Published online 24 March 2022

NAR Cancer, 2022, Vol. 4, No. 1 **1**
https://doi.org/10.1093/narcan/zcac008

Epigenetic alterations at distal enhancers are linked to proliferation in human breast cancer

Jørgen Ankill [1,2], Miriam Ragle Aure [3], Sunniva Bjørklund [3], Severin Langberg [4], Oslo Breast Cancer Consortium (OSBREAC), Vessela N. Kristensen [3], Valeria Vitelli [5], Xavier Tekpli [3] and Thomas Fleischer [1,*]

[1]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway, [2]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway, [3]Department of Medical Genetics, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway, [4]Cancer Registry of Norway, Oslo, Norway and [5]Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, Faculty of Medicine, University of Oslo, Oslo, Norway

Received November 18, 2021; Revised February 23, 2022; Editorial Decision March 04, 2022; Accepted March 14, 2022

ABSTRACT

Aberrant DNA methylation is an early event in breast carcinogenesis and plays a critical role in regulating gene expression. Here, we perform genome-wide expression-methylation Quantitative Trait Loci (emQTL) analysis through the integration of DNA methylation and gene expression to identify disease-driving pathways under epigenetic control. By grouping the emQTLs using biclustering we identify associations representing important biological processes associated with breast cancer pathogenesis including regulation of proliferation and tumor-infiltrating fibroblasts. We report genome-wide loss of enhancer methylation at binding sites of proliferation-driving transcription factors including CEBP-β, FOSL1, and FOSL2 with concomitant high expression of proliferation-related genes in aggressive breast tumors as we confirm with scRNA-seq. The identified emQTL-CpGs and genes were found connected through chromatin loops, indicating that proliferation in breast tumors is under epigenetic regulation by DNA methylation. Interestingly, the associations between enhancer methylation and proliferation-related gene expression were also observed within known subtypes of breast cancer, suggesting a common role of epigenetic regulation of proliferation. Taken together, we show that proliferation in breast cancer is linked to loss of methylation at specific enhancers and transcription factor binding and gene activation through chromatin looping.

INTRODUCTION

Epigenetic alterations, such as DNA methylation, have recently emerged as a hallmark of many cancer types including breast cancer. Previous studies have shown that changes in DNA methylation patterns are present already in pre-invasive lesions, thereby suggesting that such alterations occur early during breast cancer carcinogenesis (1–3). DNA methylation has been predominantly reported to be implicated in gene repression through promoter methylation (4), however, we have shown that DNA methylation at CpGs up to 100 kb away from a gene transcription start site could be associated with its expression (3). Furthermore, a major portion of the aberrantly methylated DNA observed in breast cancers occurs in intergenic regions. Altogether, this suggests that DNA methylation at distal *cis*-regulatory regions such as enhancers may be an important contributor to breast cancer development and progression (3,5).

Enhancers are *cis*-acting DNA sequences involved in transcriptional regulation. This process is mediated by cell-type-specific transcription factor (TF) binding and the formation of physical interactions between enhancers and promoters of their associated genes (6,7). TFs are key proteins involved in the regulation of gene expression and are linked to different functions depending on where they bind in the genome. While some TFs activate gene transcription by directly interacting with the transcriptional machinery, some TFs known as pioneer factors may regulate gene expression by remodeling the chromatin landscape to control transcriptional activity (8). TF accessibility to DNA is strictly controlled by the dynamic interplay between DNA methylation and histone modifications in a cell-type specific manner (9,10).

Several studies have reported DNA methylation at distal enhancer regions to be implicated in gene regulation mainly by interfering with TF binding to enhancer regions (11–13).

*To whom correspondence should be addressed. Tel: +47 98861883; Email: thomas.fleischer@rr-research.no

Enhancer methylation is known to be dynamic and more tissue-specific than promoter methylation, thereby suggesting that enhancers may play an essential role in contributing to cell phenotype (14–16). As for promoters, DNA methylation at enhancers tends to be associated with transcriptional inactivity, while enhancer hypomethylation is often associated with TF binding followed by transcriptional activation (7,17). However, the role of DNA methylation at enhancer regions and TF binding sites is still not fully understood.

We previously presented the genome-wide expression-methylation Quantitative Trait Loci (emQTL) analysis and showed that estrogen receptor (ER) positive breast tumors display disease-specific hypomethylation of enhancers carrying binding sites of ERα, FOXA1 and GATA3, suggesting an epigenetic regulation of estrogen signaling in breast cancer (18). The two most apparent clusters were reported: the described above estrogen cluster and a cluster related to varying immune infiltration. Here, we use a new approach, expand our analysis to include more patient samples, and use a sophisticated biclustering method to characterize novel biclusters of CpG-gene associations (Figure 1). We discover a proliferation-related bicluster in breast cancer characterized by hypomethylation at enhancers carrying transcription factor binding sites (TFBS) of proliferation-driving TFs in ER-negative (ER–) tumors. The identified CpGs and genes were found enriched in enhancer regions and to be connected through chromatin loops, thereby indicating that proliferation in breast cancer is under epigenetic regulation.

MATERIALS AND METHODS

Patient material

The OSL2 breast cancer cohort (19,20) has collected material from breast cancer patients with primary operable disease (T1–T2) in several south-eastern Norwegian hospitals. Patients were included between 2006 and 2019. The study was approved by the Norwegian Regional Committee for Medical Research Ethics (approval number 1.2006.1607, amendment 1.2007.1125). All patients have provided written consent for use of the material for research purposes.

The Cancer Genome Atlas Program (TCGA) breast cancer cohort has previously been described (21). Level 3 expression and methylation data were downloaded from the TCGA Data Portal (https://tcga-data.nci.nih.gov). CpGs and genes with >50% missing values were excluded and the remaining missing methylation values were imputed using the pamr (R function *pamr.knnimpute*) with $k = 10$. Only breast cancer tumor samples with matching expression and methylation data were included for emQTL validation in TCGA ($n = 558$).

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) has previously been described (22). METABRIC is a large gene expression cohort with a long follow-up time widely used for the investigation of breast cancer disease. Gene expression data is available from the European Genome Phenome Archive (DOI: EGAS00000000083, $n = 1980$).

Statistical computation and bioinformatical analyses

All computational analyses were performed using the R software v3.5.1 unless otherwise specified. The emQTL analysis R code can be found at https://github.com/JorgenAnkill/emQTL.

Results were considered statistically significant if the adjusted *P*-value was <0.05. Bar plots showing ChromHMM and UniBind enrichment results were generated using the R package *ggplot2* (23). Kaplan–Meier estimators and log-rank tests were performed using the *survival* R package v3.2.3 (functions *Surv* and *survfit*). Survival plots were made using the *survminer* R package (v0.4.8). The Upset plot was generated using the UpSet R-package v1.4.0 (24).

Genome-wide correlation analysis

Pearson's correlations between DNA methylation of CpGs with an interquartile range of more than 0.1 ($n = 182\,620$) were tested against all genes ($n = 18\,586$) for non-zero correlations in the OSL2 breast cancer cohort resulting in more than three billion tests. CpG-gene associations with a Bonferroni corrected *P*-value<0.05 (nominal *P*-value<1.47e–11) were considered significant. The significant CpG-gene associations in OSL2 were subsequently validated in the TCGA breast cancer cohort ($n = 558$). The significant associations were considered validated if the Bonferroni corrected *P*-value was <0.05 (nominal *P*-value < 6.70e–11). Of the 5 928 496 non-validated emQTL pairs, 28 523 associations could not be tested due to missing DNA methylation or expression data in TCGA. Only validated associations were included in the subsequent analyses. Probes and genes with less than five associations were filtered out. The remaining CpGs and genes with associations were kept in the following analyses. Before the analysis, gene symbols for expression data in the discovery and validation cohort were harmonized using the R package *HGNChelper* v0.7.1 (function *checkGeneSymbols*).

