Chromatin accessibility and active transcription factors in primary human invasive lobular and ductal breast carcinomas

Sanghoon Lee[1,3], Hatice Ulku Osmanbeyoglu[1,2,3,*]

[1]Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, USA, University of Pittsburgh, Pittsburgh, USA

[2]Department of Bioengineering, School of Engineering,

[3]UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, USA

[*] Correspondence to: Hatice Ulku Osmanbeyoglu (osmanbeyogluhu@pitt.edu)

ORCID ID: 0000-0002-4972-4347

Abstract

Invasive lobular breast carcinoma (ILC), the second most prevalent histological subtype of breast cancer, exhibits unique molecular features compared with the more common invasive ductal carcinoma (IDC). While genomic and transcriptomic features of ILC and IDC have been characterized, genome-wide chromatin accessibility pattern differences between ILC and IDC remain largely unexplored. Here, we characterized tumor-intrinsic chromatin accessibility differences between ILC and IDC using primary tumors from The Cancer Genome Atlas (TCGA) breast cancer assay for transposase-accessible chromatin with sequencing (ATAC-seq) dataset. We identified distinct patterns of genome-wide chromatin accessibility in ILC and IDC. Inferred patient-specific transcription factor (TF) motif activities revealed regulatory differences between and within ILC and IDC tumors. EGR1, RUNX3, TP63, STAT6, SOX family, and TEAD family TFs were higher in ILC, while ATF4, PBX3, SPDEF, PITX family, and FOX family TFs were higher in IDC. This study reveals the distinct epigenomic features of ILC and IDC and the active TFs driving cancer progression that may provide valuable information on patient prognosis.

Introduction

Breast cancer, the leading malignancy in women, has molecularly discrete subtypes based on the expression of estrogen receptors alpha (ESR1, also known as ER), progesterone receptors (PGR, also known as PR), and/or the amplification of human epidermal growth factor receptor 2 (*ERBB2*, also known as HER2). Of the ~200,000 newly diagnosed cases of invasive breast cancer each year, 70% are estrogen receptor-positive (ER+) (1). More patients die from advanced ER+ breast cancer than all other breast cancer types combined. ER+ breast cancer comprises two main histological subtypes with varying molecular features and clinical behaviors: 85–90% are invasive ductal carcinoma (IDC) and 10–15% are invasive lobular carcinoma (ILC) (2-4). ILC is

predominantly ER+ and PGR-positive but can, rarely, show HER2 protein overexpression. While ILC is initially associated with longer disease-free survival and a better response to adjuvant hormonal therapy than IDC, the long-term prognosis for ILC is worse than IDC; 30% of ILC patients will develop late-onset metastatic disease up to 10 years after the initial diagnosis (5). In several retrospective studies that compared clinical and pathological responses, ILC also appeared less responsive to chemotherapy than IDC (6-8).

Although ER+ ILC and IDC tumors are treated clinically as a single disease (9), recent studies have established ER+ ILC as a distinct disease with unique sites of metastasis, frequent presentation of multifocal disease, delayed relapses, and decreased long-term survival compared to ER+ IDC tumors (10-14). Large-scale studies from The Cancer Genome Atlas (TCGA) and the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) have reported genomic and transcriptomic analyses on resected IDC and ILC tumors (15,16). The distinguishing genomic feature of ILC is the loss of E-cadherin, a protein that mediates epithelial-specific cell-cell adhesion (17). The loss of E-cadherin in *CDH1* mutants is associated with phosphatidylinositol 3 kinase (PI3K)/Akt pathway activation and epidermal growth factor receptor (EGFR) overexpression, which are major drivers in breast cancer (17). E-cadherin knock-outs of IDC cell lines result in remodeling of transcriptomic membranous systems, greater resemblance to ILCs, and increased sensitivity to IFN-γ-mediated growth inhibition via activation of IRF1 (18).

The National Cancer Institute (NCI) Genomic Data Analysis Network (GDAN) generated transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) data for a subset of TCGA samples (404 patients) (19), including ER+ ILC and IDC tumors. ATAC-seq is a transformative technology for mapping the chromatin-accessible loci genome-wide and identifying nucleosome-free positions in regulatory regions. It needs only ~50,000 cells and is simpler than previous methods, such as DNase-seq (20). Epigenomic changes at the level of chromatin accessibility, potentially linked to distinct differentiation states, might reveal epigenetic reprogramming and developmental origin differences between ER+ ILC and IDC. However, chromatin accessibility landscape differences between ER+ ILC and IDC tumors based on patient samples have not been systematically characterized. Complementing genomic and transcriptomic studies, we mapped the epigenetic heterogeneity in ER+ ILC and IDC with a systematic analysis of chromatin accessibility patterns based on the primary tumor breast cancer TCGA ATAC-seq dataset (21). We defined the compendium of ~190,000 genome-wide cis-regulatory regions in breast cancer ER+ ILC and IDC with 11,762 differentially accessible (DA)

peaks between ILC and IDC, which represented 5.98% of total ATAC-seq peaks. EGR1, RUNX3, TP63, STAT6, SOX family, and TEAD family transcription factor (TF) activities were significantly higher in ILC, consistent with their role in the regulation of the extracellular matrix and growth factor signaling pathways, whereas ATF4, PBX3, SPDEF, PITX family, and FOX family TF activities were significantly higher in IDC. The inferred TF activities and context-specific target genes based on ATAC-seq data identified biological pathways that are likely distinct in ER+ ILC vs. IDC. Together, these results provide new insights into ER+ ILC and IDC biology.

Results

Chromatin accessibility differences between ER+/HER2- ILC and IDC breast cancer

Epigenetic differences at the level of chromatin accessibility, potentially linked to different differentiation states, could distinguish ER+ ILC and IDC tumors. We characterized tumor cell-intrinsic chromatin accessibility patterns using primary ER+ breast cancer ATAC-seq data from TCGA (19). Using an atlas of 196,546 peaks across all ILC and IDC breast tumors (n=67) (19), we grouped tumors according to their histological subtypes and hormone receptor status: ER+/HER2- ILCs (n=13), ER+/HER2- IDCs (n=30), and ER+/HER2+ IDCs (n=7). Principal component analysis (PCA) of peak read counts showed that all these ER+ tumors were clustered closely, the ER+/HER2- or HER2+ IDCs more closely associated, and the ILCs slightly separated **(Fig. 1A)**. In addition, we found three ER+/HER2- IDC samples and one ER+/HER2+ IDC sample that were outliers. We observed similar patterns and the same outliers through PCA analysis of RNA-seq and reverse phase protein array (RPPA) data (**Supplementary Figure 1A–B**). Because there were few ER+/HER2+ samples, we used ER+/HER2- ILCs (n=13) and ER+/HER2- IDCs (n = 27) for downstream analyses. Hereafter, we simply denote ILCs vs IDCs omitting ER+/HER-.

To understand epigenomic landscape differences between ILCs and IDCs We analyzed differential chromatin accessibility. We found 11,762 DA peaks (absolute $\log_2$ fold change > 1.0 and adjusted p-value < 0.05) between ILCs and IDCs, which represented 5.98% of all ATAC-seq peaks (**Fig. 1B–C**). Among these, 5,124 peaks (2.61%) showed increased accessibility in ILCs, and 6,638 peaks (3.38%) showed increased accessibility in IDCs. Most of the DA ATAC-seq peaks were at distal intergenic regions (48.46% for ILCs and 40.79% for IDCs); 36.26% for ILCs and 36.84% for IDCs were at introns; 9.83% for ILCs and 16.64% for IDCs were at promoters; and 3.03% for ILCs and 3.5% for IDCs were at exons (**Fig. 1D**).

We used HOMER motif analysis of DA ATAC-seq peaks to identify key TFs driving the expression differences between IDCs and ILCs (22). DA sites in ILCs were highly enriched with binding motifs for the Sry-related HMG box (SOX), TEA Domain (TEAD), runt-related transcription factor (RUNX) family, and TP63 TFs. In contrast, SPDEF and forkhead box (FOX) family binding motifs (e.g., FOXM1, FOXA1, FOXK1) were enriched in IDC sites (**Fig. 1E**). Interestingly, SOX family TFs were major predicted motifs in DA promoter peaks enriched in ILCs, but not in DA distal intergenic peaks (**Supplementary Figure 2A–B**). FOX family TFs were dominant motifs in the DA distal intergenic peaks enriched in IDCs, but not in the DA promoter peaks (**Supplementary Figure 2C–D**). Although SOX-mediated transcription regulation is active in breast cancer (23,24), no association with histological subtypes was reported. Upregulated SOX2 proteins induce chemoresistance in breast cancer cells and promote their stemness property through the recruitment of regulatory T cells (Tregs) to the tumor microenvironment (25,26). SOX4 is an oncogene that promotes PI3K/Akt signaling, angiogenesis, and resistance to cancer therapies in breast tumors; thus, SOX4 is a biomarker for PI3K-targeted therapy (27,28). TEADs interact with transcription coactivator Yes-associated protein (YAP)/transcriptional coactivator with PDZ-binding motif (TAZ), thereby affecting the Hippo pathway that plays a key role in cell proliferation, invasion, and resistance to breast cancer treatment (29,30). TP63, a member of the TP53 gene family, is highly expressed in metaplastic breast cancer (31). SPDEF function depends on the breast cancer subtype (32). SPDEF is a tumor suppressor in triple-negative breast cancer (TNBC) inhibiting tumor invasion and decreasing epithelial-mesenchymal transformation (EMT) (33). In luminal or HER2+ breast cancer, SPDEF is an oncogene (34). FOXA1 proteins enhance hormone-driven ER activity and binding to intergenic regions of DNA in ER+ breast cancer (35). FOXA1 also inhibits EMT and cell growth by modulating E-cadherin, leading to a better prognosis (36).

To identify key biological processes that drive oncogenic gene expression differences between IDCs and ILCs, we analyzed the pathways for DA ATAC-seq peaks using the Genomic Regions Enrichment of Annotations Tool (GREAT) (37). The DA peaks enriched in ILCs and IDCs were associated with different pathways. Oxidative stress response, interleukin signaling, and p53 pathways were overrepresented in ILCs, whereas endothelin signaling, EGF receptor signaling, T cell activation, inflammation, and angiogenesis pathways were overrepresented in IDCs (p-value < 0.01) (**Fig. 1F**). Interestingly, the PI3K pathway was enriched in both ILCs and IDCs. Thus, the epigenomic differences identified distinct TF motif enrichment and biological signatures between ILCs and IDCs.

