Collectively, these data sets provide unique resource for cancer researchers and provide a foundation for discovering the transcriptional and epigenomic underpinnings of breast cancer.

Using Transcription to Predict Enhancers in the Absence of Other Genomic Information

We and others have shown that actively transcribed enhancers are more likely to (1) associate with enhancer-related chromatin modifications, such as H3K4me1 and H3K27ac, (2) loop to target gene promoters, and (3) correlate with target gene activation (Kim et al. 2010; Wang et al. 2011; Orom and Shiekhattar 2013; Franco et al. 2015; Hah et al. 2015; Heinz et al. 2015). In addition, the so-called 'super enhancers' that are typically associated with oncogene expression are associated with enhancer transcription (Hah et al. 2015). Thus, enhancer transcription is a good predictor of active enhancers and can be used in the absence of other genomic information to predict active enhancers *de novo* (Hah et al. 2013; Chae et al. 2015).

We reasoned that using signatures of enhancer transcription from GRO-seq data would allow us to identify active enhancers in cell lines representing the different molecular subtypes of breast cancer that control the expression of cancer-relevant genes. We used a computational tool that we developed previously for calling unannotated transcription units from GRO-seq data, called groHMM (Hah et al. 2011; Hah et al. 2013; Danko et al. 2014; Chae et al. 2015), to identify a set of eRNA transcripts produced in each of the 13 cell lines (Figure 2, A and B). This analysis revealed both unique and shared eRNAs across the cell lines (Figure 2B), as well as an overlap of enhancer transcription with histone modifications that are typically enriched at enhancers (e.g., H3K27ac and H3K4me1) (Figure 2C).

Using the ZR-75-1 Luminal A cell line as an example (a cell line for which DNase I hypersensitivity data sets are publicly available), we compared the overlap of output from traditional enhancer prediction methods (e.g., DNase I hypersensitivity, or H3K4me1 and H3K27ac enrichment; (Shlyueva et al. 2014)) to output from an enhancer transcription-based approach (Figure 2, D and E). Using the pipeline shown in Figure 2A, we found that 65% of enhancers called based on enhancer transcription using GRO-seq data are also identified by all

three of the other methods (i.e., DNase I hypersensitivity, enrichment of H3K4me1 or H3K27ac). In contrast, only 1-2% of enhancers called based on DNase I hypersensitivity, or enrichment of H3K4me1 or H3K27ac are identified by all three of the other methods (Figure 2D). This may be due, in part, to the fact that enhancer calling based on DNase I hypersensitivity, or H3K4me1 or H3K27ac enrichment, yields much larger numbers of putative enhancers (Figure 2E), many of which may be false positives or inactive as true regulatory elements. Nonetheless, as we show below, incorporating enhancer transcription into an enhancer-calling pipeline that includes information about DNase I hypersensitivity, as well as H3K4me1 and H3K27ac enrichment, improves the fidelity of the enhancer calls.

Actively Transcribed Enhancers Track with Subtype Specific Transcriptional Programs

Next, we determined if the transcribed enhancers identified above might regulate nearby genes relevant to the biology of different subtypes of breast cancer. To do so, we identified all of the uniquely transcribed enhancers called in each cell line using our pipeline, and then determined the level of transcription for each enhancer (Figure 3A; Supplemental Figure S2A) and the nearest neighboring gene (upstream or downstream) (Figure 3B; or within 300 kb, Supplemental Figure S2B). As expected, the uniquely transcribed enhancers exhibited high levels of transcription in the cell type in which they were active, but the same loci were minimally transcribed in all other cell lines (Figure 3A; Supplemental Figure S2A), which is an important control for this analysis. Likewise, the nearest neighboring gene for each uniquely transcribed enhancer exhibited high levels of transcription in the cell type in which the enhancers were active, but the same genes were minimally transcribed in all other cell lines (Figure 3B; Supplemental Figure S2B). These results suggest that the enhancers called using GRO-seq data in our pipeline are associated with cell type-specific patterns of gene expression.

To further assess the broader biological significance, we determined the expression of these same genes in patient tumor samples. The genes nearest to the uniquely transcribed enhancers in the ZR-75-1 luminal A cell line were preferentially expressed in the corresponding patient tumor samples of the same molecular subtype (Figure 3C). Likewise, the genes nearest to the uniquely transcribed enhancers in the MDA-MB-468 TN basal cells were preferentially expressed in the corresponding patient tumor samples of the same molecular subtype (Figure 3C). These results illustrate that enhancers called using our pipeline with GRO-seq data from luminal A and TN breast cancer cell lines are associated with biologically relevant gene expression patterns in breast cancer patient samples.

A Functional Score of Enhancer Elements Identifies Subtype-Specific Enhancers and Their Cognate TFs

The analyses described above are a useful way to identify active enhancers and their putative target genes, but they tell us little about the TFs that drive enhancer formation or the relative activity of each enhancer. Thus, we sought to develop a method that would allow us to extract more detailed functional information from the genomic data. In this regard, we developed a computational pipeline that we call the Total Functional Score of Enhancer Elements (TFSEE), which integrates data from GRO-seq, RNA-seq, histone modification ChIP-seq (i.e., H3K27ac and H3K4me1), and motif searches, allowing for the simultaneous identification of active enhancers and their cognate transcription factors across all breast cancer cell lines (Figure 4A). TFSEE integrates (1) the location and magnitude of enhancer activity based on enhancer transcription (GRO-seq), (2) the enrichment of enhancer-related histone modifications (ChIP-seq), (3) the level of TF expression (RNA-seq), (4) the level of target gene expression (RNA-seq), and (5) the TF motif scores for each enhancer (Figure 4A).

For the *de novo* motif analyses, we searched a 1 kb region surrounding the transcribed enhancers for each cell line using MEME software (Bailey et al. 2009) and matched the motifs to known transcription factors from the Tomtom and JASPAR data bases (Gupta et al. 2007; Mathelier et al. 2016) (Supplemental Figure S3). We recorded the p-values for the motifs identified and assigned a higher score to more significant p-values. To ensure that the TFs whose motifs were called in our analysis were actually expressed in the corresponding cell lines, we used RNA-seq data to determine the expression levels of the mRNAs encoding them. Thus, the highest scoring TFs from TFSEE were those that had the (1) highest levels of enhancer transcription, (2) greatest enrichment of H3K4me1 and H3K27ac, (3) most significant p-values for the motifs, and (4) best correlation with the expression of the predicted TF in a given cell line.

We visualized the results from TFSEE using unsupervised hierarchical clustering (Figure 4B), which grouped the cell lines into two major clades: (1) TN / normal and (2) luminal A+B / HER2+ (Figure 4B). This analysis provided a clear demarcation of the TFs that were most enriched at transcribed enhancers between the two clades. Interestingly, previous gene expression profiling studies have also demonstrated an association or similarities between TN and 'normal' cells, both of which do not express ER, PR, or HER2 and are technically triple negative (Charafe-Jauffret et al. 2006; Neve et al. 2006; Marcotte et al. 2016). We then determined a rank order frequency distribution for all TFs within each clade (Figure 4, C and D; Supplemental Table S1). One striking observation from this analysis was the abundance of Forkhead box family TFs that ranked high in both clades (Figure 4, C and D). Importantly, the clade-specific Forkhead TFs were selectively expressed in patient samples of the same molecular subtype (Figure 4E). For example, FOXF2, FOXQ1 and FOXC1 were highly ranked in the TN /

normal clade and are also more highly expressed in TN basal breast tumors compared to other molecular subtypes (Figure 4E). The well-studied Forkhead TF FOXA1, which is expressed in luminal breast cancers (Bernardo et al. 2010; Hurtado et al. 2011), was highly enriched in the luminal A+B / HER2+ clade, as expected, lending support to the fidelity of our approach. Similar results were obtained using pairwise comparisons of TFSEE scores from the different molecular subtypes, rather than unsupervised hierarchical clustering (Supplemental Figures S4, S5, and S6). Collectively, the pairwise analyses identified most of the top TFs identified in the original unsupervised hierarchical clustering analysis (Supplemental Figure S6, A-D; compare to Figure 4, C and D).