Biclustering of the emQTL correlation coefficients

The inverse correlation coefficients ($r^* - 1$) from the emQTL analysis were biclustered using Python (v3.7.9) by applying the *SpectralCoclustering* algorithm contained within the scikit-learn library (25). The biclustering algorithm will identify biclusters in which the rows intersecting the columns within a bicluster will have a higher average value than the intersecting columns outside a bicluster. For the initial spectral co-clustering analysis, the random_state parameter was set to 0. Spectral co-clustering was performed using the inverse correlation coefficients (correlation coefficient* – 1) values obtained from the OSL2 discovery cohort. Biclustering of the absolute correlation coefficient values was also performed for comparison to using the inverse correlation coefficients (see Supplementary Text, Supplementary Table S2A, Supplementary Figures S9–S11). Python code used for the biclustering is available from https://github.com/JorgenAnkill/emQTL.

Determination of the number of biclusters for the biclustering algorithm was performed by calculating a mean square residue (MSR) score (26) when the number of biclusters was set to be a number between 2 and 20. A lower MSR

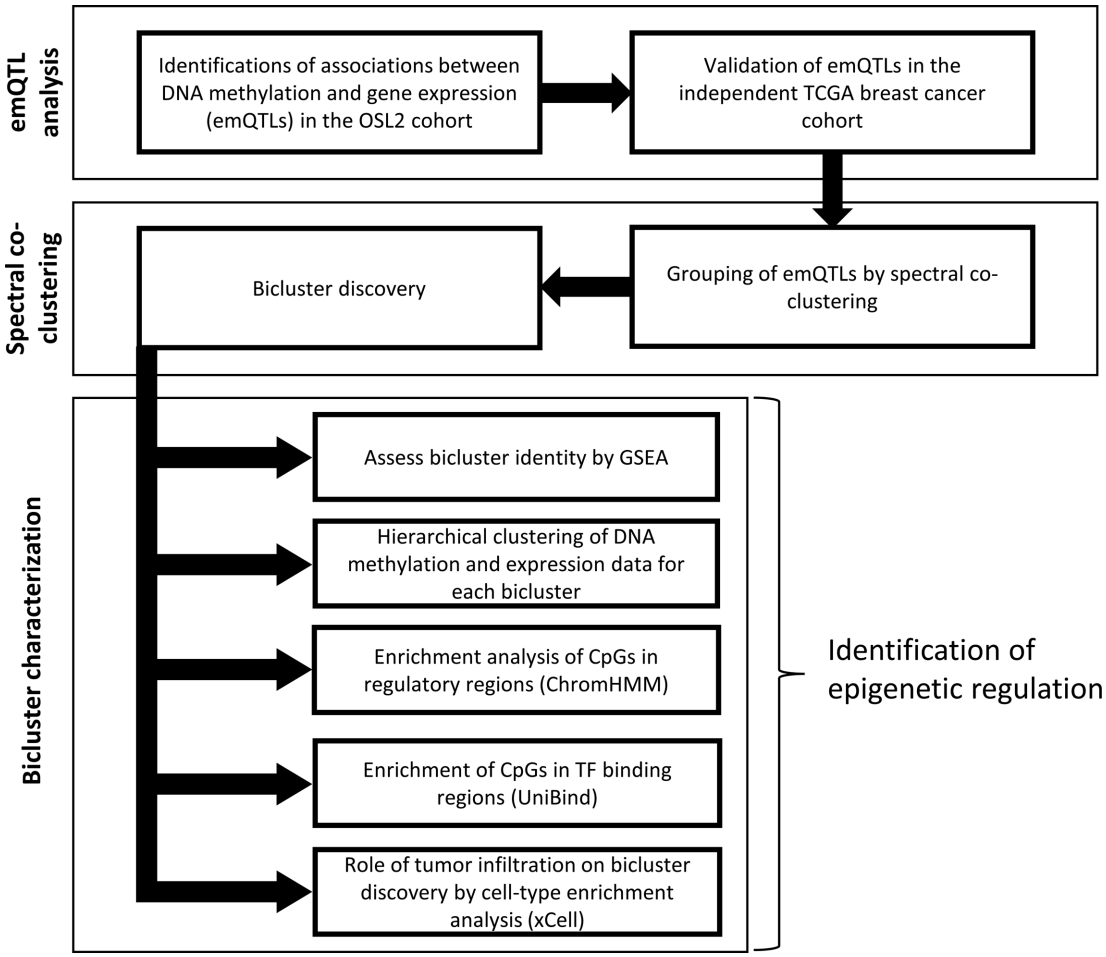

Figure 1. Study overview. Flowchart showing the pipeline used for identification of CpG-gene associations (emQTLs) and methods used for emQTL grouping and characterization.

score is associated with a stronger coherence exhibited by the biclusters and thereby indicates better biclustering. The elbow of the plot was defined to be the number of biclusters, which was therefore set to be 5. Biclustering by setting the number of biclusters to be 8 or 12 was also performed as a comparison (see Supplementary Text, Supplementary Table S2B, C).

Assessment of bicluster stability was performed using a permutation test (100 permutations) using random seeds and comparing the biclusters obtained with the biclusters from the initial biclustering when the random_state parameter was set to 0. For each run, GSEA was performed on the gene list from each bicluster identified to define their biological functions. The number of times a CpG or gene for each bicluster from the initial biclustering was found within a bicluster of similar biological functions for each permutation was then calculated (Supplementary Figure S2A, B). The estrogen biclusters (Biclusters 2 and 4) were considered as one bicluster in this analysis.

Gene set enrichment analysis

Gene sets used for GSEA analysis were downloaded from the Molecular Signatures Database v7.1 (27). Enrichment was determined by hypergeometric testing (R function *phyper*) using the Hallmark (H) and gene ontology (GO; C5) gene set collections. *P*-values were corrected for multiple testing using the Benjamini–Hochberg (BH) procedure (R function *p.adjust*).

Hierarchical clustering of DNA methylation and gene expression levels

Hierarchical clustering of the DNA methylation- and gene expression levels was performed using the R package *pheatmap* using Euclidean distance and the ward.D2 cluster agglomeration methods. For visualization purposes, gene expression values were centered and scaled by rows by dividing the centered rows by their standard deviations (R function *scale*).

Genomic segmentation and annotation

ChromHMM is a software for learning and characterizing chromatin states by using multivariate Hidden Markov Model for identifying combinatorial patterns of histone marks obtained from ChIP-seq data to functionally annotate the genome (28). ChromHMM segmentation data

from cell lines representing different breast cancer subtypes were obtained from Xi *et al.* (28), which included MCF7 and ZR751 (Luminal A), UACC812, and MB361 (Luminal B), HCC1954 and AU565 (Her2+), HCC1937 and MB469 (Basal-like). ChIP-seq peaks for key histone modifications including H3K4me3, H3K4me1, H3K27me3, H3K9me3, and H3K36me3 (GSE85158) were used to predict chromatin states across the genome of the cell lines. The genomes were annotated into thirteen distinct chromatin states including active promoter (PrAct), active promoter flanking (PrFlk), active transcription (TxAct), active transcription flanking (TxFlk), active intergenic enhancer (EhAct), active genic enhancer (EhGen), bivalent promoter (PrBiv), bivalent enhancer (EhBiv), repressive polycomb domain (RepPC), weak repressive domain (WkRep), repeat/ZNF genes (RpZNF), heterochromatin (Htchr) and quiescent state/low signals (QsLow). Subtype-specific ChromHMM annotations were made by collapsing the ChromHMM annotations from cell lines of similar subtypes and keeping the common ones.

Enrichment of CpGs in a ChromHMM defined functional region was measured as the ratio between the frequency of cell cycle bicluster-CpGs found in a specific segment type over the frequency of CpGs from the Illumina HumanMethylation450 array found within the same segment type. *P*-values were obtained by hypergeometric testing with the Illumina 450k array probes as background ($n =$ 485 512). *P*-values were corrected for multiple testing using the BH procedure.