To correlate alterations in chromatin accessibility with transcriptional output, we integrated ATAC-seq data with RNA-seq data. Consistent with the correlation of global differential accessibility and expression, differential accessibility of individual genes was often associated with significant differential expression (**Fig. 2A–B**). Genes with the greatest differential accessibility between ILCs and IDCs at their promoter, intronic, and nearby intergenic peaks are shown in **Fig. 2B**. For example, FAM83A and ERICH5 were significantly more accessible in IDCs, while FAM189A and SSPN were significantly more accessible in ILCs (**Fig. 2C**). FAM83A is involved in the chemoresistance and stemness of breast cancer through its interaction with the EGFR/PI3K/AKT signaling pathway (38,39). FAM189A is down-regulated in breast cancer (40,41) and SSPN is down-regulated in TNBC (42). Overall, we identified context-specific features, including accessibility and expression patterns associated with IDCs vs. ILCs.

The coordinated activity of many TFs characterizes ILC and IDC tumors

We inferred sample-specific TF motif activities based on genome-wide chromatin accessibility data using CREMA (Cis-Regulatory Element Motif Activities, see Methods). This allowed us to map chromatin accessibility profiles to a lower-dimensional inferred TF activity space, largely preserving the relationships between samples. Inferred activities of 29 TF motifs were significantly associated with histological subtypes by false discovery rate (FDR)-corrected p-value < 0.05 and absolute mean activity difference > 0.035 (**Fig. 3A and Supplement Figure 3**). We found that Early Growth Response 1 (EGR1) (43), TEAD family (TEAD1, TEAD3, and TEAD4), SOX family, (SOX2, SOX4, and SOX8), and RUNX3_BLC11A TFs had significantly higher activities in ILCs than IDCs (**Fig. 3B**). Similarly, FOX family (FOXA1, FOXA3, FOXC2, FOXL1, FOXK1, FOXP2, FOXP3, FOXD3, FOXI1, and FOXF1), Paired Like Homeodomain (PITX family) (PITX1 and PITX2), PBX3, and HSF4 had significantly higher activities in IDCs than ILCs (**Fig. 3C**). EGR1 mRNA is upregulated in ILCs (43), but other TFs have not been studied in the context of ILCs and IDCs. TF activities from the same families were also correlated across samples (**Fig. 3D**). Overall, these results were consistent with the motif enrichment analysis based on the DA peaks in ILCs vs. IDCs (**Fig. 1E**).

To determine whether TF activities were associated with protein expression, we compared immunohistochemically (IHC) stained images for available TFs between ILCs and IDCs obtained from the Human Protein Atlas (HPA) database (44). The HPA tissue images of breast cancer provide histological subtype but not hormone receptor subtype information. Therefore, we used

only ILC or IDC images that had ER high/medium staining intensity and HER2 low/not detected staining intensity (**Supplementary Figure 4**). The IHC images demonstrated that EGR1, TEAD4, SOX2, and BCL11A proteins were highly expressed in ILCs, but were not detected in IDCs (**Fig. 3E**) consistent with the increased TF activities in ILCs. Likewise, the IHC images of FOXA1, FOXA3, ATF4, and ZNF35 showed medium or high protein expression in IDCs but were not detected or showed low expression in ILCs (**Fig. 3F**). Overall, the staining images showed that increased TF activities in ILCs or IDCs were associated with protein expression in the corresponding histological subtypes.

We looked for functional evidence of IDC- and ILC-specific TF regulators using published breast cancer genome-wide ''dropout'' screens using pooled small hairpin (shRNA) libraries (45). The dataset included three ER+ ILC cell lines and 11 ER+ IDC cell lines (HER2 type data was not available). We ran **s**mall **i**nterfering RNA (siRNA)/shRNA **m**ixed-**e**ffect **m**odel (siMEM) for three ER+ ILC cell lines vs. other breast cancer cell lines or for 11 ER+ IDC cell lines vs others to calculate context-specific essentiality scores for IDC- or ILC-specific TFs. **Supplementary Table 1** lists the essentiality scores of the TFs in ER+ ILC (15/18 TFs) or ER+ IDC (15/19 TFs) cell lines. We identified RUNX3, SOX4, TEAD3, UBP1, NFIA, and BCL11A as essential for ER+ ILC cell proliferation, and FOXA1 and SPDEF as essential for ER+ IDC cell proliferation (FDR < 0.2). RUNX3 inhibits estrogen-dependent proliferation by targeting ERα in breast cancer cell lines and functions as a tumor suppressor, but its role has not been defined (46). Interestingly, we identified FOXA1 and SPDEF as the top essential TFs for ER+ IDC cells. The original genome-wide shRNA screening also identified FOXA1 and SPEDEF as the top luminal/HER2 essential genes out of 975 essential genes (45). In ER+ breast cancer cell lines, FOXA1 inhibits cell growth by inducing E-cadherin expression and suppressing ER pathway activity, which suggests that FOXA1 can be a favorable prognostic marker in human breast cancer (36,47,48). SPDEF expression is also enriched in luminal tumors and promotes luminal differentiation and survival of ER+ cells (49).

Gene sets for IDC- and ILC-specific TFs display coherent functions and are consistent with gene expression changes

Tables 1 and **2** summarize the enriched canonical pathways for the target genes of the TFs associated with ILCs or IDCs. Interestingly, most TFs with high activities in ILCs, including EGR1, HMGA1, NFIX_NIFB, RBPJ, SOX family, TEAD family, TP63, UBP1, and ZFHX3, were associated with genes encoding extracellular matrix (ECM)-associated proteins for structure or remodeling. IDC-specific TFs including ARNT, ATF4, and ZNF35 were associated with PI3K or

IL2 signaling mediated by PI3K. We next examined TF target gene expression in IDCs and ILCs using TCGA and METABRIC gene expression data. **Fig. 4** shows the cumulative distribution of expression changes between ILC- or IDC-specific TF activities for predicted targets based on ATAC-seq data. The TF regulators identified for ILC and IDC were associated with the upregulation of their targets. There was significant upregulation of motif-based targets of TFs based on ATAC-seq relative to all genes in the TCGA and METABRIC tumor data (p-value < 1e−3, Kolmogorov-Smirnov test). Thus, ILCs and IDCs are associated with different TFs, and the TFs regulate target gene expression and biological pathways specific for ILCs vs. IDCs.

Discussion

Many studies have examined the distinct molecular features and prognostic outcomes for ILC vs. IDC tumors**.** Using TCGA ATAC-seq data we provide here the first comprehensive genome-wide chromatin accessibility landscape analysis of ER+ ILC and IDC using primary breast cancers. We identified a chromatin accessibility signature, TFs, and biological pathways specific to ER+ ILC and IDC tumors. TFs (e.g. EGR1, SOX, and TEAD family) involved in ECM interactions and signaling pathways had higher activity in ILC compared to IDC. Little is known about the role of these TFs in ILC; however, one study reported increased EGR1 expression in ILC (43). The differences in activities of TFs in ILCs vs. IDCs based on chromatin accessibility were also consistent with TF protein expression and upregulation of TF target gene expression.

This study provides the first in-depth characterization of the genome-wide chromatin accessibility landscape of ER+ ILC and IDC primary tumors samples, identifying several differences in the epigenomic profiles between ILC and IDC and highlighting potentially clinically relevant pathways. Our deep analyses of ATAC-seq data generated a global regulatory network with the corresponding TFs in IDC and ILCs that could provide useful clinical insights into the differences between these two cancers.

Methods

Data and preprocessing

TCGA breast cancer data: We downloaded TCGA breast cancer (BRCA) ATAC-seq raw bam files (n=75) and RNA-seq raw fastq.gz files from NCI Genomic Data Commons (GDC) data portal (https://portal.gdc.cancer.gov) (50). Breast cancer peak calls and bigwig files of ATAC-seq profiles were downloaded from https://gdc.cancer.gov/about-data/publications/ATACseq-AWG. For the hormone receptor subtypes of TCGA BRCA tumors, we followed the annotation data

provided by the TCGA ATAC-seq data publication, Supplementary Data 1 (21). The RNA-seq read count and reverse phase protein array (RPPA, replicate-base normalization) data were downloaded from Xena Functional Genomics Explorer (https://xenabrowser.net/hub/) GDC and TCGA hub, respectively.

METABRIC breast cancer data: We downloaded METABRIC microarray data from cBioPortal (https://www.cbioportal.org/study/summary?id=brca_metabric) (51).

Human Protein Atlas: The Human Protein Atlas (https://www.proteinatlas.org) is a public resource that extracts and analyzes information, including images of immunohistochemistry (IHC), protein profiling, and pathologic information, from specimens and clinical material from cancer patients to determine global protein expression (52). Here, we compared the protein expression of available TFs in ILC and IDC tissues by IHC image.

Breast cancer cell lines shRNA screen: To identify and validate the TFs essential in ER+ ILC and ER+ IDC cell lines, we accessed the data for whole-genome small hairpin RNA (shRNA) ‘‘dropout screens’’ on three ER+ ILC and 11 ER+ IDC breast cancer cell lines (45) (GSE74702).

Differential peak accessibility

Reads aligning to atlas peak regions (hg19) were counted using the countOverlaps function of the R packages, GenomicAlignments (v1.30) (53) and GenomicRanges (v1.46.1) (53). To remove the bias created by low count peaks, we filtered 19,364 peaks with mean counts of less than 30 across all samples. Differential accessibility of peaks was calculated using DESeq2 (v1.34.0) (54). DA peaks were defined as significant if they had an adjusted p-value < 0.05 and the magnitude of the DESeq-normalized counts changed by a stringent factor of one or more between ER+HER2- ILC and ER+/HER2- IDC. The significant DA peaks were aggregated and represented in the hierarchical clustering heatmap using the DESeq size-factor normalized read counts and the ‘complete’ distance metric for the clustering algorithm. We used ChiPseeker (55) and clusterProfiler (56) R packages for peak region annotation and visualization of peak coverage over chromosomes.

ATAC-seq peak clustering

For visualization of ER+/HER2- ILC, ER+/HER2- IDC, and ER+/HER2+ IDC tumors by PCA, we used DESeq2 (v1.34.0) (54) to fit multi-factorial models to ATAC-seq read counts in peaks. We

used all peak counts and generated DESeq2 models including factors for hormone receptor subtypes (ER +/- and HER2 +/-) and histological classes (ILC vs IDC). Then, we calculated a variance stabilizing transformation from the DESeq2 model and performed PCA.