To confirm experimentally that the predicted TF did indeed bind to the transcribed enhancers, we performed ChIP-qPCR experiments for two transcription factors in each clade. These analyses revealed the enrichment of PLAG1 and RUNX2 (i.e., TN / normal clade) at transcribed enhancers in HCC-1937 TN cells (Figure 5A; Supplemental Figure S7, A and B), as well as FOXA1 and HLF (i.e., luminal A+B / HER2+ clade) at transcribed enhancers in MCF-7 luminal A cells (Figure 5B; Supplemental Figure S7, A and B). These TFs did not bind to a negative control region that exhibited no features indicative of an active enhancer (Figure 5, A and B; Supplemental Figure S7, A-D). Of the 24 potential positive or negative binding events tested for these four TFs (12 sites total tested in two cell lines), 22 (~90%) gave the expected result. Taken together, our TFSEE analysis lead to the identification of key enhancers and their cognate transcription factors that may control the gene expression programs that dictate the cellular phenotypes of the different molecular subtypes of breast cancers.

FOS-like 1 (FOSL1) is Enriched at Transcribed Enhancers in TN cells, Regulates Cellular Proliferation, and is Predictive of Breast Cancer Patient Outcomes

To confirm that subtype-specific TFs identified using TFSEE play a role in the biology of the cognate cell types, we performed a series of functional analyses on FOS-like 1 (FOSL1), a TF highly enriched in the TN / normal clade (Figure 4C). FOSL1 is a member of the Fos gene family of leucine zipper proteins that dimerize with Jun proteins to form the AP-1 TF complex (Shaulian and Karin 2002; Shaulian 2010; Zhao et al. 2014). AP-1 is a key component of many signal transduction pathways, which regulate a variety of cellular processes including differentiation, apoptosis, migration, and transformation (Shaulian and Karin 2002; Shaulian 2010; Zhao et al. 2014).

To confirm that FOSL1 binds at predicted enhancers in TN cells (e.g., Figure 6A; Supplemental Figure S8A), we performed ChIP-qPCR at predicted enhancers in several TN cell lines. We observed a significant enrichment of FOSL1 at its cognate predicted enhancers, but not at a negative control region that exhibited no features indicative of an active enhancer (Figure 6B; Supplemental Figure S8, B and C; Supplemental Figure S9). Of the 33 potential positive or negative FOSL1 binding events tested (11 sites total in three cell lines), 28 (~85%) gave the expected result. Furthermore, siRNA-mediated knockdown of FOSL1 caused a significant reduction in enhancer transcription at the cognate predicted enhancers as measured by RT-qPCR (Figure 6C). These results indicate that FOSL1 binds to the genomic loci and nucleates enhancer formation.

Extending these analyses to patient samples, we observed elevated expression of FOSL1 mRNA in TN basal tumors relative to other breast cancer subtypes, which increases with tumor grade (Figure 6D). Furthermore, elevated expression of FOSL1 is associated with worse clinical outcomes (i.e., overall survival) in patients with ER-negative breast cancers, but not ER-positive breast cancers (Figure 6E), suggesting a role for these FOSL1 as an oncogene. Finally, to

directly assess a subtype-specific functional role of FOSL1 in the biology of breast cancers, we knocked down FOSL1 in breast cancer cells and monitored the proliferation and viability of the cells, using knockdown of Polo-like Kinase 1 (PLK1) as a positive control. Knockdown of FOSL1 inhibited the proliferation and decreased the viability of several TN cell lines (Figure 6F; Supplemental Figure S8, D and E), but had no effect on MCF-7 luminal A cells (Figure 6G).

A parallel set of experiments with PLAG1, another TF highly enriched in the TN / normal clade (Figure 4C), confirmed that PLAG1 also plays a key role in enhancer formation and the biology of TN cells (Supplemental Figure S10). In addition, we found that elevated expression of three other TFs in the TN / normal clade (i.e., PRDM1, IRF1, and RUNX2) is associated with better clinical outcomes (i.e., distant metastasis-free survival) in patients with ER-negative breast cancers, but not ER-positive breast cancers (Supplemental Figure S11), suggesting roles for these TFs as tumor suppressors. Taken together, our results show that TFSEE can be used to identify breast cancer subtype-specific TFs that control the biology of those subtypes. In addition, our analyses have led to the discovery of TN-specific TFs, such as FOSL1, that control the proliferation and viability of TN cells and whose expression is predictive of clinical outcomes in patients.

DISCUSSION

The results from our analyses demonstrate that enhancer transcription can be used to improve the fidelity of functional enhancer identification. In addition, our study has identified TFs that play critical roles in the biology of specific types of breast cancers, including Forkhead TFs, FOSL1, and PLAG1. More broadly, our study has described a genomic and computational approach for identifying enhancers and their cognate TFs that play a critical role in the biology of particular cell types, which should be applicable to a wide variety of biological systems.

Enhancer Transcription Defines Functional Regulatory Elements and Their Cognate TFs

We and others have shown previously that enhancer transcription, as defined by GRO-seq and related methods (e.g., PRO-seq), is a robust way to identify regulatory elements, such as enhancers, on a global scale (Hah et al. 2011; Wang et al. 2011; Hah et al. 2013; Core et al. 2014; Chae et al. 2015; Heinz et al. 2015). In addition, we have shown that the presence of enhancer transcription can be used to distinguish between active enhancers and inactive TF binding sites (Hah et al. 2013; Franco et al. 2015). Other genomic methods that have been used to identify enhancers include (1) DNase-seq, which measures accessibility ('hypersensitivity') of the genome (although indiscriminately between open promoters, gene bodies, and enhancers), (2) ChIP-seq for histone modifications typically enriched at enhancers, such as H3K4me1 and H3K27ac, and (3) ChIP-seq for common enhancer-enriched coregulators, such as the acetyltransferases EP300 (a.k.a. p300) and CBP (Shlyueva et al. 2014). A directed approach using ChIP-seq for a TF of interest can also be used, but this approach is biased and requires prior knowledge about the TFs that are functioning in a given cell type. These approaches yield thousands of putative enhancers (see for example Figure 2E); the challenge is determining which of the putative enhancers are functional in the cell type of interest.

TFSEE using GRO-seq data allows for the simultaneous identification of enhancers, assessment of enhancer activity and gene expression, and identification of TFs that drive cell type-specific gene expression programs (Figure 4A). TFSEE improves enhancer identification by focusing the analysis on those enhancers that are most likely to be functionally active in a given cell type. Although TFSEE calls fewer enhancers from GRO-seq data than is typically called using other types of enhancer-related genomic data (Figure 2E), the enhancers called from GRO-seq data exhibit the greatest overlap with enhancers called using other approaches (Figure 2D). Transcribed enhancers called by three or four genomic assays are more likely to be functional in the cell type in which they are called than enhancers called by a single genomic assay. Moreover, the functional indications from enhancer transcription, which are not evident with other enhancer features (e.g., DNase I hypersensitivity, histone modification enrichment), increase the likelihood of identifying functional enhancers, as opposed to inactive TF binding sites.

Forkhead Family TFs in Luminal and Triple Negative Breast Cancers

Our analyses identified TFs that play subtype-specific roles in breast cancers, including Forkhead TFs. The Forkhead (FOX) proteins comprise a diverse family of TFs with roles in a variety of biological systems, including cancers (Benayoun et al. 2011; Lam et al. 2013). We found FOXF2, FOXQ1 and FOXC1 to be enriched in TN and ER-negative breast tumors compared to other molecular subtypes, whereas FOXI1, FOXA1 and FOXP2 were enriched in luminal and ER-negative breast tumors compared to other molecular subtypes (Figure 4E). The FOX-related results from our TFSEE analyses are well supported by the literature. For example, previous studies have shown that FOXC1 regulates basal-like breast cancer cells by activating NF-κB signaling (Wang et al. 2012) and predicts poor overall survival in this breast cancer

subtype (Ray et al. 2010). Likewise, FOXA1 is a well characterized TF critical for the estrogen-dependent growth and proliferation of luminal breast tumors (Bernardo et al. 2010; Hurtado et al. 2011). Our results suggest that additional exploration of the functions of the Forkhead family of TFs in breast cancers is warranted.