TF enrichment analysis in UniBind-defined TF binding regions

Enrichment of CpGs in TF binding regions was assessed using data obtained from the UniBind 2018 (29) database. UniBind is a database storing direct TF-DNA interactions for 231 unique human TFs obtained from 1983 ChIP-seq datasets performed on 315 different cell lines and tissues. Maps of direct TF–DNA interactions were downloaded from the UniBind website (https://unibind2018.uio.no) for the prediction model PWM. The genomic positions of all CpGs from the Illumina 450k array were lifted over from hg19 to hg38 using the LiftOver web tool from the UCSC genome browser (https://genome.ucsc.edu) and were extended with 150 bp upstream and downstream. Since each TF can have binding regions derived from multiple ChIP-seq experiments, we merged the TF binding regions for all ChIP-seq experiments for each TF. Enrichment of CpGs in proximity to TF binding regions was computed using hypergeometric testing (R function *phyper*) with IlluminaMethylation450 Bead Chip CpGs as background. False discovery rate was estimated by BH correction using the R function *p.adjust*.

scRNA-seq data

Count matrix from single-cell RNA-seq obtained from Qian *et al.* (30) was analyzed using the Seurat R package v3.2.1 (31) to obtain UMAP. In brief, the count matrix was already filtered for dying cells by the authors. It was further normalized and scaled regressing out potential confounding factors (number of UMIs, number of genes detected in cell, percentage of mitochondrial RNA). After scaling, variably expressed genes were used to construct principal components (PCs). PCs covering the highest variance in the dataset were selected based on elbow and Jackstraw plots to build the UMAP. Clusters were calculated by the *FindClusters* function with a resolution between 0.8 and 1.8 and visualized using the UMAP dimensional reduction method. Four main cell types were identified on these UMAP, combining both the information obtained from the UMAP clustering and cell-type annotation from the authors. The main cell types were immune-, cancer-, endothelial cells and fibroblasts.

xCell analysis

The xCell (32) algorithm was used to deconvolute the cellular composition of the tumor samples. xCell is a powerful machine learning framework trained on 64 immune and stromal cell datasets used to generate cell-type specific enrichment scores and adjust them to cell-type proportions. The algorithm uses 10 808 genes as signatures to identify specific cell types from bulk tissue. The cell type enrichment scores were calculated for the OSL2 cohort ($n = 272$) using the xCell (32) web tool (http://xcell.ucsf.edu/). Gene names from the expression data of the OSL2 cohort were harmonized with the gene list provided by the xCell tool before the analysis using the HGNChelper v0.7.1 R package. Pre-calculated xCell scores for TCGA tumor samples were downloaded from http://xcell.ucsf.edu/xCell_TCGA_RSEM.txt.

Chromatin interaction mapping

Chromatin Interaction Analysis by Paired-End Tag sequencing (ChIA-PET) and Integrated Methods for Predicting Enhancer Targets (IM-PET) are methods used to identify such interactions on a genome-wide scale (33,34). ChIA-PET data defining long-range chromatin interactions in the ER+ MCF7 breast cancer cell line was obtained from ENCODE (Accession number ENCR000CAA; (33)). Only *cis* loops were included in the analysis. An emQTL was considered to be in a ChIA-PET Pol2 loop if the CpG and transcription start site of its associated gene were found within the genomic intervals of two opposite feet of the same loop. Computational chromatin interactions predicted by the IM-PET algorithm for the ER- HCC1954 breast cancer cell line was retrieved from the 4Dgenome data portal (35). BEDTools v2.27.1 (36) was used to intersect the CpG and gene positions with the genomic intervals defining the feet of the chromatin loops for the ChIA-PET and IM-PET data. Chromatin interaction plots were made using the Gviz v1.32.0 (37) and GenomicRanges v1.40.0 (38) R packages. Genome interaction tracks were made using the R package *GenomicInteractions* v1.22.0 (39).

Enrichment of *in cis* (i.e. on the same chromosome) emQTLs in ChIA-PET and IM-PET loops was determined by hypergeometric tests (R function *phyper*) using all possible *in cis* CpG-gene pairs as background.

Identification of proliferation-promoting emQTLs

A supervised approach was used to identify proliferation-promoting emQTLs by using the characteristics of the cell cycle bicluster in the candidate search; (i) the CpG-gene pair must be on opposite sides of chromatin loops defined by ChIA-PET (33) and/or IM-PET (34) loops and (ii) be located in enhancers according to ChromHMM segmentation (28) of either subtype. (iii) The CpG must be in the binding region of the top enriched TFs as defined by UniBind (29) and (iv) the gene must be a part of a curated gene set associated with proliferation. There must also be a significant correlation between DNA methylation at the candidate CpG and the candidate gene.

RESULTS

Expanded expression-methylation Quantitative Trait Loci (emQTL) analysis

Genome-wide in *cis* and in *trans* correlations between the gene expression and DNA methylation at CpGs was performed in the OSL2 breast cancer cohort ($n = 277$, Supplementary Figure S1). We identified 16 193 303 significant CpG-gene associations (Bonferroni corrected P-values < 0.05) of which 10 264 807 (63.4%) were validated in the independent The Cancer Genome Atlas (TCGA) breast cancer cohort (BRCA, $n = 558$). Among these associations, 613 600 were cis-emQTLs and they were significantly more enriched than trans-emQTLs (9 324 057 associations, P-value $< 2.2e{-}16$, fold enrichment $= 1.19$). The validated associations involved the expression level of 6803 genes and methylation level of 64 439 CpGs. To focus on hub associations, emQTL-CpGs and emQTL-genes with less than five associations were filtered out. The remaining CpGs ($n = 44\,263$) and genes ($n = 4904$) with associations after filtering were included in downstream analyses. To identify emQTLs with similar biological features, we grouped the emQTLs using Spectral co-clustering (40) of the inverse correlation coefficients values (correlation coefficient* – 1; see Supplementary Text) obtained from the genome-wide emQTL analysis.

The number of biclusters was set to five based on a mean square residue score (MSR, Figure 2A: see also Materials and Methods). The five biclusters were (Figure 2B, Supplementary Table S1A, B): Bicluster 1 (8641 CpGs and 1085 genes), Bicluster 2 (9398 CpGs and 870 genes), Bicluster 3 (6910 CpGs and 936 genes), Bicluster 4 (10 564 CpGs and 1087 genes) and Bicluster 5 (8750 CpGs and 926 genes). To confirm that the identified biclusters were not artifacts of the selected seed parameter used by the biclustering algorithm, we assessed bicluster stability by using a permutation test (see Materials and Methods). The biclusters were found to be very stable as only 15 genes and 113 CpGs were found <70% of the time within a bicluster of similar biological characteristic (Supplementary Figure S2A, B).

To elucidate the biological role of the biclusters, gene set enrichment analysis (GSEA) was performed based on the genes of each bicluster (Figure 2C and Supplementary Table S1C). As expected, we rediscovered the estrogen- (Biclusters 2 and 4) and immune cluster (Bicluster 5) first described by Fleischer, Tekpli et al. (18). The majority of the previous paper immune cluster genes (94.5%) and CpGs (53.5%) were found in the newly discovered immune bicluster and the same was true for the estrogen cluster genes (53.9%) and CpGs (56.7%). Moreover, the median correlation coefficients from OSL2 between emQTL-CpGs in Bicluster 2 and expression of the emQTL genes in Bicluster 4 were negative (Supplementary Figure S3), thereby suggesting that DNA methylation at the CpGs show similar trends in both DNA methylation and expression in both biclusters. This suggests that these two biclusters, separated by the biclustering algorithm recapitulate the same biological pathway. In addition to rediscovering the immune- and estrogen clusters, we now identify two novel biclusters with distinct biological functions: cell cycle regulation (Bicluster 1) and epithelial-mesenchymal transition (EMT), extracellular matrix (ECM), and cell locomotion (Bicluster 3) as shown in Figure 2C.

Enhancer methylation, TF binding and a proliferative phenotype of human breast tumors

To understand the functional link between DNA methylation at CpG sites and expression of genes in the cell cycle bicluster (Bicluster 1), we first aimed to characterize the functional genomic location of the CpGs using ChromHMM segmentation (see Material and methods) including breast cancer cell lines representing different breast cancer subtypes (28). CpGs in the cell cycle bicluster were significantly enriched (P-value $= 1.08e{-}77$) in active intergenic enhancer regions of breast tumors across all subtypes (Figure 3A, Supplementary Table S1D). Moreover, we found that 46% of the CpGs overlapped with at least one active intergenic enhancer region of another subtype (Figure 3B).