De novo TF motif analysis

The HOMER v.4.11.1 utility findMotifsGenome.pl (22) was used to identify the top 10 TF motifs enriched in differential accessible peaks. We set 100-bp-wide regions around the DA peak summits with hg19 as the reference genome. We generated custom background regions with a 150-bp-wide range around the peak summits. The top motifs were reported and compared to the HOMER database of known motifs and then manually curated to restrict them to TFs that are expressed based on RNA-seq data and to similar motifs from TFs belonging to the same family.

Pathway enrichment analysis

We used GREAT (Genomic Regions Enrichment of Annotations Tool, v1.26) to associate the subcluster of the DA peaks to genes and used pathway analysis to identify the functional significance of the DA peaks (37).

TF essentiality analyses in ER+ ILC and ER+ IDC cell lines

We used **s**mall **i**nterfering RNA (siRNA)/shRNA **m**ixed-**e**ffect **m**odel (siMEM) (45) to score the screening results of the TFs and identify their significant context-specific essentiality between ER+ ILC and ER+ IDC from the shRNA screening data. The significantly essential TFs had an FDR q-value < 0.2 in the siMEM results. The annotation data for ER subtype and histological types in the breast cancer cell lines are available at https://github.com/neellab/simem.

Cis-Regulatory Element Motif Activity analysis

We used the CREMA (Cis-Regulatory Element Motif Activities, https://crema.unibas.ch/) to analyze genome-wide DNA-accessibility, calculate TF motif activities, and identify active cis-regulatory elements (CREs). CREMA first identifies all CREs genome-wide that are accessible in at least one sample, quantifies the accessibility of each CRE in each sample, predicts TF binding sites (TFBSs) for hundreds of TFs in all CREs, and then models the observed accessibilities across samples in terms of these TFBS, inferring the activity of each TF in each sample. A Wilcoxon rank sum test was used to compare TF activities and assess the association between TF and histologic subtypes. Then, the resulting p-values were adjusted for multiple hypothesis testing (across TFs). This analysis was visualized with a scatterplot where the x-axis represents

mean TF activity difference, and the y-axis represents FDR-corrected p-value. The significant TF motifs were selected by absolute mean TF activity difference > 0.035 and FDR-corrected p-value < 0.05.

The TF targets identified by CREMA are CREs, not genes directly. After identifying TF target CREs, the gene-CRE association probabilities are calculated on the basis of distance to transcription start sites (TSSs) of gene within +/- 1,000,000 bp, using a weighing function. The weighing function quantifies the prior probability that a CRE will regulate a TSS at distance d and is a mixture of two Lorentzian distributions with length-scales 150 bp (corresponding to promoter regions) and 50Kb (corresponding to enhancer regions). This weighing function is used to weigh log-likelihood score per possible CRE-TSS interaction. The target gene score is a sum of the log-likelihood scores of all CREs associated with the gene weighted with the association probability. Then, the scores were used to predict over-represented canonical pathways in the TF's target genes.

Differential gene expression analysis

We ran DESeq2 on the TCGA RNA-seq read count data between ER+/HER2- ILCs (n=100) and ER+/HER2- IDCs (n=297), which include all available tumors for hormone receptor and histological subtypes. We used the Limma (v3.48) (57) package to calculate the log2 fold change of differentially expressed genes between ER+/HER2- ILCs (n=121) and ER+/HER2- IDCs (n=1,030) for the METABRIC dataset.

We calculated the cumulative distribution of expression changes for the target genes and background genes and ran the Kolmogorov-Smirnov (K-S) statistic to quantify the distance between empirical cumulative distribution function (eCDF) of target genes and cumulative distribution function (CDF) of background genes and determine its significance. We used all 16,537 genes as background genes after removing genes with low mean counts across samples.

Statistical analysis and data visualization

All statistical analyses were performed using R version 4.1.1 (R Foundation for Statistical Computing, Vienna, Austria) (58). Heatmaps were generated using the R package ComplexHeatmap v2.10.0 (59). Graphs were generated using the R package ggplot2 v3.3.5 (60). Genome track images were generated using the IGV (v2.11.1) (61). P-values in multiple comparisons were adjusted using the Benjamini–Hochberg (BH) method.

Author contributions

S.L. performed all the computational experiments, analyzed results, and helped to write the manuscript. H.U.O. conceived the project, advised on the analysis, and supervised the research and wrote the manuscript.

Data and code availability

This study includes no data or code deposited in external repositories.

Disclosure of Potential Conflict of Interests

The Authors declare that there is no conflict of interest.

Acknowledgments

The results published here are in whole or part based on the data generated by The Cancer Genome Atlas project established by the NCI and NHGRI (accession number: phs000178.v7p6). Information about TCGA and the investigators and institutions that constitute the TCGA research network can be found at http://cancergenome.nih.gov/. This work was supported by NIH award R00CA207871. ATAC-seq data analyses in this research were supported by the University of Pittsburgh Center for Research Computing and the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation grant OCI-1053575. Specifically, it used the Bridges2 system, which is supported by NSF award ACI-1445606 at the Pittsburgh Supercomputing Center. We thank Steffi Oesterreich, Adrian Lee, Jacqueline Bromberg, Jing Hong Wang, Micheal Gatza, Devin Dikec, and Kristi Rothermund for helpful discussions and Mikhail Pachkov and Erik van Nimwegen for their help in CREMA analysis.

Figure Legends

Figure 1. Differential chromatin accessibility between ER+/HER2- ILC and ER+/HER2- IDC.

(A) PCA of ATAC-seq signal in all peaks (n=196,546). All tumors of ER+ were clustered, but ILC and IDC tumors were slightly separated. The three outliers of ER+/HER2- IDCs and one outlier of ER+/HER2+ IDCs are annotated in the plot.

(B) Volcano plot of ATAC-seq peaks comparing ILCs (n=13) to IDCs (n=27). Significant peaks with differential chromatin accessibility are highlighted in red. The vertical dotted line indicates an absolute log2 fold change of 1.0 and the horizontal dotted line indicates an FDR-corrected p-value 0.05 criterion; the DA peaks enriched in ILCs (n=5,124) vs IDCs (n=6,638). FDR-corrected p-values were obtained using DESeq2.

(C) Hierarchical clustering of the 11,762 DA peaks. The significant DA peaks identified in Fig. 1B were aggregated to 11,762 peaks and represented as chromatin accessibility patterns in ILCs and IDCs. Colors represent log2-transformed peak count data and the z-score row was normalized.

(D) Pie charts show the percentage of DA ATAC-seq peaks (FDR < 0.05) at the promoter, intronic, intergenic, and exonic regions for ILCs vs. IDCs.

(E) Enrichment of TF-binding motifs for the subclusters of DA regions of ILCs and IDCs. The top 10 enriched motifs are shown.

(F) Enrichment of PANTHER pathways for subclusters of DA regions of ILCs and IDCs. In the bar plot, the grey line indicates the significance of the PANTHER pathways (hypergeometric test, adjusted p-value < 0.05). GREAT tool was used to identify the PANTHER pathways overrepresented in the DA peaks.

Figure 2. ILC and IDC tumors share a common chromatin state space.

(A) Scatter plot of differential expression (RNA-seq log2FC, x-axis) and differential accessibility (mean ATAC-seq log2FC over all peaks associated with a gene, y-axis) between IDCs and ILCs tumors. Significantly DA genes are highlighted with red or blue color.

(B) Differential accessibility and differential expression between ILCs and IDCs. Left: ATAC-seq signal log2 fold change for peaks of significantly DA genes; right: log2 fold change of RNA-seq gene expression (color for significantly decreased/increased individual peaks or genes; adjusted p-value < 0.05).

(C) The upper panel depicts genome browser tracks (GRCh38) showing chromatin accessibility at ERICH5 and FAM18A2 gene loci in ILCs and IDCs. The lower panel of genome browser tracks shows chromatin accessibility at FAM189A2 and SSPN gene loci which have DA peaks enriched

in ILCs. The dotted line boxes highlight the ATAC-seq peaks of DA between ILCs and IDCs. All the track lines have the same y-axis limits and the peak height is scaled over all samples.

Figure 3. ATAC-seq analysis identifies key TFs in ILC and IDC tumors.

(A) Inferred TF motif activity differences between ILCs and IDCs. The x-axis is the mean TF activity differences and the y-axis is –log10 (FDR-corrected p-values). Multiple hypothesis testing correction was done using the Benjamini–Hochberg procedure. The vertical dotted line indicates an absolute mean TF activity difference of 0.035 and the horizontal dotted line indicates the FDR-corrected p-value = 0.05 for significant TFs.
(B) EGR1, TEAD4, SOX2, RUNX3_BCL11A had inferred high TF activities in ILCs. **(C)** FOXA1, FOXA3, ATF4, and ZNF35 had inferred high TF activity in IDCs. The significance of the TF motif activity difference was determined by the Wilcoxon rank-sum test adjusted p-value.
(D) The Pearson correlation for TF activities in all ILC and IDC tumors.
(E) Immunohistochemical staining for TF protein expression. EGR1, BCL11A, TEAD4, and SOX2 had high activities in ILCs. Their protein expressions were high or medium in ILCs, but not detected in IDCs.
(F) FOXA1, FOXA3, ATF4, and ZNF35 had high activities in IDCs. Their protein expression was high or medium in IDCs, but not detected or low in ILCs. All tumors had high or medium ER expression, but HER2 expression was not detected or low.

Figure 4. Gene sets for candidate ILC- and IDC-specific TFs display coherent functional annotations and consistent expression changes in tumors.
(A) Targets of EGR1, TEAD4, SOX2, and RUNX3_BCL11A, ILC-specific candidate TFs, showed significant upregulation in ILC tumors relative to IDC tumors (p-value < 1e−3, Kolmogorov-Smirnov test) compared to background genes. The upper panel depicts the upregulation of TF target genes expression in TCGA RNA-seq data and the bottom panel depicts METABRIC microarray data. **(B)** Targets of FOXA1, HSF4, PBX3, and PITX1, IDC-specific candidate TFs showed significant upregulation of expression in TCGA and METABRIC data. The background genes were all genes identified in the gene expression dataset after removing low or non-expression genes. The yellow lines are empirical Cumulative Distribution Functions (eCDF) for the target gene log2 fold changes between ILCs and IDCs. The blue lines are CDFs for background gene log2 fold changes between ILCs and IDCs. The p-values are from the Kolmogorov-Smirnov (K-S) test between the target and the background distributions.

Tables

Table 1: Candidate TF regulators selected at 5% FDR for ILC. Functional annotations were determined from terms overrepresented from the canonical pathway gene sets associated with the candidate regulator. The p-values are from the Kolmogorov-Smirnov (K-S) test between the target and the background distributions for TCGA and METABRIC datasets.