FOSL1 and PLAG1: TFs Driving the Triple Negative Breast Cancer Phenotype

Our studies also identified and verified FOSL1 (a.k.a. Fra1) as a key TF in the biology of TN cells. In this regard, we found that FOSL1 enhancers are enriched in TN cells (Figure 4C), and knockdown of FOSL1 inhibits the proliferation and viability of TN cells (Figure 6F). Furthermore, FOSL1 expression is elevated TN basal tumors (Figure 6D), a feature that is associated with poor clinical outcomes in patients (Figure 6E). Thus, FOSL1 exhibits hallmarks of a cancer 'driver," which may be particularly active in TN cells. FOSL1 dimerizes with Jun proteins to form AP-1 TF complexes (Shaulian and Karin 2002; Shaulian 2010; Zhao et al. 2014), which function as key regulators of gene expression in cancers (Eferl and Wagner 2003; Verde et al. 2007). FOSL1 has previously been implicated in a variety of cancer types, including those of the colon, lung, and breast (Young and Colburn 2006). FOSL1 acts to induce epithelial-to-mesenchymal transitions and drive metastasis in breast and other cancers (Diesch et al. 2014; Risolino et al. 2014; Bakiri et al. 2015; Dhillon and Tulchinsky 2015; Iskit et al. 2015; Liu et al. 2015).

Finally, our studies identified and verified PLAG1 as a key TF in the biology of TN cells (Supplemental Figure S10). PLAG1 (Pleomorphic adenoma gene 1 protein) is a zinc-finger transcription factor, which has been implicated in cancer (Abdollahi 2007; Van Dyck et al. 2007). Unlike FOSL1, little if any previous direct evidence exists linking PLAG1 to breast cancers. Rather, the gene encoding PLAG1 is consistently rearranged and subject to activating

reciprocal chromosomal translocations involving 8q12 in pleomorphic adenomas of the salivary glands (Kas et al. 1997; Abdollahi 2007; Van Dyck et al. 2007). Thus, TFSEE can be a useful approach for identifying new TF drivers in cancers.

Collectively, our results provide a clear example of how integrative computational analyses of genomic data can be used to understand the molecular mechanisms supporting the function and biology of specific cell types.

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. Cell proliferation was assessed using a crystal violet staining assay and cell viability was assessed using Cell Titer Glo reagent (Promega) for cells transfected with siRNAs using reverse transfection methodology. Additional details about the cell culture conditions, cell proliferation and viability assays, and siRNA-mediated knockdown are provided in the Supplemental Methods.

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.

Global run-on sequencing (GRO-seq)

Cells were collected at ~70-80% confluence and nuclei were isolated as described previously (Luo et al. 2014). 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). 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. Additional details about the preparation of nuclei, nuclear run-ons, GRO-seq library preparation, and sequencing are provided in the Supplemental Methods.

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. 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). Reads >32 bp long were retained for alignment to the human reference genome using the BWA aligner v 0.6.1 (Li and Durbin 2010). 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. Enhancer transcripts were classified into (1) short paired eRNAs and (2) short unpaired eRNAs as described previously (Hah et al. 2013). The universe of expressed eRNAs (short paired and short unpaired) was assembled and used for further analyses.

The universe of expressed genes in each cell line was determined from 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. De novo motif analyses were performed on a 1 kb region ($\pm$ 500 bp) surrounding the peak summit or the transcription start site for short paired and short unpaired eRNAs, respectively, using MEME (Bailey et al. 2009). The predicted motifs were matched to known motifs using Tomtom (Gupta et al. 2007).

Additional details about quality control and trimming, read alignment and gene annotation, transcript calling using groHMM, enhancer transcript calling, nearest neighboring gene analyses, motif analyses, and the generation of box plots are provided in the Supplemental

Methods.

RNA isolation, RT-qPCR, and RNA-seq

Cells were collected at ~70-80% confluence and total RNA 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). mRNA-seq libraries were prepared using methods described previously (Zhong et al. 2011). 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 to achieve a minimum of ~65 M raw reads per cell line. Additional details about RNA isolation, RT-qPCR and primers, RNA-seq library preparation, and sequencing are provided in the Supplemental Methods.

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, FPKM values were calculated per gene using Cufflinks (v.2.1.2) (Trapnell et al. 2010). Additional details about quality control and trimming, read alignment, and expression analyses are provided in the Supplemental Methods.

ChIP-qPCR and ChIP-seq

ChIP was performed as previously described (Franco et al. 2015) with a few modifications. Cells were grown to ~70-80% confluence and then cross-linked with 1% formaldehyde for 10 min at 37^{o}C. ChIPed DNA was analyzed by qPCR using enhancer- or

gene-specific primers, or used for ChIP-seq. ChIP libraries were prepared using a modified Kapa LTP Library Preparation kit (KAPA Biosystems) for Illumina Platforms (Xi et al. 2017). The quality of the final libraries was assessed using a 2200 TapeStation (Agilent Technologies). The libraries 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. Additional details about the antibodies used for ChIP, the ChIP method, qPCR and primers, ChIP-seq library preparation, and sequencing are provided in the Supplemental Methods.

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/). Relaxed peaks were called using MACS (v2.1.0) (Feng et al. 2012) with a p-value = 1×10^{-2}. Additional details about quality control and trimming, read alignment, and peak calling are provided in the Supplemental Methods.

Predicting breast cancer subtype-specific TFs using TFSEE

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 scripts are included in a supplemental file). The algorithm does the following (details provided in the Supplemental Material): (1) normalizes enhancer expression using GRO-seq, (2) normalizes enhancer expression using ChIP-seq, (3)

determines enhancer activity, (4) normalizes motif predictions, (5) normalizes TF expression using RNA-seq, and (6) determines the Total Functional Score of Enhancer Elements (TFSEE) and generates a heatmap.

Additional details about (1) normalization of enhancer activity, motif predictions, and transcription factor expression, (2) defining total enhancer activity, (3) determining the total functional score of enhancer elements (TFSEE), (4) pairwise Pearson's correlation analyses, and (5) the generation of heatmaps are provided in the Supplemental Methods.

Oligonucleotide Sequences

The sequences of all oligonucleotides used for RT-qPCR, ChIP-qPCR, and siRNA-mediated knockdown are provided in the Supplemental Methods.

DATA ACCESS

Deep sequencing data from this study are available from the NCBI's Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/) under the following accession numbers: GSE96859 (GRO-seq), GSE96860 (RNA-seq), and GSE85158 (ChIP-seq). The custom scripts for TFSEE are provided in the Supplemental Material.

SUPPLEMENTAL MATERIALS

• The LONESTAR Consortium.

• Supplemental Figures S1 through S11.

• Supplemental Table S1.

• Supplemental Methods.

• Supplemental References.

• TFSEE Files (Scripts and Readme file; provided in a separate compressed folder)

ACKNOWLEDGMENTS

The authors thank Xiaole Shirley Liu, Jean-Pierre Issa, Brad Cairns, Jeff Rosen, and members of the Kraus lab, for helpful comments and discussions. This work was supported by a grant from the Cancer Prevention and Research Institute of Texas (CPRIT) (RP110471-P1) to S.Y.R.D., W.L.K., W.L.[2], X.S., M.T.B., M.C.B., and K.K., a grant from the NIH/NCI (R00CA204628) to H.L.F., and a grant from NIH/NIDDK (DK058110) to W.L.K.