Having found the cell cycle bicluster-CpGs to be enriched in intergenic enhancer regions, we next sought to identify transcription factor binding regions (TFBR) overlapping the cell cycle bicluster-CpGs using direct TF-DNA interaction data obtained from UniBind (29). The genomic regions of the cell cycle bicluster-CpGs were found enriched in the binding region of TFs previously described to regulate proliferation in breast cancers including CEBP-β (41) and several of the FOS family of proteins including FOS (42), FOSL1 (43,44), and FOSL2 (45) (Figure 3C, Supplementary Table S1E).

Using unsupervised hierarchical clustering, we further investigated the level of DNA methylation of cell cycle bicluster-CpGs in regard to histopathological features including ER status and PAM50 subtype. DNA methylation level of the cell cycle bicluster CpGs ($n = 8641$) was clearly associated with the breast cancer subtypes (Chi-squared test P-value $= 0.0005$ (three patient subclusters), Figure 3D), and the CpGs in this bicluster showed lower methylation levels in the Basal-like, Her2-enriched, and Normal-like tumors in both the OSL2 (Figure 3E) and TCGA (Figure 3F) breast cancer cohorts. Moreover, DNA methylation at the cell cycle bicluster-CpGs in TF binding regions of the top six most enriched TFs was lower in Basal-like and Her2-enriched breast tumors (Supplementary Figure S4A-F). Altogether, these results show that CpGs in the cell cycle bicluster are enriched for enhancer regions overlapping TFBR of TFs associated with proliferation such as CEBP-β, FOSL1, and FOSL2. Moreover, their TFBR were found

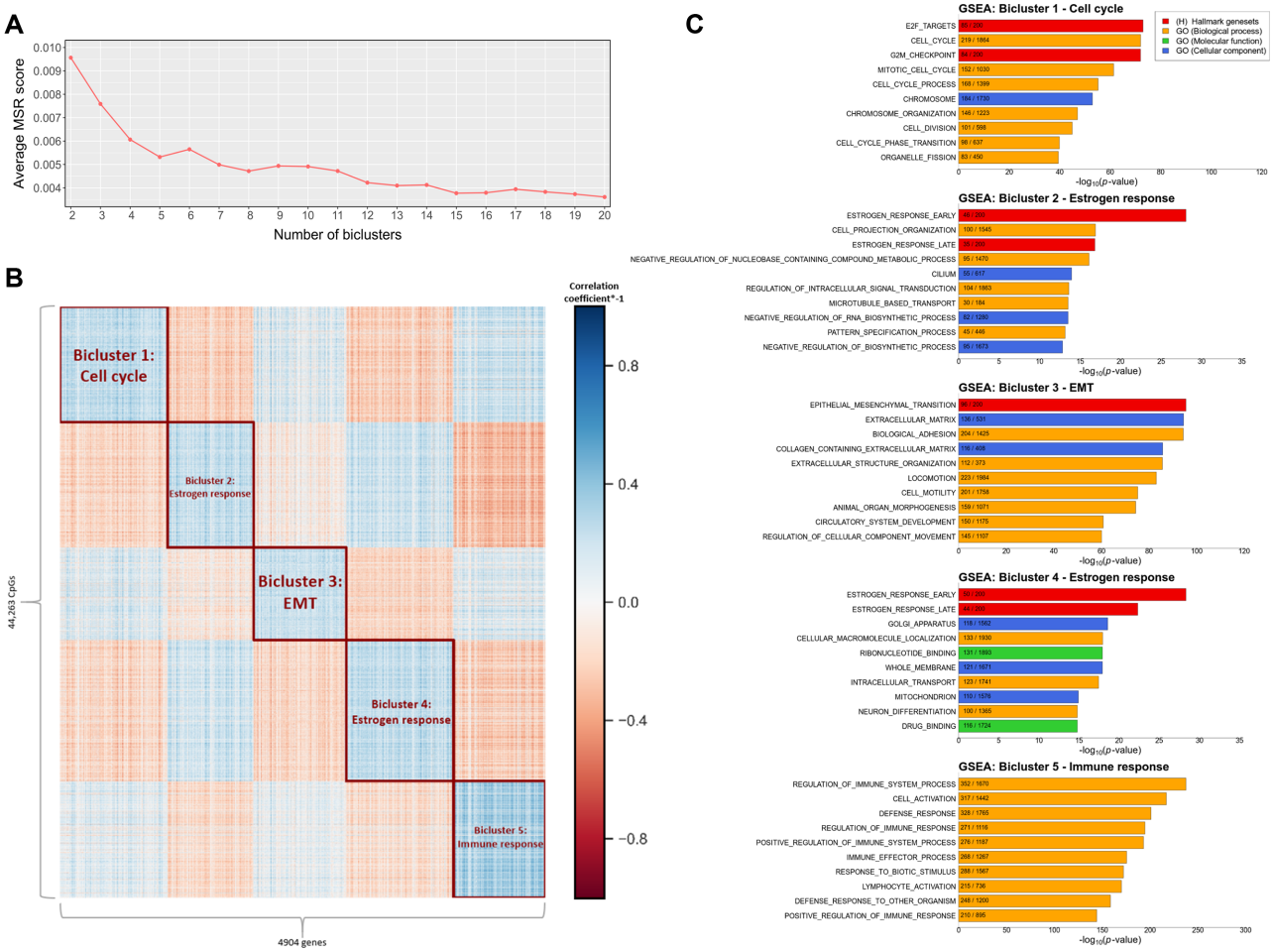

Figure 2. Identification and characterization of the emQTL biclusters. (**A**) Line chart showing the average MSR score for the biclusters obtained by biclustering of the inverse correlation coefficients obtained from the emQTL analysis in OSL2 when the number of biclusters k were set to be a number between 2 and 20. (**B**) Heatmap showing the inverse correlation coefficients of the emQTL-CpGs ($n = 44\,263$) and emQTL-genes ($n = 4904$) from OSL2 after biclustering. Rows represent CpGs and columns represent genes. Each of the five biclusters is annotated. Blue points indicate negative correlations between the variables while red points represent positive correlations. White points indicate little or no correlation. (**C**) GSEA of the genes in Bicluster 1 ($n = 1085$), Bicluster 2 ($n = 870$), Bicluster 3 ($n = 936$), Bicluster 4 ($n = 1087$) and Bicluster 5 ($n = 926$) using gene sets obtained from the MSigDB. The length of the bars represents the log-transformed Benjamini-Hochberg (BH) corrected P-values obtained by hypergeometric distribution. Red bars indicate Hallmark gene sets while GO biological process, GO molecular function, and GO cellular compartment GO gene sub-collections are colored in orange, green and blue, respectively. Overlap between the gene list of the bicluster and each MSigDB gene set is annotated within each bar.

to be less methylated in Basal-like and Her2-enriched compared to Luminal A and B tumors.

One of the most known markers of cell proliferation is the MKI67 gene which is a non-histone nuclear protein expressed during the active phase of the cell cycle (46). We found a significant negative correlation between average methylation of the cell cycle bicluster CpGs and expression of MKI67 both in OSL2 and TCGA within the ER- tumors (P-value = 0.012 and 0.0005) and for all breast tumors (P-value = 1.79e–13 and 2.78e–13, Supplementary Figure S5A, B). Interestingly, we also find MKI67 to reside within the cell cycle bicluster (Supplementary Table S1B). These observations support the link between DNA methylation at the cell cycle bicluster-CpGs and proliferation.