TF symbol	Pathways associated with TF target genes (top 3)	p-value TCGA	p-value METABRIC	Relation to breast cancer	Reference
EGR1	Genes encoding, enzymes and their regulators involved in the remodeling of the ECM Genes encoding secreted soluble factors Genes encoding structural ECM glycoproteins	< 1.0E-16	< 1.0E-16	Overexpression induces E-cadherin transcription inhibition	(62)
HMGA1	E2F transcription factor network Genes encoding structural ECM glycoproteins Regulation of Ras family activation	< 3.4E-03	< 1.0E-16	Overexpression promotes metastasis	(63)
NFIA	Genes encoding secreted soluble factors Ephrin B reverse signaling ErbB receptor signaling network	< 1.0E-16	< 1.0E-16	Interacts to affect chromatin remodeling	(64)
NFIX_ NFIB	Genes encoding structural ECM glycoproteins Genes encoding secreted soluble factors Genes encoding enzymes and their regulators involved in the remodeling of the ECM	< 1.0E-16	< 1.0E-16	Upregulated in ER+ tumors and acts as an oncogene	(65)
RBPJ	Genes encoding enzymes and their regulators involved in the remodeling of the ECM Genes encoding structural components of basement membranes GMCSF-mediated signaling events	< 7.6E-01	< 3.2E-13	Regulates the NOTCH1 pathway via transcriptional repression resulting in recurrence of tumors	(66)
RUNX3_ BCL11A	E-cadherin signaling in keratinocytes TCR signaling in naïve CD4+ T cells IL12 signaling mediated by STAT4	< 1.0E-16	< 1.0E-16	RUNX3 inhibits cell proliferation by targeting ERα. BCL11A highly expressed in TNBC, drives metastasis	(46,67)
SOX2	Alpha6 beta4 integrin-ligand interactions Ephrin B reverse signaling Genes related to regulation of the actin cytoskeleton	< 2.2E-04	< 1.0E-16	Relates to cancer cell stemness, tumorigenicity, and transcription regulation of the *CCND1* gene	(68)
SOX4	Genes encoding structural ECM glycoproteins Signaling events mediated by the Hedgehog family Wnt/beta-catenin Pathway	< 8.4E-03	< 1.0E-16	Regulation of EMT-related genes, increased clonogenicity, angiogenesis, and tumor cell dissemination	(69)
SOX8	Signaling events mediated by the Hedgehog family Genes encoding collagen proteins Genes encoding structural ECM glycoproteins	< 7.9E-07	< 1.0E-16	Relates to cancer cell stemness in TNBC cells	(70)
STAT6	E-cadherin signaling in keratinocytes	< 2.4E-07	< 1.0E-16		(71)

	Netrin-mediated signaling events Genes encoding structural components of basement membranes			Mediates Interleukin-4 (IL-4) growth inhibition, induction of apoptosis	
TEAD3_ TEAD1	Ensemble of genes encoding ECM-associated proteins including ECM-affiliated proteins, ECM regulators and secreted factors Genes encoding enzymes and their regulators involved in the remodeling of the ECM AMB2 Integrin signaling	< 1.0E-16	< 1.0E-16	Bind with HIPPO pathway co-activators (YAP, TAZ) creating oncogenic transformation and tumorigenesis	(29)
TEAD4	Notch-mediated HES/HEY network PDGFR-beta signaling pathway Genes encoding structural ECM glycoproteins	< 8.2E-07	< 1.0E-16	Overexpressed in BC stem cells and correlate with poor survival	(29)
TP63	Genes encoding secreted soluble factors Genes encoding structural ECM glycoproteins Genes encoding enzymes and their regulators involved in the remodeling of the ECM	< 1.0E-16	< 1.0E-16	Enhances endocrine treatment responses in ER+ tumors	(72)
UBP1	RhoA signaling pathway Genes encoding collagen proteins Genes encoding structural ECM glycoproteins	< 1.0E-16	< 1.0E-16	Overexpressed in breast invasive cancer	(73)
ZFHX3	Validated targets of C-MYC transcriptional repression Genes encoding proteins affiliated structurally or functionally to ECM E-cadherin signaling in keratinocytes	< 1.4E-03	< 1.0E-16	Promotes proliferation and tumorigenesis in ER+ cells by increasing stemness of cancer cells	(74)

Table 2: Candidate TF regulators selected at 5% FDR for IDC. Functional annotations were determined from terms overrepresented from the canonical pathway gene sets associated with the candidate regulator. The p-values are from the Kolmogorov-Smirnov (K-S) test between the target and the background distributions for TCGA and METABRIC datasets.

TF symbol	Pathways associated with TF target genes (top 3)	p-value TCGA	p-value METABRIC	Relation to breast cancer	Reference
ARNT	HIF-1-alpha transcription factor network IL2 signaling events mediated by PI3K EPHA2 forward signaling	< 2.1E-04	< 9.1E-10	Downregulation promotes cancer cell migration and invasion	(75)
ATF4	PI3K Pathway Osteopontin-mediated events Validated transcriptional targets of deltaNp63 isoforms	< 1.1E-01	< 5.9E-10	Critical regulator of the unfolded protein response (UPR) pathway and is implicated in tumorigenesis	(76,77)
FOXA1	FOXA1 transcription factor network FOXA2 and FOXA3 transcription factor networks Regulation of CDC42 activity	< 4.7E-10	< 5.9E-03	Inhibits cell growth via E-cadherin and suppression of ER pathway activation	(47)
FOXA3_ FOXC2	FOXA2 and FOXA3 transcription factor networks Signaling events mediated by HDAC Class II Regulation of CDC42 activity	< 1.0E-16	< 9.8E-04	Induces EMT and cancer cell stemness	(78)
FOXD3_ FOXI1_ FOXF1	FOXA1 transcription factor network JNK MAPK Pathway Hedgehog signaling events mediated by Gli proteins	< 2.2E-16	< 5.7E-06	Down-regulation associated with lymph node metastasis in IDC. Potential tumor suppressor affecting the cell cycle. Overexpression associated with EMT	(79-82)
FOXJ3	FOXA1 transcription factor network Regulation of CDC42 activity BMP receptor signaling	< 1.0E-16	< 6.4E-05	Protected motif in chromatin landscape in drug resistant cancer cells	(83)
FOXK1_ FOXP3	FOXA1 transcription factor network Genes encoding enzymes and their regulators involved in the remodeling of the extracellular matrix Genes encoding proteins affiliated structurally or functionally to extracellular matrix proteins	< 1.0E-16	< 1.8E-05	Promotes cell proliferation, migration, EMT, and invasion	(84)
FOXL1	BMP receptor signaling FOXA1 transcription factor network ErbB4 signaling events	< 1.0E-16	< 2.1E-11	Functions as a tumor suppressor to inhibit cell proliferation and invasion	(85)
HSF4	Calcineurin-regulated NFAT-dependent transcription in lymphocytes Genes encoding structural ECM glycoproteins Plasma membrane estrogen receptor signaling	< 1.0E-07	< 1.0E-16	Promotes HIF-1α expression and tumor progression	(86)
PBX3	FOXA1 transcription factor network Signaling events mediated by Hepatocyte Growth Factor Receptor (c-Met) Regulation of Androgen receptor activity	< 1.7E-06	< 1.0E-16	Attenuates response to Letrozole by potentiating breast cancer cell survival and anchorage-independent growth	(87)

PITX1	Coregulation of Androgen receptor activity LKB1 signaling events Validated targets of C-MYC transcriptional activation	< 1.0E-16	< 1.0E-16	Tumor suppressor that is regulated by ERα	(88)
PITX2	ErbB4 signaling events Arf6 trafficking events RXR and RAR heterodimerization with other nuclear receptors	< 2.9E-04	< 1.0E-16	Hypermethylation of PITX2 promoter reduced expression and induced cancer cell progression.	(88)
SPDEF	ErbB2/ErbB3 signaling events Nectin adhesion pathway Hedgehog signaling events mediated by Gli proteins	< 1.3E-01	< 1.6E-13	Expression is enriched in luminal tumors and promotes differentiation and survival of ER+ cells	(32)
ZNF35	Notch signaling pathway E-cadherin signaling in the nascent adherens junction IL2 signaling events mediated by PI3K	< 1.4E-02	< 5.3E-12	Overexpression indicates poor prognosis and lymph node metastasis.	(89)

References

1. Clark, G.M., Osborne, C.K. and McGuire, W.L. (1984) Correlations between estrogen receptor, progesterone receptor, and patient characteristics in human breast cancer. *Journal of clinical oncology*, **2**, 1102-1109.
2. Sreekumar, S., Levine, K.M., Sikora, M.J., Chen, J., Tasdemir, N., Carter, D., Dabbs, D.J., Meier, C., Basudan, A. and Boone, D. (2020) Differential regulation and targeting of estrogen receptor α turnover in invasive lobular breast carcinoma. *Endocrinology*, **161**, bqaa109.
3. Rakha, E.A., Reis-Filho, J.S., Baehner, F., Dabbs, D.J., Decker, T., Eusebi, V., Fox, S.B., Ichihara, S., Jacquemier, J. and Lakhani, S.R. (2010) Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Research*, **12**, 1-12.
4. Li, C.I., Anderson, B.O., Daling, J.R. and Moe, R.E. (2003) Trends in incidence rates of invasive lobular and ductal breast carcinoma. *Jama*, **289**, 1421-1424.
5. Sflomos, G., Schipper, K., Koorman, T., Fitzpatrick, A., Oesterreich, S., Lee, A.V., Jonkers, J., Brunton, V.G., Christgen, M. and Isacke, C. (2021) Atlas of Lobular Breast Cancer Models: Challenges and Strategic Directions. *Cancers*, **13**, 5396.
6. Tubiana-Hulin, M., Stevens, D., Lasry, S., Guinebretiere, J., Bouita, L., Cohen-Solal, C., Cherel, P. and Rouesse, J. (2006) Response to neoadjuvant chemotherapy in lobular and ductal breast carcinomas: a retrospective study on 860 patients from one institution. *Annals of oncology*, **17**, 1228-1233.
7. Cocquyt, V.F., Blondeel, P., Depypere, H., Praet, M., Schelfhout, V., Silva, O.E., Hurley, J., Serreyn, R., Daems, K. and Van Belle, S. (2003) Different responses to preoperative chemotherapy for invasive lobular and invasive ductal breast carcinoma. *European Journal of Surgical Oncology (EJSO)*, **29**, 361-367.
8. Mouabbi, J.A., Hassan, A., Lim, B., Hortobagyi, G.N., Tripathy, D. and Layman, R.M. (2022) Invasive lobular carcinoma: an understudied emergent subtype of breast cancer. *Breast Cancer Research and Treatment*, 1-12.
9. Adachi, Y., Ishiguro, J., Kotani, H., Hisada, T., Ichikawa, M., Gondo, N., Yoshimura, A., Kondo, N., Hattori, M. and Sawaki, M. (2016) Comparison of clinical outcomes between luminal invasive ductal carcinoma and luminal invasive lobular carcinoma. *BMC cancer*, **16**, 1-9.
10. Cristofanilli, M., Gonzalez-Angulo, A., Sneige, N., Kau, S.-W., Broglio, K., Theriault, R.L., Valero, V., Buzdar, A.U., Kuerer, H. and Buccholz, T.A. (2005) Invasive lobular carcinoma classic type: response to primary chemotherapy and survival outcomes. *Journal of Clinical Oncology*, **23**, 41-48.
11. Chamalidou, C., Fohlin, H., Albertsson, P., Arnesson, L.-G., Einbeigi, Z., Holmberg, E., Nordenskjöld, A., Nordenskjöld, B., Karlsson, P. and Linderholm, B. (2021) Survival patterns of invasive lobular and invasive ductal breast cancer in a large population-based cohort with two decades of follow up. *The Breast*, **59**, 294-300.
12. Delpech, Y., Coutant, C., Hsu, L., Barranger, E., Iwamoto, T., Barcenas, C., Hortobagyi, G., Rouzier, R., Esteva, F. and Pusztai, L. (2013) Clinical benefit from neoadjuvant chemotherapy in oestrogen receptor-positive invasive ductal and lobular carcinomas. *British journal of cancer*, **108**, 285-291.
13. Biglia, N., Maggiorotto, F., Liberale, V., Bounous, V., Sgro, L., Pecchio, S., D'Alonzo, M. and Ponzone, R. (2013) Clinical-pathologic features, long term-outcome and surgical treatment in a large series of patients with invasive lobular carcinoma (ILC) and invasive ductal carcinoma (IDC). *European Journal of Surgical Oncology (EJSO)*, **39**, 455-460.
14. Duraker, N., Hot, S., Akan, A. and Nayır, P.Ö. (2020) A comparison of the clinicopathological features, metastasis sites and survival outcomes of invasive lobular,