AUTHOR CONTRIBUTIONS

PIs of the LONESTAR Consortium Project 1 (S.Y.R.D., W.L.K., W.L.[†], X.S., M.T.B., M.C.B., K.K.) conceived of the overall project and directed its execution in their labs. H.L.F., A.N., V.M., and W.L.K. developed the specific project described herein with input from

S.Y.R.D., W.L.[†], X.S., M.T.B., M.C.B., and K.K., and performed the experiments and data analyses in the Kraus lab. H.L.F. generated the GRO-seq and RNA-seq data sets, and performed the cellular and molecular assays. V.M., A.N. and H.L.F. developed TFSEE with input from W.L.K., processed and analyzed the GRO-seq and RNA-seq data, and performed the integrative computational analyses of the genomic data. D.R. and K.K. grew the cells and ensured quality control. W.L.[‡], K.L.A, J.L., K.T., S.M., M.T.B., X.S., M.C.B., and S.Y.R.D. generated the ChIP-seq data sets. Y.X. and W.L.[‡] processed the ChIP-seq data and performed the initial analyses. All authors contributed to the data interpretation and presentation. H.L.F., V.M., and A.N. prepared the figures and wrote the methods, and H.L.F. wrote the remaining text. The figures and text were edited and finalized by W.L.K. with input from H.L.F., V.M., A.N. and the rest of the group.

W.L.[†] = Wei Li, Baylor College of Medicine, Houston, Texas

W.L.[‡] = Wenqian Li, University of Texas M.D. Anderson Cancer Center, Smithville, Texas

THE LONESTAR CONSORTIUM

A description of the LONESTAR Consortium is provided in the Supplemental Materials

DISCLOSURES

None.

REFERENCES

Abdollahi A. 2007. LOT1 (ZAC1/PLAGL1) and its family members: mechanisms and functions. *J Cell Physiol* **210**: 16-25.

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.

Bakiri L, Macho-Maschler S, Custic I, Niemiec J, Guio-Carrion A, Hasenfuss SC, Eger A, Muller M, Beug H, Wagner EF. 2015. Fra-1/AP-1 induces EMT in mammary epithelial cells by modulating Zeb1/2 and TGFbeta expression. *Cell Death Differ* **22**: 336-350.

Benayoun BA, Caburet S, Veitia RA. 2011. Forkhead transcription factors: key players in health and disease. *Trends Genet* **27**: 224-232.

Bernardo GM, Lozada KL, Miedler JD, Harburg G, Hewitt SC, Mosley JD, Godwin AK, Korach KS, Visvader JE, Kaestner KH et al. 2010. FOXA1 is an essential determinant of ERalpha expression and mammary ductal morphogenesis. *Development* **137**: 2045-2054.

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.

Charafe-Jauffret E, Ginestier C, Monville F, Finetti P, Adelaide J, Cervera N, Fekairi S, Xerri L, Jacquemier J, Birnbaum D et al. 2006. Gene expression profiling of breast cell lines identifies potential new basal markers. *Oncogene* **25**: 2273-2284.

Chipumuro E, Marco E, Christensen CL, Kwiatkowski N, Zhang T, Hatheway CM, Abraham BJ, Sharma B, Yeung C, Altabef A et al. 2014. CDK7 inhibition suppresses super-enhancer-linked oncogenic transcription in MYCN-driven cancer. *Cell* **159**: 1126-1139.

Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, Zhang H, McLellan M, Yau C, Kandoth C et al. 2015. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163**: 506-519.

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**: 1311-1320.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**: 1845-1848.

Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D et al. 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* **16**: 123-131.

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.

De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei CL, Natoli G. 2010. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol* **8**: e1000384.

Dhillon AS, Tulchinsky E. 2015. FRA-1 as a driver of tumour heterogeneity: a nexus between oncogenes and embryonic signalling pathways in cancer. *Oncogene* **34**: 4421-4428.

Diesch J, Sanij E, Gilan O, Love C, Tran H, Fleming NI, Ellul J, Amalia M, Haviv I, Pearson RB et al. 2014. Widespread FRA1-dependent control of mesenchymal transdifferentiation programs in colorectal cancer cells. *PLoS One* **9**: e88950.

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

Eferl R, Wagner EF. 2003. AP-1: a double-edged sword in tumorigenesis. *Nat Rev Cancer* **3**: 859-868.

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

Franco HL, Kraus WL. 2015. No Driver behind the Wheel? Targeting Transcription in Cancer. *Cell* **163**: 28-30.

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, Benner C, Chong LW, Yu RT, Downes M, Evans RM. 2015. Inflammation-sensitive super enhancers form domains of coordinately regulated enhancer RNAs. *Proc Natl Acad Sci U S A* **112**: E297-302.

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.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**: 108-112.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**: 311-318.

Heinz S, Romanoski CE, Benner C, Glass CK. 2015. The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol* **16**: 144-154.

Hurtado A, Holmes KA, Ross-Innes CS, Schmidt D, Carroll JS. 2011. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat Genet* **43**: 27-33.

Iskit S, Schlicker A, Wessels L, Peeper DS. 2015. Fra-1 is a key driver of colon cancer metastasis and a Fra-1 classifier predicts disease-free survival. *Oncotarget* **6**: 43146-43161.

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

Kas K, Voz ML, Roijer E, Astrom AK, Meyen E, Stenman G, Van de Ven WJ. 1997. Promoter swapping between the genes for a novel zinc finger protein and beta-catenin in pleiomorphic adenomas with t(3;8)(p21;q12) translocations. *Nat Genet* **15**: 170-174.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**: 182-187.

Kim TK, Shiekhattar R. 2015. Architectural and functional commonalities between enhancers and promoters. *Cell* **162**: 948-959.

Lam EW, Brosens JJ, Gomes AR, Koo CY. 2013. Forkhead box proteins: tuning forks for transcriptional harmony. *Nat Rev Cancer* **13**: 482-495.

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.

Li W, Notani D, Ma Q, Tanasa B, Nunez E, Chen AY, Merkurjev D, Zhang J, Ohgi K, Song X et al. 2013. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**: 516-520.

Liu H, Ren G, Wang T, Chen Y, Gong C, Bai Y, Wang B, Qi H, Shen J, Zhu L et al. 2015. Aberrantly expressed Fra-1 by IL-6/STAT3 transactivation promotes colorectal cancer aggressiveness through epithelial-mesenchymal transition. *Carcinogenesis* **36**: 459-468.

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.

Marcotte R, Sayad A, Brown KR, Sanchez-Garcia F, Reimand J, Haider M, Virtanen C, Bradner JE, Bader GD, Mills GB et al. 2016. Functional Genomic Landscape of Human Breast Cancer Drivers, Vulnerabilities, and Resistance. *Cell* **164**: 293-309.

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

Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R et al. 2016. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**: D110-115.

Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe JP, Tong F et al. 2006. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**: 515-527.

Northcott PA, Lee C, Zichner T, Stutz AM, Erkek S, Kawauchi D, Shih DJ, Hovestadt V, Zapatka M, Sturm D et al. 2014. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**: 428-434.

Orom UA, Shiekhattar R. 2013. Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell* **154**: 1190-1193.

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA et al. 2000. Molecular portraits of human breast tumours. *Nature* **406**: 747-752.

Ray PS, Wang J, Qu Y, Sim MS, Shamonki J, Bagaria SP, Ye X, Liu B, Elashoff D, Hoon DS et al. 2010. FOXC1 is a potential prognostic biomarker with functional significance in basal-like breast cancer. *Cancer Res* **70**: 3870-3876.

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

Risolino M, Mandia N, Iavarone F, Dardaei L, Longobardi E, Fernandez S, Talotta F, Bianchi F, Pisati F, Spaggiari L et al. 2014. Transcription factor PREP1 induces EMT and metastasis by controlling the TGF-beta-SMAD3 pathway in non-small cell lung adenocarcinoma. *Proc Natl Acad Sci U S A* **111**: E3775-3784.

Shaulian E. 2010. AP-1--The Jun proteins: Oncogenes or tumor suppressors in disguise? *Cell Signal* **22**: 894-899.

Shaulian E, Karin M. 2002. AP-1 as a regulator of cell life and death. *Nat Cell Biol* **4**: E131-136.

Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, Furey TS. 2013. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res* **23**: 777-788.

Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**: 272-286.

Sotiriou C, Pusztai L. 2009. Gene-expression signatures in breast cancer. *N Engl J Med* **360**: 790-800.

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.

Van Dyck F, Declercq J, Braem CV, Van de Ven WJ. 2007. PLAG1, the prototype of the PLAG gene family: versatility in tumour development (review). *Int J Oncol* **30**: 765-774.

Verde P, Casalino L, Talotta F, Yaniv M, Weitzman JB. 2007. Deciphering AP-1 function in tumorigenesis: fra-ternizing on target promoters. *Cell Cycle* **6**: 2633-2639.

Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, Qiu J, Liu W, Kaikkonen MU, Ohgi KA et al. 2011. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**: 390-394.

Wang J, Ray PS, Sim MS, Zhou XZ, Lu KP, Lee AV, Lin X, Bagaria SP, Giuliano AE, Cui X. 2012. FOXC1 regulates the functions of human basal-like breast cancer cells by activating NF-kappaB signaling. *Oncogene* **31**: 4798-4802.

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)*.

Young MR, Colburn NH. 2006. Fra-1 a target for cancer prevention or intervention. *Gene* **379**: 1-11.

Zhao C, Qiao Y, Jonsson P, Wang J, Xu L, Rouhi P, Sinha I, Cao Y, Williams C, Dahlman-Wright K. 2014. Genome-wide profiling of AP-1-regulated transcription provides insights into the invasiveness of triple-negative breast cancer. *Cancer Res* **74**: 3983-3994.

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.

FIGURE LEGENDS

Figure 1. Transcriptional and epigenomic profiling identifies putative enhancers across the breast cancer genome.

(**A)** *(Top)* Features of cell lines representing five distinct molecular subtypes of breast cancer, which were used in this study, including ER, PR, and HER2 status. CL, claudin low. (*Bottom)* Depiction of the transcriptional (GRO-seq and RNA-seq) and epigenomic (ChIP-seq) profiles generated for each cell line.

(B) Genome browser views of GRO-seq and histone modification ChIP-seq data from a normal breast epithelial cell line (76N-F2V) showing the features of a typical bi-directionally transcribed enhancer *(red box with dashed line)* and its nearest neighboring gene (*SIRPA*). Key features include transcription *(red/blue)*, as well as histone modifications typically enriched at enhancers *(green)*, promoters *(brown)*, gene bodies *(purple)*, and repressed chromatin *(turquoise)*.

Figure 2. Unbiased, genome-wide prediction of active enhancers using GRO-seq data.

(A) Overview of the computational pipeline used for the genome-wide annotation of enhancer transcripts (eRNAs) and prediction of active enhancers using GRO-seq data.

(B) Catalogue of all predicted active enhancers in the breast cancer cell lines listed in Figure 1A defined by enhancer transcription using the pipeline shown in panel (A). Red indicates cell type-specific enhancers and blue represents enhancers transcribed in at least one other cell line. The colored circles indicate the molecular subtype of each cell line.

(C) Metaplot analyses showing correlations among enhancer transcription, DNase I hypersensitivity, H3K27ac enrichment, and H3K4me1 enrichment for the uniquely transcribed enhancers identified in ZR-75-1 cells in panel (B). Metaplots for the same genomic loci in

MCF-10A cells are shown for comparison *(grey lines)*.

(D) Stacked bar chart comparing enhancer prediction methods in ZR-75-1 cells. Intergenic enhancers were called using prediction methods based on four different enhancer features: enhancer transcription (using GRO-seq and the pipeline shown in panel A), DNase I hypersensitivity, H3K4me1 enrichment, or H3K27ac enrichment. The percentage of called enhancers from each prediction method overlapping enhancers called using one or more of the other methods is shown.

E) Venn diagrams showing the number of enhancers called in ZR-75-1 cells and the overlap of enhancers called using DNase I and H3K4me1 *(left)*, H3K4me1 and H3K27ac *(middle)*, and H3K27ac and enhancer transcription *(right)*.

Figure 3. Actively transcribed enhancers dictate subtype-specific transcriptional programs.

(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. 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 the nearest neighboring genes to uniquely transcribed enhancers (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, $p < 0.05$).

(C) The nearest neighboring genes to the uniquely transcribed enhancers in each cell line are preferentially expressed in patient tumor samples of the same molecular subtype. *(Left box plot)* ZR-75-1 cells represent the luminal A breast cancer molecular subtype. The nearest neighboring

genes to the uniquely transcribed enhancers for ZR-75-1 cells are more highly expressed in luminal A patient tumor samples compared to the other tumors types. *(Right box plot)* For comparison, MDA-MB-468 cells represent the TN basal breast cancer molecular subtype. The nearest neighboring genes to the uniquely transcribed enhancers in MDA-MB-468 cells are more highly expressed in TN basal tumor samples compared to the other tumor types. Observed differences are significant as determined by an ANOVA comparison of the means (p-value < 0.00001).

Figure 4. A functional score of enhancer elements identifies subtype-specific enhancers and their cognate TFs that drive subtype-specific gene expression in breast cancer cells.

(A) Diagram of the data used for determining the Total Functional Score of Enhancer Elements (TFSEE) across breast cancer cell lines. TFSEE simultaneously identifies putative subtype-specific enhancers and their cognate TFs by integrating the magnitude of enhancer transcription (GRO-seq), TF mRNA expression levels (RNA-seq), TF motif p-values (MEME/Tomtom), and enrichment of H3K4me1 and H3K27ac (ChIP-seq). This analysis yields the location, activity level, and predicted TFs at each enhancer in all breast cancer cells.

(B) Unsupervised hierarchical clustering of cell line-normalized TFSEE scores shown in a heatmap representation. Two major clades arise from this analysis, highlighting key TFs for TN / Normal subtypes versus Luminal/HER2+ subtypes.

(C and D) Rank order frequency distribution of TFs enriched in the TN/Normal-like clade (panel C) and the Luminal/HER2+ clade (panel D) identified using TFSEE. The top TFs in each clade are noted.

(E) Box plots of expression values for members of the Forkhead Box family of TFs in patient breast tumor samples, confirming the differential enrichment of these TFs in the TN/Normal-

Like versus the Luminal/HER2+ clades shown in panels C and D. Observed differences are significant as determined by an ANOVA comparison of the means (p-value < 0.00001).

Figure 5. TFSEE-Predicted transcription factors are enriched at sites of enhancer transcription.

(A) Genome browser views of transcribed enhancers (GRO-seq; H3K27ac and H3K4me1) (*left*) and corresponding ChIP-qPCR experiments for their cognate predicted TFs (*right*) for two TFs highly enriched in the TN/Normal-Like clade based on TFSEE: PLAG1 (*top row*) and RUNX2 (*bottom row*). The data shown are from TN basal breast cancer cells (HCC-1937). Enhancer transcription provides a measure of activity and the locations of the enhancers, while ChIP-qPCR confirms the binding of the predicted TF at that site. The enhancers are designated by their genomic coordinates. A negative control region not bound by the predicted TFs is shown for comparison. 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).

(B) A set of experiments similar to those shown in panel A for two highly enriched TFs in the Luminal/HER2+ clade: FOXA1 (*top row*) and HLF (*bottom row*). The data shown are from luminal A breast cancer cells (MCF-7). 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).

Figure 6. FOSL1 is enriched at transcribed enhancers in TN cells, regulates cell proliferation, and correlates with breast cancer patient outcomes.

(A) Genome browser views of a transcribed enhancer predicted to be bound by FOSL1 in TN

cells (GRO-seq; H3K27ac and H3K4me1). The data shown are from TN basal breast cancer cells (HCC-1937).

(B) ChIP-qPCR for FOSL1 at two transcribed enhancers predicted to be bound by FOSL1, shown in a TN basal cell line (HCC-1937). The enhancers are designated by their genomic coordinates. Genome browser shots for the enhancer found on chr 5 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).