To assess the link between DNA methylation and gene expression in the cell cycle bicluster, we performed unsupervised clustering using the expression of genes within the bicluster and observed that expression was higher in the subtypes known to have higher proliferation rates (Figure 4A). Basal-like tumors showed the highest expression, followed by Her2-enriched and Normal-like (Figure 4B). Luminal A tumors showed the lowest expression of cell cycle bicluster genes in OSL2 (Figure 4B) and TCGA (Figure 4C). A correlation analysis between average methylation and average expression of the CpGs and genes in the cell cycle bicluster separately within ER+ and ER– tumors showed a significant correlation (Figure 4D, E). Taken together, these results show a statistically significant association between enhancer methylation and expression of proliferation-related genes, and that ER– breast tumors have low methylation at enhancers potentially driving proliferation. The varying degree of enhancer methylation may be related to the proliferative potential in ER– tumors.

To investigate the functional relationship between DNA methylation and gene expression in the cell cycle bicluster we assessed the extent to which CpGs within this bicluster

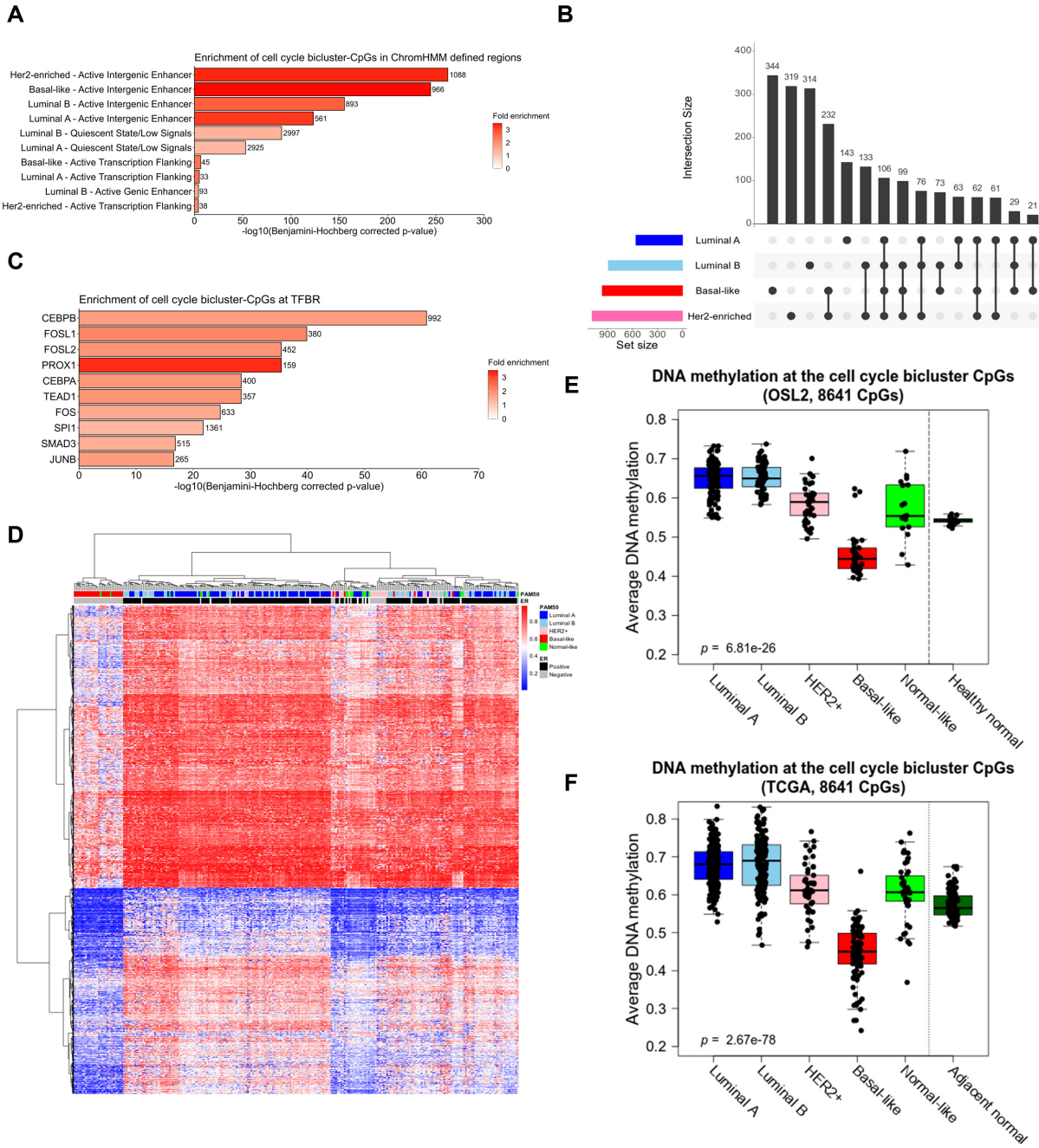

Figure 3. Functional characterization of the emQTL-CpGs in the cell cycle bicluster. (**A**) Bar plot the showing enrichment of the cell cycle bicluster-CpGs in ChromHMM-defined genomic regions by subtype. The length of the bars represents the log-transformed BH corrected *P*-values. The color gradient of the bars represents fold enrichment (FE) in which a red color indicates FE close to 3.5 while white bars are genomic regions by subtype with FE close to 0. An enrichment was considered to be significant if the BH-corrected *P*-value was less than 0.05. (**B**) UpSet plot showing the overlap between CpGs in the cell cycle bicluster found within ChromHMM-defined active intergenic enhancer regions by breast cancer subtype. (**C**) Bar plot representing enrichment of the cell cycle bicluster-CpGs in the binding region of specific TFs according to UniBind. Bar length displays the log-transformed BH-corrected *P*-value obtained by hypergeometric testing for each TF. Red color indicates FE close to 3.5 while a white color indicates FE close to 0. (**D**) Unsupervised hierarchical clustering of DNA methylation levels of the 8641 cell cycle bicluster-CpGs for the tumor tissue from OSL2 with PAM50 status available ($n = 272$). Rows represent CpGs and columns represent histopathological features including PAM50 subtype and ER status of the tumor samples. Red points indicate methylated CpGs while blue points represent unmethylated CpGs. Boxplots showing the average DNA methylation of the cell cycle bicluster-CpGs in (**E**) OSL2 ($n = 272$) and (**F**) TCGA ($n = 562$) by PAM50 subtype. Kruskal-Wallis test *P*-values are denoted in the lower-left corner.