invasive ductal and mixed invasive ductal and lobular breast carcinoma. *European Journal of Breast Health*, **16**, 22.

15. Du, T., Zhu, L., Levine, K.M., Tasdemir, N., Lee, A.V., Vignali, D.A., Houten, B.V., Tseng, G.C. and Oesterreich, S. (2018) Invasive lobular and ductal breast carcinoma differ in immune response, protein translation efficiency and metabolism. *Scientific reports*, **8**, 1-11.
16. Ciriello, G., Gatza, M.L., Beck, A.H., Wilkerson, M.D., Rhie, S.K., Pastore, A., Zhang, H., McLellan, M., Yau, C. and Kandoth, C. (2015) Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, **163**, 506-519.
17. Teo, K., Gómez-Cuadrado, L., Tenhagen, M., Byron, A., Rätze, M., van Amersfoort, M., Renes, J., Strengman, E., Mandoli, A. and Singh, A.A. (2018) E-cadherin loss induces targetable autocrine activation of growth factor signalling in lobular breast cancer. *Scientific reports*, **8**, 1-14.
18. Chen, F., Ding, K., Priedigkeit, N., Elangovan, A., Levine, K.M., Carleton, N., Savariau, L., Atkinson, J.M., Oesterreich, S. and Lee, A.V. (2021) Single-Cell Transcriptomic Heterogeneity in Invasive Ductal and Lobular Breast Cancer Cells. *Cancer Res*, **81**, 268-281.
19. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W. *et al.* (2018) The chromatin accessibility landscape of primary human cancers. *Science*, **362**.
20. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, **10**, 1213-1218.
21. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K. and Cho, S.W. (2018) The chromatin accessibility landscape of primary human cancers. *Science*, **362**, eaav1898.
22. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*, **38**, 576-589.
23. Chen, Y., Shi, L., Zhang, L., Li, R., Liang, J., Yu, W., Sun, L., Yang, X., Wang, Y. and Zhang, Y. (2008) The molecular mechanism governing the oncogenic potential of SOX2 in breast cancer. *Journal of Biological Chemistry*, **283**, 17969-17978.
24. Mehta, G.A., Khanna, P. and Gatza, M.L. (2019) Emerging role of SOX proteins in breast Cancer development and maintenance. *Journal of mammary gland biology and neoplasia*, **24**, 213-230.
25. Xu, Y., Dong, X., Qi, P., Ye, Y., Shen, W., Leng, L., Wang, L., Li, X., Luo, X. and Chen, Y. (2017) Sox2 communicates with tregs through CCL1 to promote the stemness property of breast cancer cells. *Stem Cells*, **35**, 2351-2365.
26. Dey, A., Kundu, M., Das, S., Jena, B.C. and Mandal, M. (2022) Understanding the function and regulation of Sox2 for its therapeutic potential in breast cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 188692.
27. Mehta, G.A., Parker, J.S., Silva, G.O., Hoadley, K.A., Perou, C.M. and Gatza, M.L. (2017) Amplification of SOX4 promotes PI3K/Akt signaling in human breast cancer. *Breast cancer research and treatment*, **162**, 439-450.
28. Moreno, C.S. (2020), *Seminars in cancer biology*. Elsevier, Vol. 67, pp. 57-64.
29. Wu, Y., Li, M., Lin, J. and Hu, C. (2021) Hippo/TEAD4 signaling pathway as a potential target for the treatment of breast cancer. *Oncology Letters*, **21**, 1-6.

30. He, L., Yuan, L., Sun, Y., Wang, P., Zhang, H., Feng, X., Wang, Z., Zhang, W., Yang, C. and Zeng, Y.A. (2019) Glucocorticoid receptor signaling activates TEAD4 to promote breast cancer progression. *Cancer research*, **79**, 4399-4411.
31. Koker, M.M. and Kleer, C.G. (2004) p63 expression in breast cancer: a highly sensitive and specific marker of metaplastic carcinoma. *The American journal of surgical pathology*, **28**, 1506-1512.
32. Ye, T., Feng, J., Wan, X., Xie, D. and Liu, J. (2020) Double agent: SPDEF gene with both oncogenic and tumor-suppressor functions in breast cancer. *Cancer management and research*, **12**, 3891.
33. Turner, D.P., Findlay, V.J., Kirven, A.D., Moussa, O. and Watson, D.K. (2008) Global gene expression analysis identifies PDEF transcriptional networks regulating cell migration during cancer progression. *Molecular biology of the cell*, **19**, 3745-3757.
34. Sood, A.K., Saxena, R., Groth, J., Desouki, M.M., Cheewakriangkrai, C., Rodabaugh, K.J., Kasyapa, C.S. and Geradts, J. (2007) Expression characteristics of prostate-derived Ets factor support a role in breast and prostate cancer progression. *Human pathology*, **38**, 1628-1638.
35. Carroll, J.S., Liu, X.S., Brodsky, A.S., Li, W., Meyer, C.A., Szary, A.J., Eeckhoute, J., Shao, W., Hestermann, E.V. and Geistlinger, T.R. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, **122**, 33-43.
36. Wolf, I., Bose, S., Williamson, E.A., Miller, C.W., Karlan, B.Y. and Koeffler, H.P. (2007) FOXA1: Growth inhibitor and a favorable prognostic factor in human breast cancer. *International journal of cancer*, **120**, 1013-1022.
37. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M. and Bejerano, G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, **28**, 495-501.
38. Grant, S. (2012) FAM83A and FAM83B: candidate oncogenes and TKI resistance mediators. *The Journal of clinical investigation*, **122**, 3048-3051.
39. Liu, C., Jiang, Y. and Han, B. (2020) miR-613 suppresses chemoresistance and stemness in triple-negative breast cancer by targeting FAM83A. *Cancer Management and Research*, **12**, 12623.
40. Xu, Y.-H., Deng, J.-L., Wang, L.-P., Zhang, H.-B., Tang, L., Huang, Y., Tang, J., Wang, S.-M. and Wang, G. (2020) Identification of candidate genes associated with breast cancer prognosis. *DNA and Cell Biology*, **39**, 1205-1227.
41. Tsunoda, T., Riku, M., Yamada, N., Tsuchiya, H., Tomita, T., Suzuki, M., Kizuki, M., Inoko, A., Ito, H. and Murotani, K. (2021) ENTREP/FAM189A2 encodes a new ITCH ubiquitin ligase activator that is downregulated in breast cancer. *EMBO reports*, e51182.
42. Chuan, T., Li, T. and Yi, C. (2020) Identification of CXCR4 and CXCL10 as potential predictive biomarkers in triple negative breast cancer (TNBC). *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, **26**, e918281-918281.
43. Weigelt, B., Geyer, F.C., Natrajan, R., Lopez-Garcia, M.A., Ahmad, A.S., Savage, K., Kreike, B. and Reis-Filho, J.S. (2010) The molecular underpinning of lobular histological growth pattern: a genome-wide transcriptomic analysis of invasive lobular carcinomas and grade-and molecular subtype-matched invasive ductal carcinomas of no special type. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, **220**, 45-57.
44. Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K. and Hober, S. (2010) Towards a knowledge-based human protein atlas. *Nature biotechnology*, **28**, 1248-1250.