(C) siRNA-mediated knockdown of FOSL1 in a TN basal 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 *FOSL1* mRNA expression levels in patient tumor samples confirm enrichment of FOSL1 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) *FOSL1* mRNA expression is predictive of clinical outcomes in ER-negative (ER-) breast tumor patients. Kaplan-Meier survival analyses of patients expressing high levels of *FOSL1* mRNA *(maroon line)* exhibit a poorer outcome compared to patients expressing low levels of *FOSL1* 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.

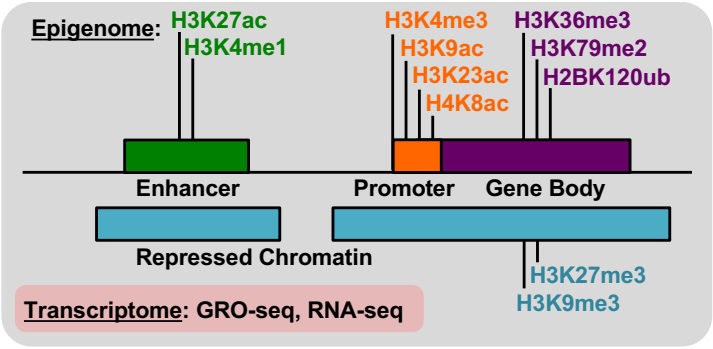
(F) siRNA-mediated knockdown of FOSL1 in two TN cell lines (MDA-MB-468 and HCC-1937) causes decreased proliferation and viability, as observed in proliferation assays (*left panels*) and

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).

(G) siRNA-mediated knockdown of FOSL1 in Luminal A cell line (MCF-7) shows no significant effects on proliferation or viability compared to TN cells. siRNA-mediated knockdown of Polo-like Kinase 1 (PLK1) serves as a positive control. Each bar represents the mean + SEM, n = 3. The 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

Cell Line	Subtype	ER	PR	HER2	Type
76N-F2V	Normal	–	–	–	Mammary Breast Epithelium
MCF-10A	Normal	–	–	–	Mammary Gland, Fibrocystic
MCF-7	Luminal A	+	+	–	Invasive Ductal Carcinoma
ZR-75-1	Luminal A	+	–	–	Invasive Ductal Carcinoma
MDA-MB-361	Luminal B	+	–	+	Adenocarcinoma
UACC812	Luminal B	+	–	+	Invasive Ductal Carcinoma
SKBR3	HER2+	–	–	+	Adenocarcinoma
AU565	HER2+	–	–	+	Adenocarcinoma
HCC-1954	HER2+	–	–	+	Ductal Carcinoma
MDA-MB-468	TN Basal	–	–	–	Adenocarcinoma
HCC-1937	TN Basal	–	–	–	Ductal Carcinoma
MDA-MB-231	TN CL	–	–	–	Adenocarcinoma
MDA-MB-436	TN CL	–	–	–	Invasive Ductal Carcinoma



B

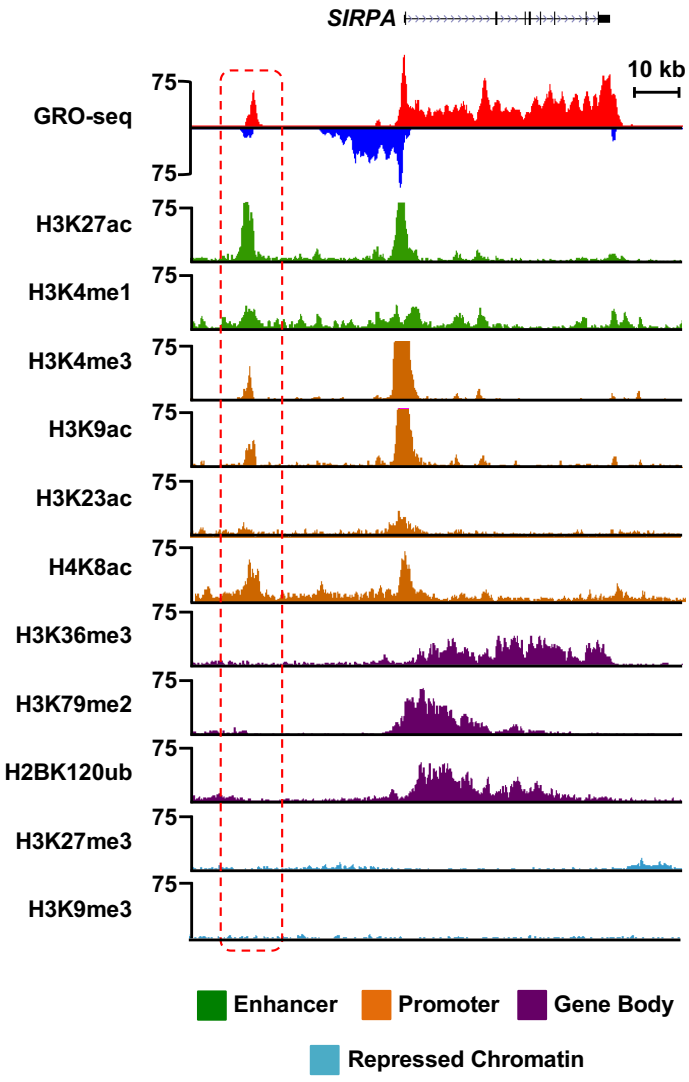


Figure 1 - Franco *et al.* (2017)