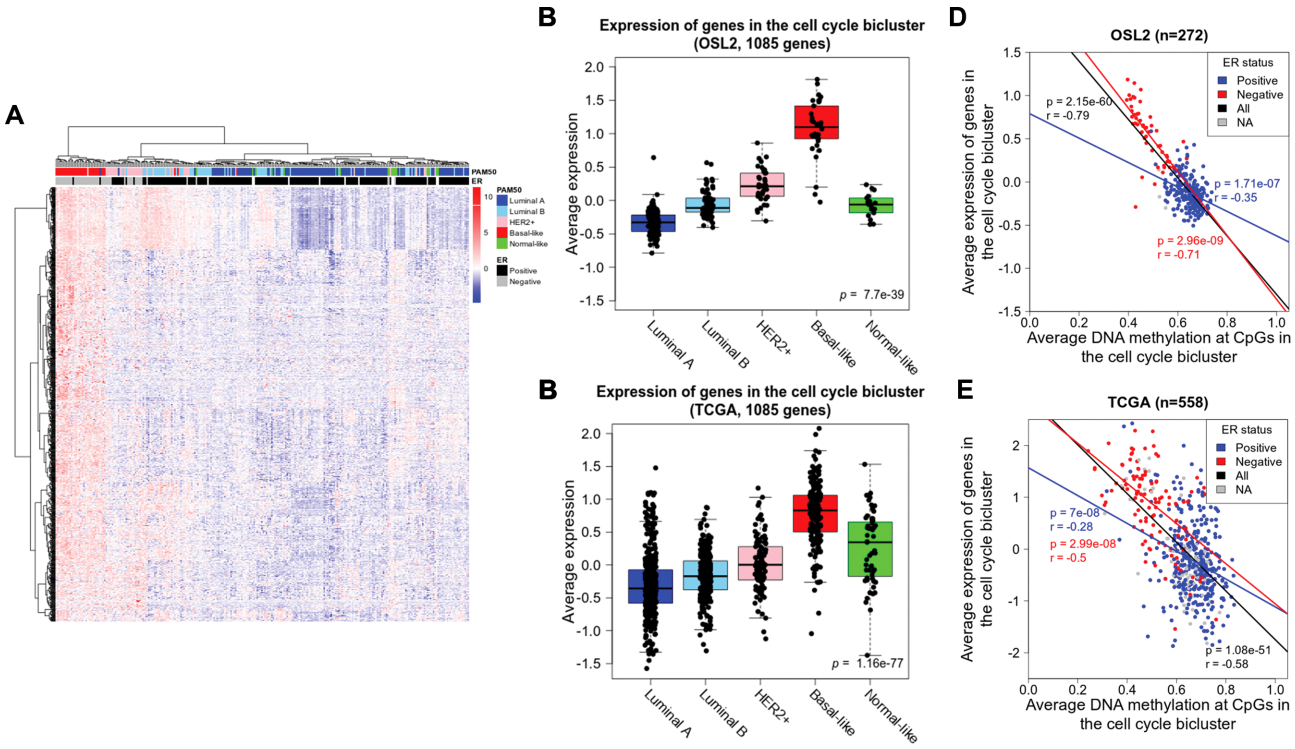

Figure 4. Expression of genes in the cell cycle bicluster. (**A**) Unsupervised clustering of the expression levels of the 1085 genes in the cell cycle bicluster for the tumors in the OSL2 breast cancer cohort ($n = 272$). Rows represent genes and columns represent samples annotated with histopathological features including PAM50 subtype and ER status. Red color indicates high expression levels and blue color indicates low. Boxplots showing the average expression of genes in the cell cycle bicluster in the (**B**) OSL2 ($n = 272$) and (**C**) TCGA ($n = 981$) breast cancer cohorts. Kruskal-Wallis test *P*-values are denoted. Scatterplots showing the association between average DNA methylation of the cell cycle bicluster-CpGs versus average expression of the genes contained within the same bicluster by ER status in the OSL2 (**D**) and TCGA (**E**) breast cancer cohorts. Pearson correlation coefficients and *P*-values are denoted and colored by ER status.

were located nearby (±10 kb) any of the genes contained within the same bicluster. We found that 36% of the genes in the cell cycle bicluster were located nearby at least one CpG in the same bicluster suggesting that many genes in the cell cycle bicluster may be locally regulated by DNA methylation in enhancer regions at TFBR of the enriched TFs including CEBP-β, FOSL1 and FOSL2.

Enhancers can promote gene expression of distant genes by interacting with promoter regions of their associated genes through chromatin loop formation (6,7). To assess the potential physical contact between enhancers with loss of methylation and expression of their target genes, we obtained ChIA-PET Pol2 data (33) from MCF7- (ER+) and IM-PET interaction (34) from the HCC1954 (ER–) breast cancer cell lines. We found that the CpGs-gene pairs in emQTLs in the cell cycle bicluster were significantly enriched in ChIA-PET and IM-PET defined loops (hypergeometric test *P*-value = 5.1e–3 and 4.7e–4 respectively, Figure 5A). Altogether, 59 CpGs were experimentally confirmed by ChIA-PET to form physical interactions with 39 unique genes in the cell cycle bicluster (Supplementary Table S1F), and 22 unique emQTL CpG-gene loops were confirmed by IM-PET (Supplementary Table S1G). Altogether, these results suggest that emQTLs represent direct regulatory links between DNA methylation at enhancer regions targeted by proliferation-associated TFs and the expression of the cell cycle bicluster-genes (Figure 5B).

Identification of potential key drivers of proliferative signaling in breast cancer

Knowing that enhancer methylation at TF binding regions is associated with the regulation of proliferation-related genes, we next performed a supervised emQTL approach to more efficiently identify key drivers of carcinogenic signaling. With the supervised approach, we can include all emQTLs prior to the filtering step and identify proliferation-promoting emQTLs that are independent of the performance of the biclustering algorithm. Using the supervised approach, we select emQTLs based on the identified cell cycle bicluster characteristics (see Material and Methods). We identified 53 proliferation-promoting candidate emQTLs in which the majority (79%) show negative correlations between DNA methylation and gene expression (Table 1, Supplementary Table S1H). Figure 5C shows one of the potential proliferation-promoting emQTL in which the CpG (cg00733115) is found experimentally by ChIA-PET to interact with the Pim-1 Proto-Oncogene, Serine/Threonine Kinase (*PIM1*) gene found in the GO_CELL_CYCLE gene set from the MSigDB (27). The CpG itself is located in a region with a high abundance of active intergenic enhancer chromatin marks in Basal-like, Her2-enriched, and Luminal A subtype according to ChromHMM (28). Moreover, the CpG overlaps with the binding region of several members of the FOS fam-