45. Marcotte, R., Sayad, A., Brown, K.R., Sanchez-Garcia, F., Reimand, J., Haider, M., Virtanen, C., Bradner, J.E., Bader, G.D. and Mills, G.B. (2016) Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance. *Cell*, **164**, 293-309.
46. Huang, B., Qu, Z., Ong, C.W., Tsang, Y., Xiao, G., Shapiro, D., Salto-Tellez, M., Ito, K., Ito, Y. and Chen, L.-F. (2012) RUNX3 acts as a tumor suppressor in breast cancer by targeting estrogen receptor α. *Oncogene*, **31**, 527-534.
47. BenAyed-Guerfali, D., Dabbèche-Bouricha, E., Ayadi, W., Trifa, F., Charfi, S., Khabir, A., Sellami-Boudawara, T. and Mokdad-Gargouri, R. (2019) Association of FOXA1 and EMT markers (Twist1 and E-cadherin) in breast cancer. *Molecular biology reports*, **46**, 3247-3255.
48. Habashy, H.O., Powe, D.G., Rakha, E.A., Ball, G., Paish, C., Gee, J., Nicholson, R.I. and Ellis, I.O. (2008) Forkhead-box A1 (FOXA1) expression in breast cancer and its prognostic significance. *European journal of cancer*, **44**, 1541-1551.
49. Buchwalter, G., Hickey, M.M., Cromer, A., Selfors, L.M., Gunawardane, R.N., Frishman, J., Jeselsohn, R., Lim, E., Chi, D. and Fu, X. (2013) PDEF promotes luminal differentiation and acts as a survival factor for ER-positive breast cancer cells. *Cancer cell*, **23**, 753-767.
50. Heath, A.P., Ferretti, V., Agrawal, S., An, M., Angelakos, J.C., Arya, R., Bajari, R., Baqar, B., Barnowski, J.H. and Burt, J. (2021) The NCI genomic data commons. *Nature genetics*, **53**, 257-262.
51. Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S. and Yuan, Y. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346-352.
52. Pontén, F., Jirström, K. and Uhlen, M. (2008) The Human Protein Atlas—a tool for pathology. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, **216**, 387-393.
53. Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T. and Carey, V.J. (2013) Software for computing and annotating genomic ranges. *PLoS computational biology*, **9**, e1003118.
54. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, **15**, 1-21.
55. Yu, G., Wang, L.-G. and He, Q.-Y. (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382-2383.
56. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W. and Zhan, L. (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, **2**, 100141.
57. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, **43**, e47-e47.
58. Team, R.C. (2013) R: A language and environment for statistical computing.
59. Gu, Z., Eils, R. and Schlesner, M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847-2849.
60. Wickham, H. (2016) *ggplot2: elegant graphics for data analysis*. springer.
61. Robinson, J.T., Thorvaldsdóttir, H., Wenger, A.M., Zehir, A. and Mesirov, J.P. (2017) Variant review with the integrative genomics viewer. *Cancer research*, **77**, e31-e34.
62. Wang, B., Guo, H., Yu, H., Chen, Y., Xu, H. and Zhao, G. (2021) The role of the transcription factor EGR1 in cancer. *Frontiers in Oncology*, **11**, 775.
63. Wang, Y., Hu, L., Zheng, Y. and Guo, L. (2019) HMGA1 in cancer: Cancer classification by location. *Journal of Cellular and Molecular Medicine*, **23**, 2293-2302.

64. Fane, M., Harris, L., Smith, A.G. and Piper, M. (2017) Nuclear factor one transcription factors as epigenetic regulators in cancer. *International Journal of Cancer*, **140**, 2634-2641.
65. Chen, H., Yu, C., Shen, L., Wu, Y., Wu, D., Wang, Z., Song, G., Chen, L. and Hong, Y. (2020) NFIB functions as an oncogene in estrogen receptor-positive breast cancer and is regulated by miR-205-5p. *Pathology-Research and Practice*, **216**, 153236.
66. Kulic, I., Robertson, G., Chang, L., Baker, J.H., Lockwood, W.W., Mok, W., Fuller, M., Fournier, M., Wong, N. and Chou, V. (2015) Loss of the Notch effector RBPJ promotes tumorigenesis. *Journal of Experimental Medicine*, **212**, 37-52.
67. Seachrist, D.D., Hannigan, M.M., Ingles, N.N., Webb, B.M., Weber-Bonk, K.L., Yu, P., Bebek, G., Singh, S., Sizemore, S.T. and Varadan, V. (2020) The transcriptional repressor BCL11A promotes breast cancer metastasis. *Journal of Biological Chemistry*, **295**, 11707-11719.
68. Schaefer, T. and Lengerke, C. (2020) SOX2 protein biochemistry in stemness, reprogramming, and cancer: the PI3K/AKT/SOX2 axis and beyond. *Oncogene*, **39**, 278-292.
69. Zhang, J., Xiao, C., Feng, Z., Gong, Y., Sun, B., Li, Z., Lu, Y., Fei, X., Wu, W. and Sun, X. (2020) SOX4 promotes the growth and metastasis of breast cancer. *Cancer cell international*, **20**, 1-11.
70. Tang, H., Chen, B., Liu, P., Xie, X., He, R., Zhang, L., Huang, X., Xiao, X. and Xie, X. (2019) SOX8 acts as a prognostic factor and mediator to regulate the progression of triple-negative breast cancer. *Carcinogenesis*, **40**, 1278-1287.
71. Gooch, J.L., Christy, B. and Yee, D. (2002) STAT6 mediates interleukin-4 growth inhibition in human breast cancer cells. *Neoplasia*, **4**, 324-331.
72. Hanker, L., Karn, T., Ruckhäberle, E., Gaetje, R., Solbach, C., Schmidt, M., Engels, K., Holtrich, U., Kaufmann, M. and Rody, A. (2010) Clinical relevance of the putative stem cell marker p63 in breast cancer. *Breast cancer research and treatment*, **122**, 765-775.
73. Zhao, Y., Kaushik, N., Kang, J.-H., Kaushik, N.K., Son, S.H., Uddin, N., Kim, M.-J., Kim, C.G. and Lee, S.-J. (2020) A feedback loop comprising EGF/TGFα sustains TFCP2-mediated breast cancer progression. *Cancer research*, **80**, 2217-2229.
74. Hu, Q., Zhang, B., Chen, R., Fu, C., Fu, X., Li, J., Fu, L., Zhang, Z. and Dong, J.-T. (2019) ZFHX3 is indispensable for ERβ to inhibit cell proliferation via MYC downregulation in prostate cancer cells. *Oncogenesis*, **8**, 1-15.
75. Hanieh, H., Mohafez, O., Hairul-Islam, V.I., Alzahrani, A., Bani Ismail, M. and Thirugnanasambantham, K. (2016) Novel aryl hydrocarbon receptor agonist suppresses migration and invasion of breast cancer cells. *PloS one*, **11**, e0167650.
76. Zeng, P., Sun, S., Li, R., Xiao, Z.-X. and Chen, H. (2019) HER2 upregulates ATF4 to promote cell migration via activation of ZEB1 and downregulation of E-cadherin. *International Journal of Molecular Sciences*, **20**, 2223.
77. Nagelkerke, A., Bussink, J., Mujcic, H., Wouters, B.G., Lehmann, S., Sweep, F.C. and Span, P.N. (2013) Hypoxia stimulates migration of breast cancer cells via the PERK/ATF4/LAMP3-arm of the unfolded protein response. *Breast Cancer Research*, **15**, 1-13.
78. Hollier, B.G., Tinnirello, A.A., Werden, S.J., Evans, K.W., Taube, J.H., Sarkar, T.R., Sphyris, N., Shariati, M., Kumar, S.V. and Battula, V.L. (2013) FOXC2 expression links epithelial–mesenchymal transition and stem cell properties in breast cancer. *Cancer research*, **73**, 1981-1992.
79. Zhao, H., Chen, D., Wang, J., Yin, Y., Gao, Q. and Zhang, Y. (2014) Downregulation of the transcription factor, FoxD3, is associated with lymph node metastases in invasive ductal carcinomas of the breast. *International Journal of Clinical and Experimental Pathology*, **7**, 670.

80. Onodera, Y., Takagi, K., Neoi, Y., Sato, A., Yamaguchi, M., Miki, Y., Ebata, A., Miyashita, M., Sasano, H. and Suzuki, T. (2021) Forkhead Box I1 in Breast Carcinoma as a Potent Prognostic Factor. *Acta histochemica et cytochemica*, 21-00034.
81. Lo, P.-K., Lee, J.S., Liang, X., Han, L., Mori, T., Fackler, M.J., Sadik, H., Argani, P., Pandita, T.K. and Sukumar, S. (2010) Epigenetic inactivation of the potential tumor suppressor gene FOXF1 in breast cancer. *Cancer research*, **70**, 6047-6058.
82. Katoh, M., Igarashi, M., Fukuda, H., Nakagama, H. and Katoh, M. (2013) Cancer genetics and genomics of human FOX family genes. *Cancer letters*, **328**, 198-206.
83. Murad, R., Avanes, A., Ma, X., Geng, S., Mortazavi, A. and Momand, J. (2021) Transcriptome and chromatin landscape changes associated with trastuzumab resistance in HER2+ breast cancer cells. *Gene*, **799**, 145808.
84. Gao, F. and Tian, J. (2020) FOXK1, regulated by miR-365-3p, promotes cell growth and EMT indicates unfavorable prognosis in breast cancer. *OncoTargets and therapy*, **13**, 623.
85. Zhong, J., Wang, H., Yu, J., Zhang, J. and Wang, H. (2017) Overexpression of Forkhead box L1 (FOXL1) inhibits the proliferation and invasion of breast Cancer cells. *Oncology research*, **25**, 959.
86. Chen, R., Liliental, J., Kowalski, P., Lu, Q. and Cohen, S. (2011) Regulation of transcription of hypoxia-inducible factor-1α (HIF-1α) by heat shock factors HSF2 and HSF4. *Oncogene*, **30**, 2570-2580.
87. Pang, Z.-y., Wei, Y.-t., Shang, M.-y., Li, S., Li, Y., Jin, Q.-x., Liao, Z.-x., Cui, M.-k., Liu, X.-y. and Zhang, Q. (2021) Leptin-elicited PBX3 confers letrozole resistance in breast cancer. *Endocrine-related cancer*, **28**, 173-189.
88. Stender, J.D., Stossi, F., Funk, C.C., Charn, T.H., Barnett, D.H. and Katzenellenbogen, B.S. (2011) The estrogen-regulated transcription factor PITX1 coordinates gene-specific regulation by estrogen receptor-alpha in breast cancer cells. *Molecular Endocrinology*, **25**, 1699-1709.
89. Yan, D., Shen, M., Du, Z., Cao, J., Tian, Y., Zeng, P. and Tang, Z. (2021) Developing ZNF Gene Signatures Predicting Radiosensitivity of Patients with Breast Cancer. *Journal of oncology*, **2021**.