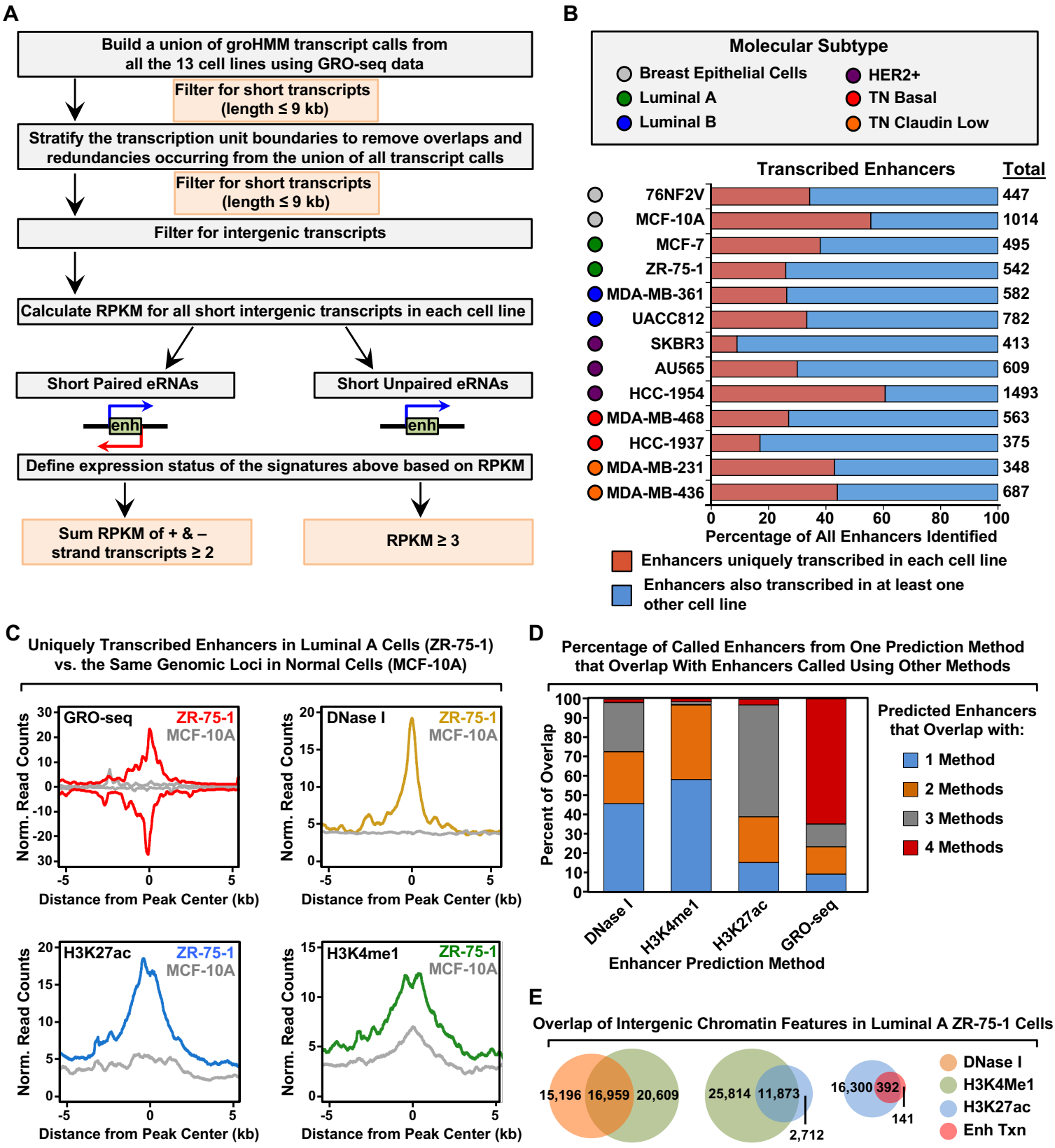


Figure 2 - Franco *et al.* (2017)

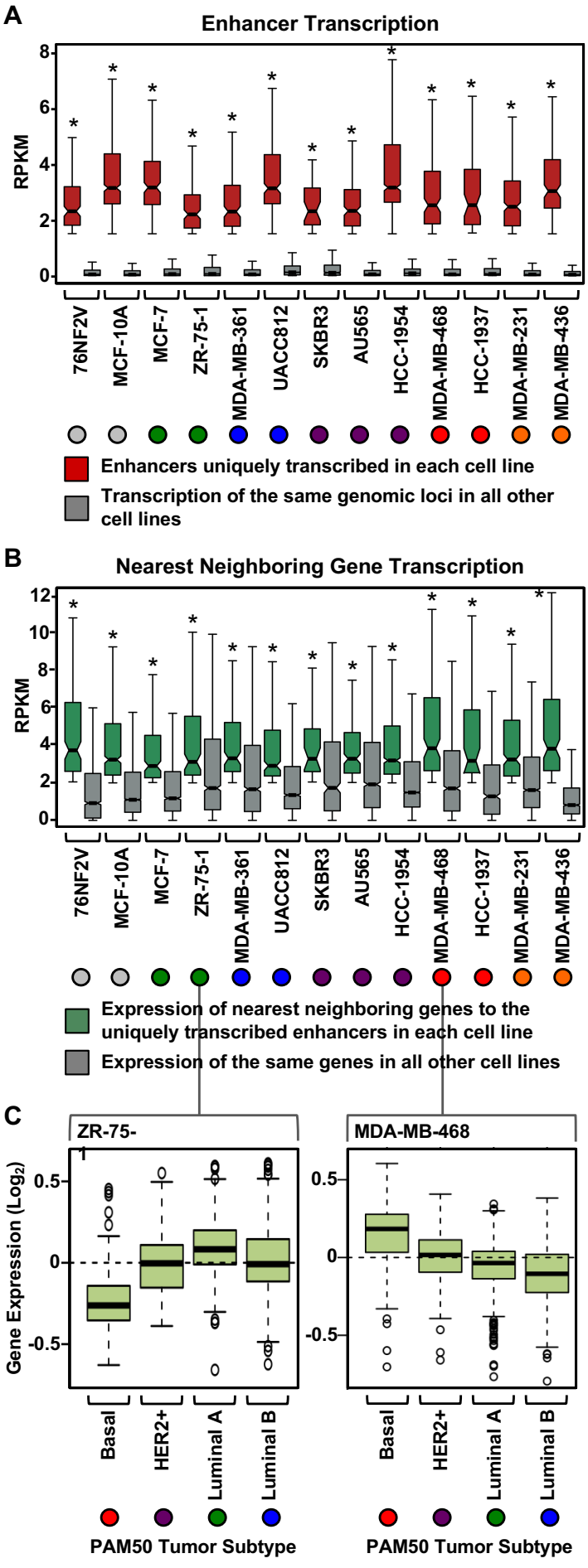


Figure 3 - Franco *et al.* (2017)

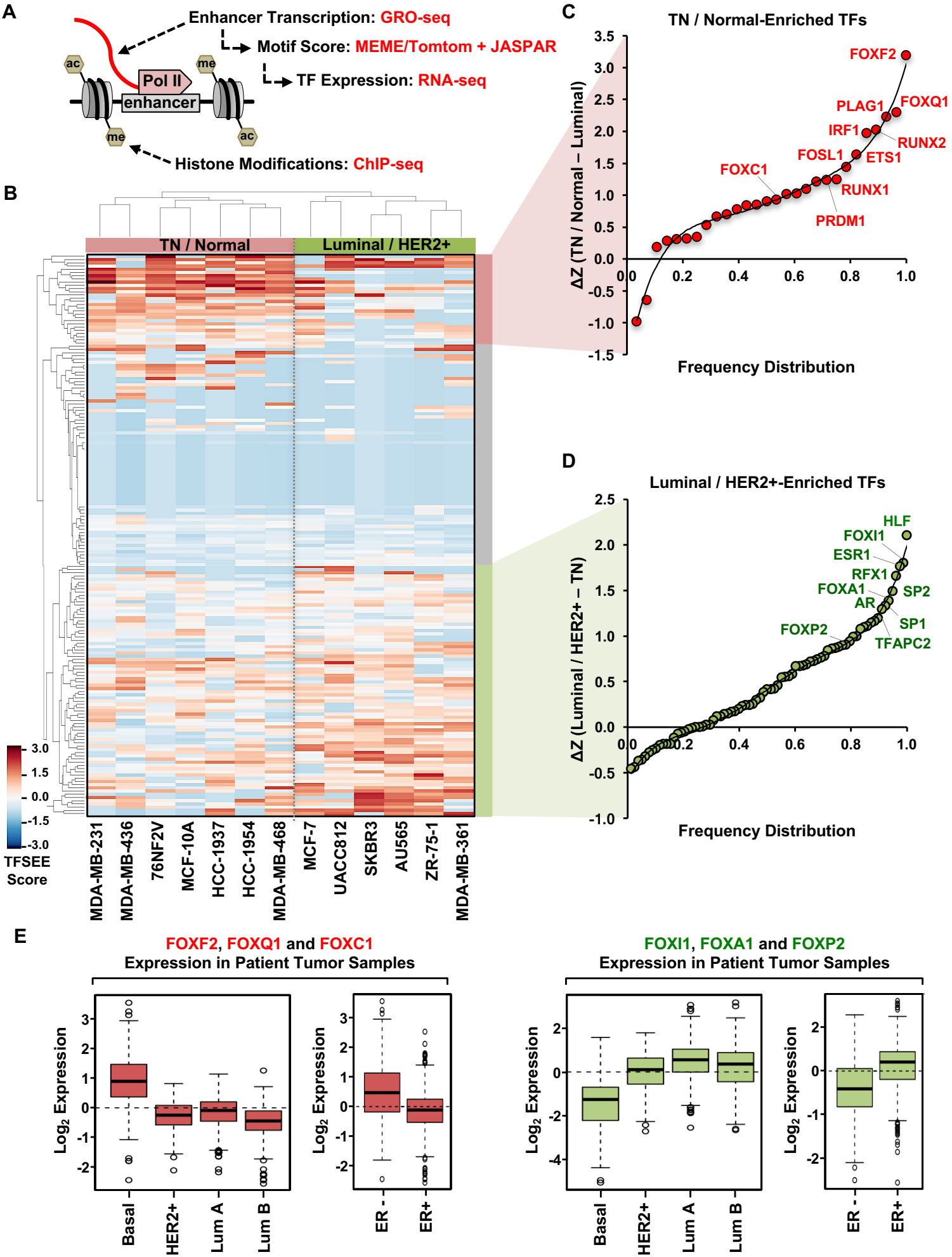


Figure 4 - Franco *et al.* (2017)