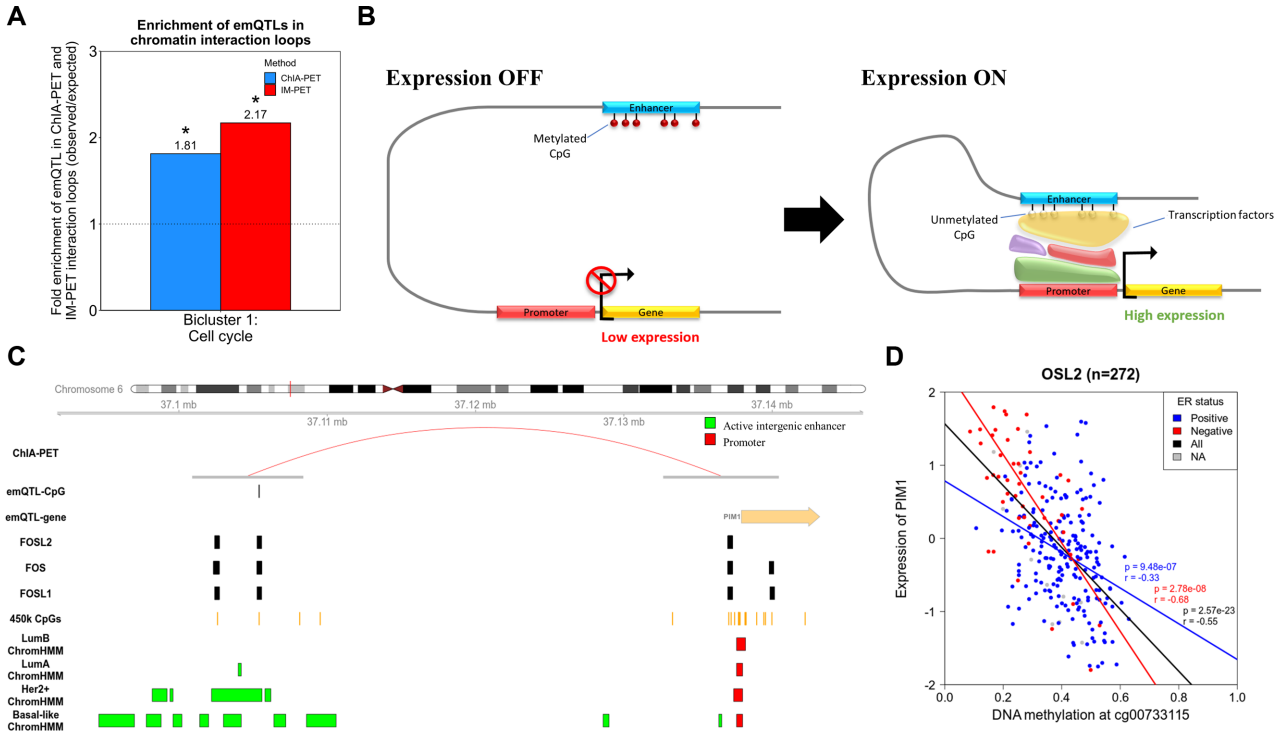

Figure 5. DNA methylation at enhancers facilitates target gene expression through enhancer-promoter interactions. (**A**) Bar plot showing the enrichment of emQTLs in ChIA-PET Pol2 loops and IM-PET loops for the ER+ MCF7 and ER– HCC1954 breast cancer cell lines, respectively. Bar height represents the enrichment level measured as the ratio between the frequency of emQTLs (CpG-gene pairs) found in the head and tail of a loop over the expected frequency if such overlaps were to occur at random. Enrichments that are statistically significant (hypergeometric test, BH corrected P-value < 0.05) are marked with an asterisk. (**B**) Enhancer hypomethylation at specific enhancers allows TF binding and the transcriptional activation of enhancer target genes through physical enhancer-promoter interactions by chromatin looping. (**C**) An example of a potential proliferation-promoting alteration in which the CpG (cg00733115) has been found in one foot of a ChIA-PET Pol2 loop (red arc) and a gene associated with proliferation (PIM1) is found in the other. Annotations for active intergenic enhancer regions and active promoters according to ChromHMM that are conserved across the cell lines of a similar subtype are shown in green and blue color respectively by breast cancer subtype. The binding sites of FOS, FOSL1/2 are also shown. (**D**) Scatterplot showing the association between DNA methylation at the emQTL-CpG cg00733115 and its associated gene (PIM1) by ER status in OSL2. Pearson's correlation coefficients and P-values are denoted.

ily of proteins including the FOS, FOSL1/2 TFs. The TF binding region of FOS as indicated in Figure 5C is obtained from the breast epithelial MCF10A cell line. A negative correlation was observed between DNA methylation of cg00733115 and the expression of *PIM1* in both ER– and ER+ tumors (Figure 5D). These results suggest that we can identify promising proliferation-promoting emQTLs using the supervised emQTL approach.

The cell cycle bicluster associates with prognosis

To investigate the prognostic impact of the identified genes, we performed survival analysis in the METABRIC breast cancer cohort ($n = 1904$). When stratifying tumors by PAM50 subtype and dividing the patients into two groups based on the median of the average expression values of the genes in the cell cycle bicluster, we observe high expression to be associated with worse prognosis within the Luminal A, Luminal B and Normal-like breast subtypes (Figure 6A–E, log-rank P-value $= 0.00044$, 0.0019 and 0.01 respectively). When performing the survival analysis independent of subtype, we observe a significant association between survival and expression (Figure 6F, log-rank P-value < 0.0001).

Rediscovery of the immune- and estrogen response related biclusters

Both the immune and the estrogen biclusters rediscovered in this study were found to overlap with the immune and estrogen-related clusters first described by Fleischer, Tekpli *et al.* (18). The immune bicluster-CpGs were found enriched in close proximity to TF binding regions of several TFs involved in immune cell homeostasis such as RUNX1, FLI1 and ERG (Supplementary Table S1E). DNA methylation and gene expression levels of the rediscovered immune bicluster was associated with varying degree of immune infiltration (Supplementary Figure S6A, B). Furthermore, immune cells including leukocytes, monocytes, T-cells, and B-cells showed similar methylation levels at the immune bicluster CpGs as the tumors with high immune infiltration. Contrary, the ER-positive MCF7 and ER-negative MDAMB453 breast cancer cell lines showed methylation levels similar to the tumors with low immune infiltration (Supplementary Figure S6A).

Our emQTL-CpGs and genes associated with estrogen response separated into two biclusters (Figure 2C, Supplementary Table S1A, B). This is likely due to the predominance of estrogen response-related emQTLs, as the biclus-

Table 1. Top potential cancer-promoting alterations identified using the supervised emQTL approach. The strength of the correlations and the Bonferroni corrected *P*-value is shown for the OSL2 ($n = 272$) and TCGA ($n = 558$) breast cancer cohort. The table is ordered by the strength of the negative correlation in OSL2

	Pearson's correlation coefficient		Adjusted *P*-value		
emQTL ID	OSL2	TCGA	OSL2	TCGA	Method
cg00733115_PIM1	–0.55	–0.32	1.36E–21	7.38E–13	ChIA-PET
cg02976539_SLC9A3R1	–0.54	–0.52	2.02E–20	2.69E–38	IM-PET
cg18037834_KRT18	–0.54	–0.51	3.14E–20	2.66E–37	ChIA-PET
cg15880704_PDCD4	–0.54	–0.29	3.78E–20	2.00E–10	ChIA-PET
cg04482712_SLC9A3R1	–0.54	–0.54	7.70E–20	4.59E–42	IM-PET
cg00484122_RHOB	–0.54	–0.42	8.75E–20	1.15E–23	ChIA-PET
cg16729850_KRT18	–0.53	–0.51	5.44E–19	1.41E–36	IM-PET
cg21359793_KRT18	–0.52	–0.50	9.20E–19	2.27E–35	IM-PET
cg20812370_PBX1	–0.51	–0.39	6.70E–18	6.52E–20	ChIA-PET
cg12610744_KRT18	–0.51	–0.51	8.58E–18	2.98E–36	IM-PET