Figure 1

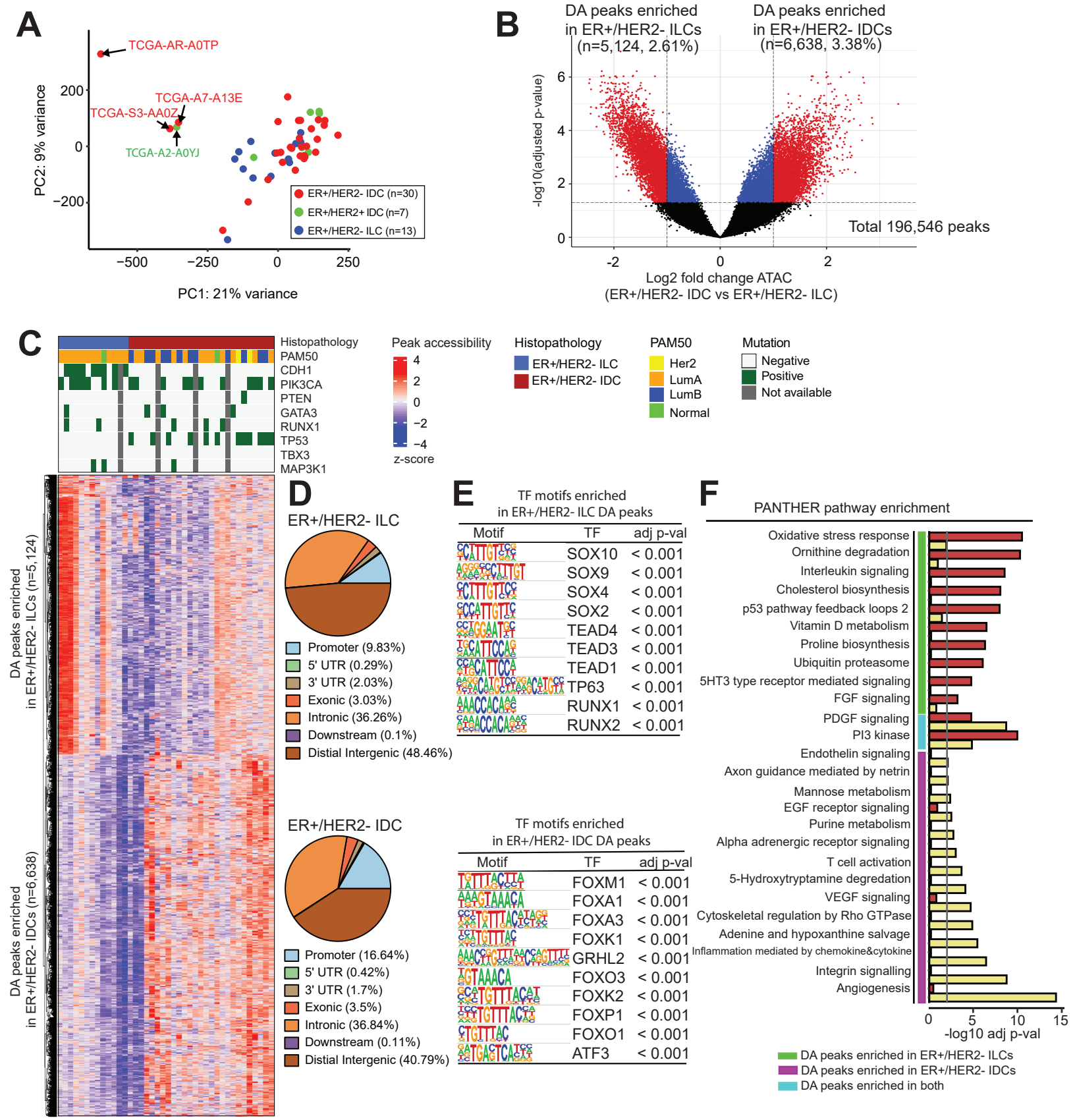

A
TCGA-AR-A0TP
TCGA-A7-A13E
TCGA-S3-AA0Z
TCGA-A2-A0YJ
PC2: 9% variance
PC1: 21% variance
ER+/HER2- IDC (n=30)
ER+/HER2+ IDC (n=7)
ER+/HER2- ILC (n=13)
B
DA peaks enriched in ER+/HER2- ILCs (n=5,124, 2.61%)
DA peaks enriched in ER+/HER2- IDCs (n=6,638, 3.38%)
-log10(adjusted p-value)
Total 196,546 peaks
Log2 fold change ATAC (ER+/HER2- IDC vs ER+/HER2- ILC)
C
Histopathology
PAM50
CDH1
PIK3CA
PTEN
GATA3
RUNX1
TP53
TBX3
MAP3K1
Peak accessibility
z-score
Histopathology
ER+/HER2- ILC
ER+/HER2- IDC
PAM50
Her2
LumA
LumB
Normal
Mutation
Negative
Positive
Not available
DA peaks enriched in ER+/HER2- ILCs (n=5,124)
DA peaks enriched in ER+/HER2- IDCs (n=6,638)
D
ER+/HER2- ILC
Promoter (9.83%)
5' UTR (0.29%)
3' UTR (2.03%)
Exonic (3.03%)
Intronic (36.26%)
Downstream (0.1%)
Distial Intergenic (48.46%)
ER+/HER2- IDC
Promoter (16.64%)
5' UTR (0.42%)
3' UTR (1.7%)
Exonic (3.5%)
Intronic (36.84%)
Downstream (0.11%)
Distial Intergenic (40.79%)
E
TF motifs enriched in ER+/HER2- ILC DA peaks
Motif TF adj p-val
SOX10 < 0.001
SOX9 < 0.001
SOX4 < 0.001
SOX2 < 0.001
TEAD4 < 0.001
TEAD3 < 0.001
TEAD1 < 0.001
TP63 < 0.001
RUNX1 < 0.001
RUNX2 < 0.001
TF motifs enriched in ER+/HER2- IDC DA peaks
Motif TF adj p-val
FOXM1 < 0.001
FOXA1 < 0.001
FOXA3 < 0.001
FOXK1 < 0.001
GRHL2 < 0.001
FOXO3 < 0.001
FOXK2 < 0.001
FOXP1 < 0.001
FOXO1 < 0.001
ATF3 < 0.001
F
PANTHER pathway enrichment
Oxidative stress response
Ornithine degradation
Interleukin signaling
Cholesterol biosynthesis
p53 pathway feedback loops 2
Vitamin D metabolism
Proline biosynthesis
Ubiquitin proteasome
5HT3 type receptor mediated signaling
FGF signaling
PDGF signaling
PI3 kinase
Endothelin signaling
Axon guidance mediated by netrin
Mannose metabolism
EGF receptor signaling
Purine metabolism
Alpha adrenergic receptor signaling
T cell activation
5-Hydroxytryptamine degredation
VEGF signaling
Cytoskeletal regulation by Rho GTPase
Adenine and hypoxanthine salvage
Inflammation mediated by chemokine&cytokine
Integrin signalling
Angiogenesis
-log10 adj p-val
DA peaks enriched in ER+/HER2- ILCs
DA peaks enriched in ER+/HER2- IDCs
DA peaks enriched in both

Figure 2

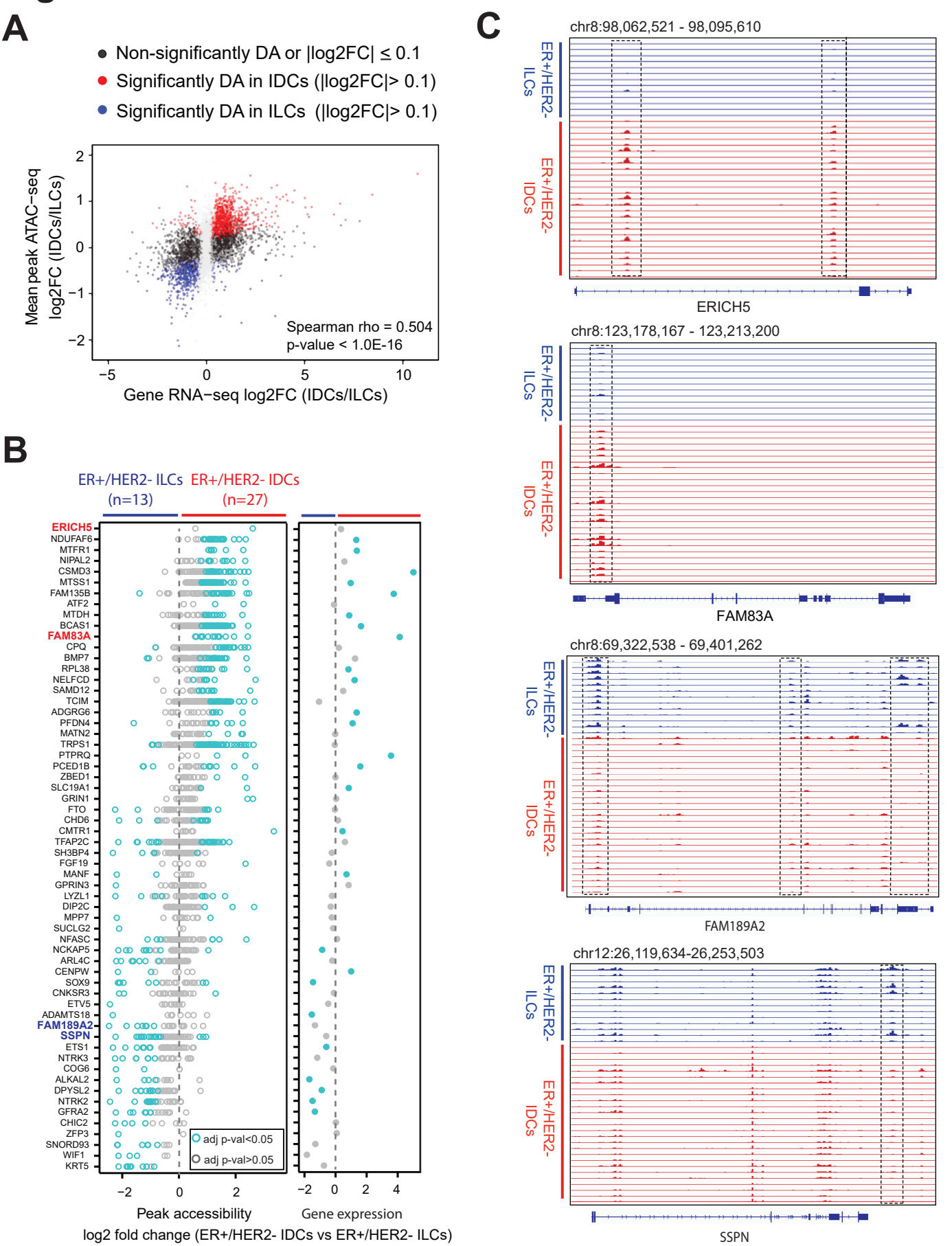

A
Non-significantly DA or |log2FC| ≤ 0.1
Significantly DA in IDCs (|log2FC|> 0.1)
Significantly DA in ILCs (|log2FC|> 0.1)
Mean peak ATAC−seq log2FC (IDCs/ILCs)
Gene RNA−seq log2FC (IDCs/ILCs)
Spearman rho = 0.504
p-value < 1.0E-16
B
ER+/HER2- ILCs (n=13)
ER+/HER2- IDCs (n=27)
ERICH5
FAM83A
FAM189A2
SSPN
adj p-val<0.05
adj p-val>0.05
Peak accessibility
Gene expression
log2 fold change (ER+/HER2- IDCs vs ER+/HER2- ILCs)
C
chr8:98,062,521 - 98,095,610
ERICH5
chr8:123,178,167 - 123,213,200
FAM83A
chr8:69,322,538 - 69,401,262
FAM189A2
chr12:26,119,634-26,253,503
SSPN
ER+/HER2- ILCs
ER+/HER2- IDCs