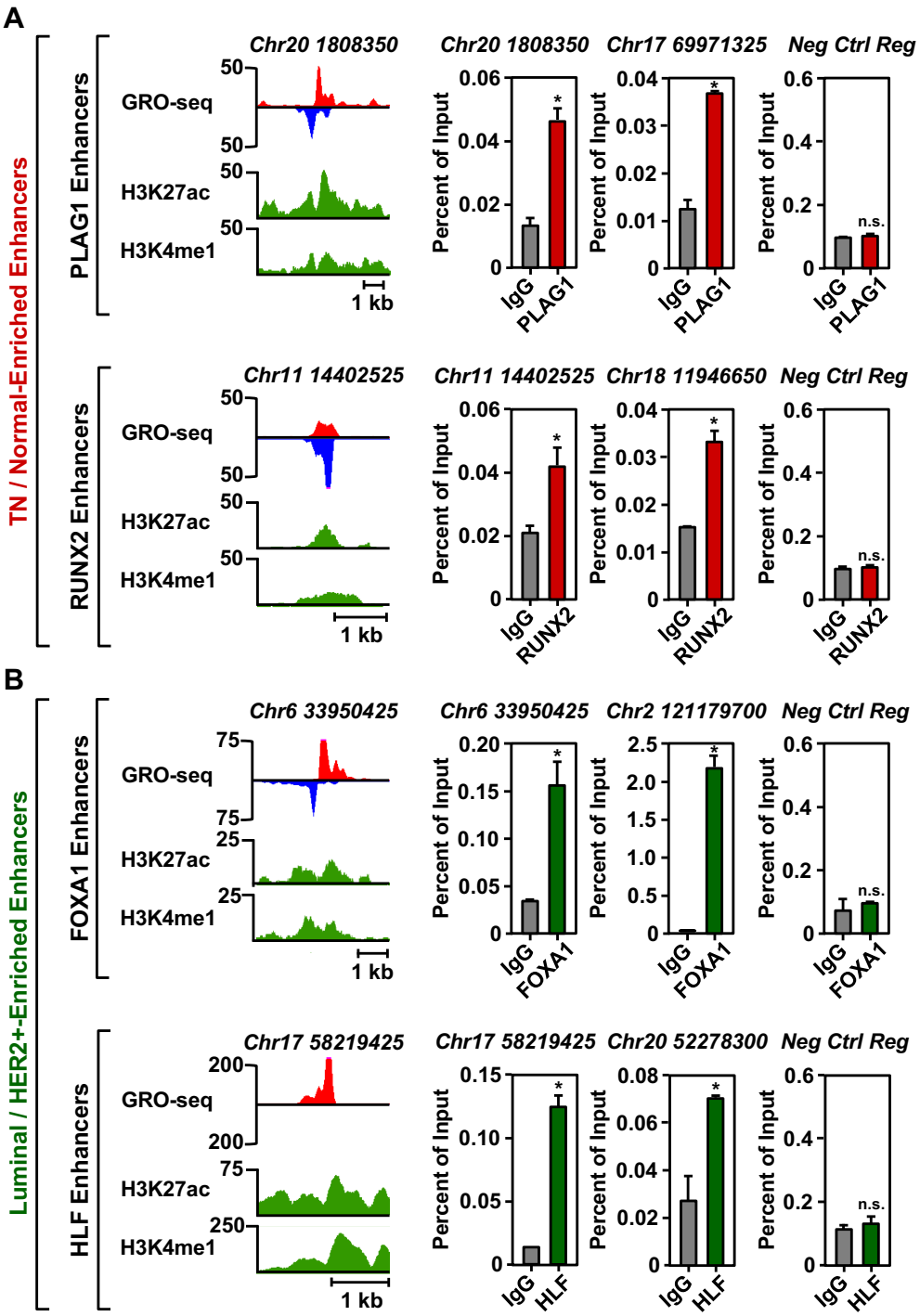


Figure 5 - Franco *et al.* (2017)

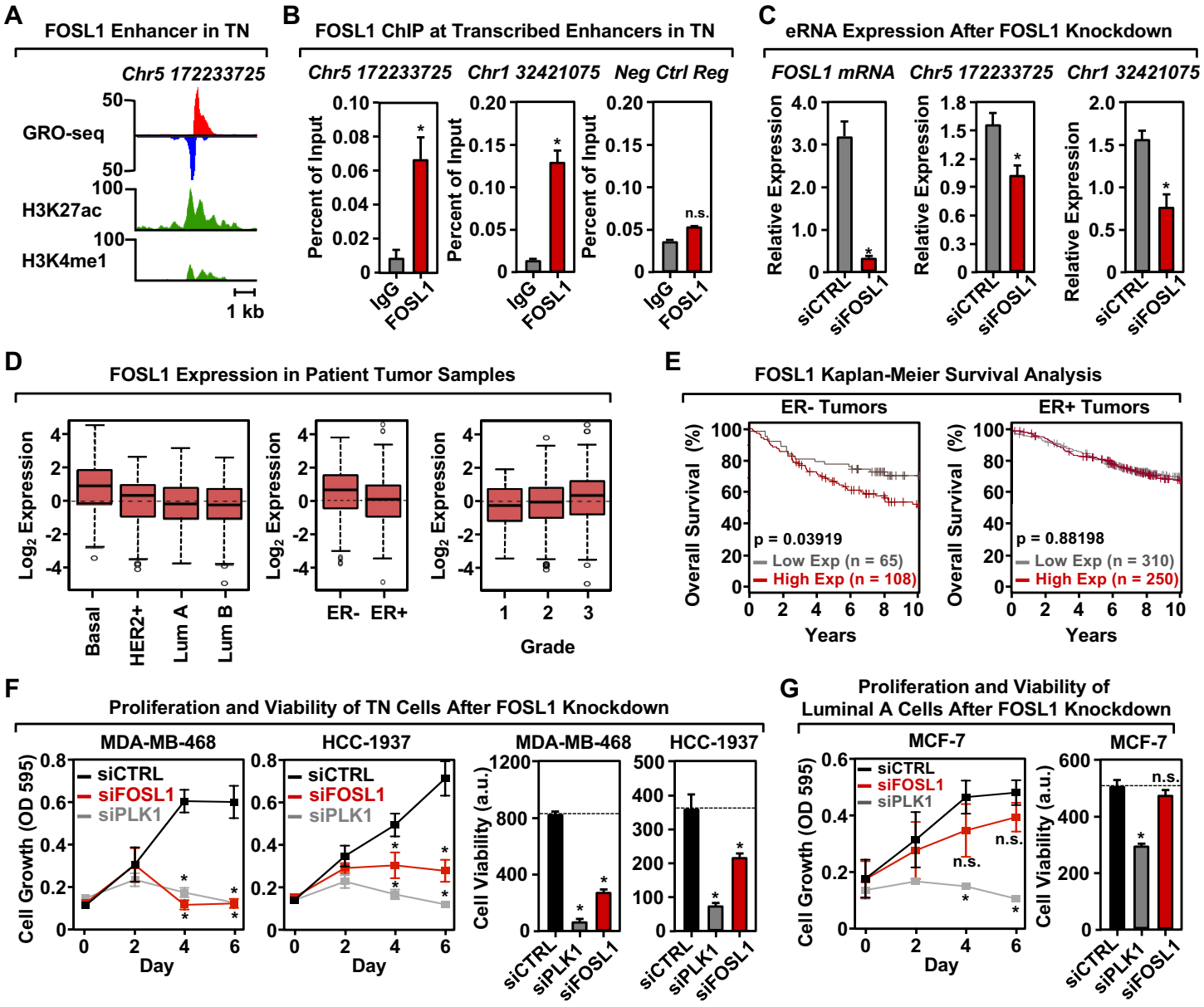


Figure 6 - Franco *et al.* (2017)