Figure 3

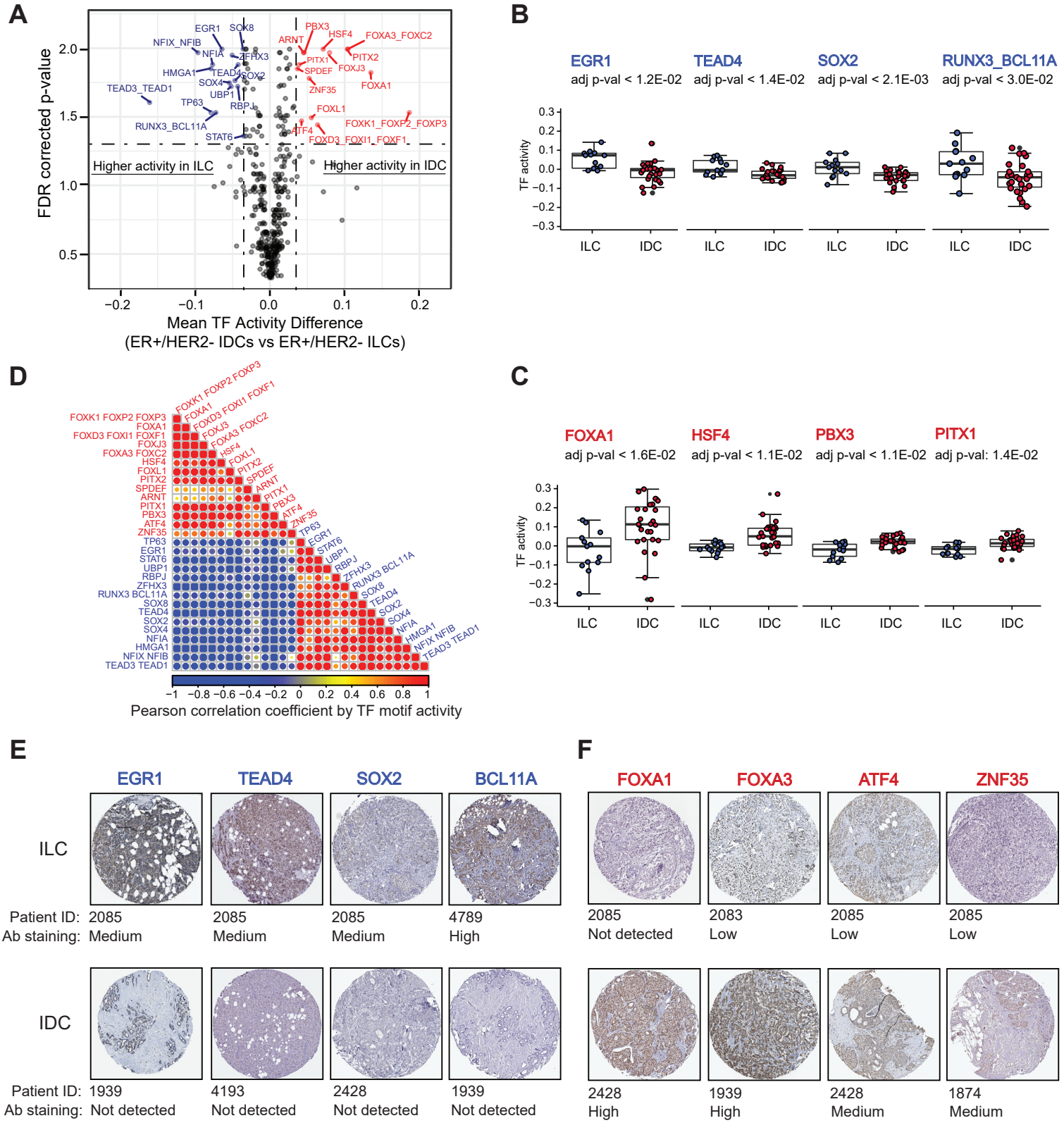
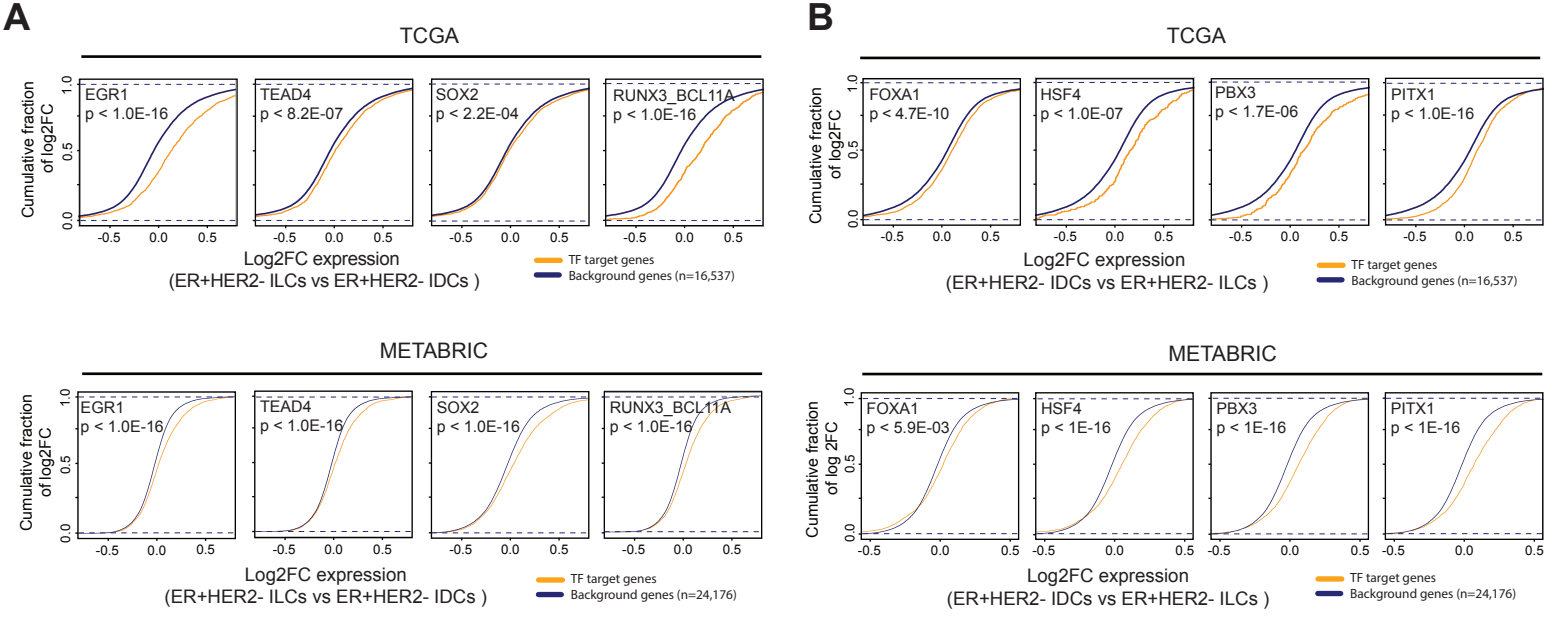

A
FDR corrected p-value
Mean TF Activity Difference
(ER+/HER2- IDCs vs ER+/HER2- ILCs)
Higher activity in ILC
Higher activity in IDC
EGR1
SOX8
NFIX_NFIB
NFIA
ZFHX3
HMGA1
TEAD4
SOX2
SOX4
TEAD3_TEAD1
UBP1
TP63
RBPJ
RUNX3_BCL11A
STAT6
PBX3
HSF4
FOXA3_FOXC2
ARNT
PITX1
PITX2
SPDEF
FOXJ3
ZNF35
FOXA1
FOXL1
FOXK1_FOXP2_FOXP3
ATF4
FOXD3_FOXI1_FOXF1
B
EGR1
adj p-val < 1.2E-02
TEAD4
adj p-val < 1.4E-02
SOX2
adj p-val < 2.1E-03
RUNX3_BCL11A
adj p-val < 3.0E-02
TF activity
ILC
IDC
D
Pearson correlation coefficient by TF motif activity
C
FOXA1
adj p-val < 1.6E-02
HSF4
adj p-val < 1.1E-02
PBX3
adj p-val < 1.1E-02
PITX1
adj p-val: 1.4E-02
TF activity
ILC
IDC
E
EGR1
TEAD4
SOX2
BCL11A
ILC
Patient ID: 2085 2085 2085 4789
Ab staining: Medium Medium Medium High
IDC
Patient ID: 1939 4193 2428 1939
Ab staining: Not detected Not detected Not detected Not detected
F
FOXA1
FOXA3
ATF4
ZNF35
2085 2083 2085 2085
Not detected Low Low Low
2428 1939 2428 1874
High High Medium Medium

Figure 4


A
TCGA
EGR1
p < 1.0E-16
TEAD4
p < 8.2E-07
SOX2
p < 2.2E-04
RUNX3_BCL11A
p < 1.0E-16
Cumulative fraction of log2FC
Log2FC expression
(ER+HER2- ILCs vs ER+HER2- IDCs)
TF target genes
Background genes (n=16,537)
B
TCGA
FOXA1
p < 4.7E-10
HSF4
p < 1.0E-07
PBX3
p < 1.7E-06
PITX1
p < 1.0E-16
Cumulative fraction of log2FC
Log2FC expression
(ER+HER2- IDCs vs ER+HER2- ILCs)
TF target genes
Background genes (n=16,537)
METABRIC
EGR1
p < 1.0E-16
TEAD4
p < 1.0E-16
SOX2
p < 1.0E-16
RUNX3_BCL11A
p < 1.0E-16
Cumulative fraction of log2FC
Log2FC expression
(ER+HER2- ILCs vs ER+HER2- IDCs)
TF target genes
Background genes (n=24,176)
METABRIC
FOXA1
p < 5.9E-03
HSF4
p < 1E-16
PBX3
p < 1E-16
PITX1
p < 1E-16
Cumulative fraction of log 2FC
Log2FC expression
(ER+HER2- IDCs vs ER+HER2- ILCs)
TF target genes
Background genes (n=24,176)