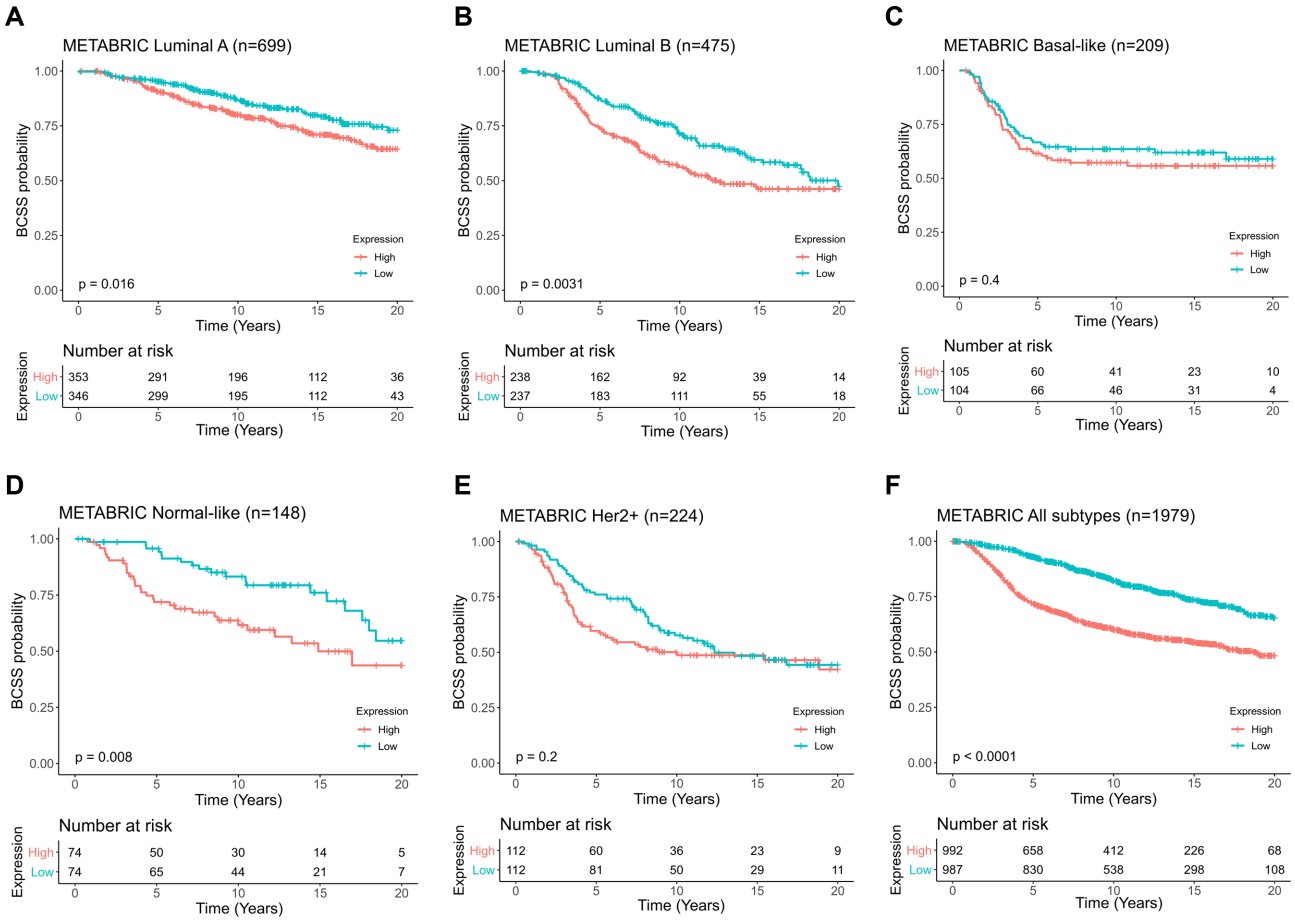

Figure 6. Expression of genes in the cell cycle bicluster associates with prognosis. Kaplan-Meier survival curves for the cell cycle bicluster in METABRIC cohort, for Luminal A (**A**), Luminal B (**B**), Basal-like (**C**), Normal-like (**D**), Her2-enriched (**E**) and all breast cancer subtypes (**F**). Tumors were divided into two groups based on the median of the average expression of genes in the cell cycle bicluster. *P*-values obtained by log-rank test are denoted.

tering algorithm favors more equally sized biclusters. The CpGs in estrogen bicluster 2 and 4 were significantly less methylated in ER+ compared to ER– tumors (Wilcoxon rank-sum test, $P = 1.06e–18$, and 1.48e–18 respectively, Supplementary Figure S7A, B). The estrogen-related genes in both biclusters were overexpressed in ER+ tumors compared to the ER– (Wilcoxon rank-sum test, $P = 3.16e–24$ and 9.25e–20, Supplementary Figure S7C, D). Moreover, CpGs within each of the estrogen-related biclusters were enriched in enhancer regions and genomic regions in proximity to TFBR of several TFs associated with estrogen-response such as ERα, FOXA1 and GATA3 (Supplementary Table S1D, E). This was observed in both estrogen biclusters which suggests that these two biclusters represent the same biological pathway. Altogether, these results are in concordance with the corresponding findings regarding the estrogen cluster previously described by Fleischer, Tekpli *et al.* (18).

Bicluster 3 reflects a varying degree of fibroblast infiltration

GSEA indicated that genes in bicluster 3 were related to processes including EMT, ECM and cell locomotion (Supplementary Table S1C). Contrary to the cell cycle bicluster, genes and CpGs in the EMT bicluster (Bicluster 3) seems to a lesser extent to segregated the breast cancer patients according to the PAM50 subtypes (Figure 7A, Supplementary Figure S8A). Fibroblasts are prominent cell types of the tumor microenvironment and carry out functions related to ECM remodeling while also being able to migrate (47). We therefore, hypothesized that this bicluster was linked to fibroblast infiltration. To examine this, we estimated the relative amount of fibroblasts for each tumor sample using the xCell (32) deconvolution tool which is based on mRNA expression. By dividing the tumors into quartile groups based on the amount of fibroblast infiltration, we found the EMT bicluster gene expression levels to be associated with fibroblast infiltration in OSL2 and TCGA (Figure 7B, C, Kruskal–Wallis test P-value = 1.74e–27 and 3.35e–20, respectively), i.e. high expression of the EMT bicluster genes is associated with high fibroblast infiltration. Altogether this suggests that the expression levels of these genes may be caused by a high expression of the EMT bicluster genes in tumor-infiltrating fibroblasts rather than the cancer cells themselves.

Furthermore, we characterized the CpGs in the EMT bicluster and found them to be enriched in active intergenic enhancer regions, but to a lower extent than the cell cycle bicluster CpGs (Supplementary Table S1D). No significant enrichment of EMT bicluster-CpGs in emQTL with EMT bicluster genes was observed from the ChIA-PET Pol2 and IM-PET data (Supplementary Figure S8B). TF enrichment analysis revealed significant enrichment of the CpGs in the TFBR of several TFs previously linked to EMT such as FOSL1 (48), TEAD1 (49), NFIC (50), and TWIST1 (51) (Supplementary Table S1E). Mean methylation of CpGs in the EMT bicluster was associated with varying degrees of fibroblast infiltration in OSL2 and TCGA (Figure 7D, E), i.e. increasing fibroblast infiltration was associated with decreased DNA methylation.

To further support the hypothesis that the EMT bicluster was related to varying degrees of fibroblast infiltration, rather than differences in the EMT potential of breast tumors we obtained DNA methylation data (Illumina 450k) from the PMC42 breast cancer cell line before and after EGF-induces EMT. No prominent change in DNA methylation levels at the EMT bicluster CpGs was observed after EGF-induced EMT in the PMC42 breast cancer cell line (Figure 7D, E). Moreover, the cell line displayed similar methylation levels as the tumors with low fibroblast infiltration, i.e. high methylation. By contrast, human mammary fibroblasts were unmethylated at the EMT bicluster-CpGs. Taken together, these results show that DNA methylation and the expression level of genes in the EMT bicluster are mainly caused by fibroblast infiltration.

Cell-type-specific expression of genes in the emQTL-biclusters by scRNA-seq

Since the tumor microenvironment consists of a highly dynamic and heterogeneous collection of cells, we used scRNA-seq data from 14 breast cancer patients (30) to investigate the cell-type specific expression of a subset of genes from each bicluster. For the analysis, we selected 10 genes from each bicluster showing the strongest negative correlation coefficient with an associated emQTL-CpG within the same bicluster. We found most of the genes from the cell cycle bicluster to be cancer-specific compared to other cells types such as immune cells, fibroblasts, and endothelial cells which are prominent cell types of the tumor microenvironment (Figure 8A, B). Moreover, these genes were highly expressed by cancer cells from tumors classified as Her2-enriched and triple-negative breast cancer (TNBC) subtypes compared to Luminal A and Luminal B (Figure 8C–F). Altogether, this supports the hypothesis that the cell cycle bicluster genes are important regulators of proliferation in breast cancer cells. Similarly, to the cell cycle bicluster, the estrogen-related genes were almost exclusively expressed by cancer cells from ER+ tumors (Figure 8C–F). Contrary, the genes associated with the EMT- and immune biclusters were mainly expressed by fibroblasts and immune cells respectively (Figure 8B).

DISCUSSION

Cancer initiation and progression involve altered proliferation rates that play an important role in breast cancer pathogenesis (52,53). Today, little is known about how DNA methylation contributes to the proliferative phenotype of breast tumors. By performing genome-wide emQTL analysis before biclustering of the correlation coefficients, we identify a previously unreported gene regulatory network involved in breast cancer carcinogenesis. In ER– breast tumors, we observe hypomethylation at enhancers carrying TFBR of key proliferation-driving TFs with concomitant high expression of proliferation-related genes in tumor cells as confirmed by scRNA-seq. We show that the identified CpGs and genes were connected through chromatin loops. Taken together, we show that proliferation in breast cancer is linked to loss of enhancer methylation and TF binding through chromatin loops. The causal effects the candidates have on the observed associations regarding the cancer phenotype will be of great interest for future studies.

The proliferation-related CpGs were found significantly enriched in active intergenic enhancer regions of all breast cancer subtypes, but most pronouncedly in Basal-like, Her2-enriched, and Luminal B tumors according to ChromHMM (28), which are also the most proliferative subtypes. Using chromatin loop enrichment analysis, we found the CpGs in the cell cycle bicluster to be significantly enriched in chromatin loops in both ER+ and ER– cell lines, thereby strengthening the hypothesis that the transcriptional network associated with proliferation could be regulated by DNA methylation independent of ER status. Because the loop data from the ER+ and ER– breast cancer cell lines are generated using different technologies the ER+ and ER– cell lines are not directly comparable and the results should be interpreted independently. TF enrichment analysis showed an enrichment of the proliferation-related CpGs nearby TFBR of several TFs known to be implicated in breast cancer tumorigenesis including CEBP-β, FOSL1 and FOSL2. The TFBSs of these TFs stored

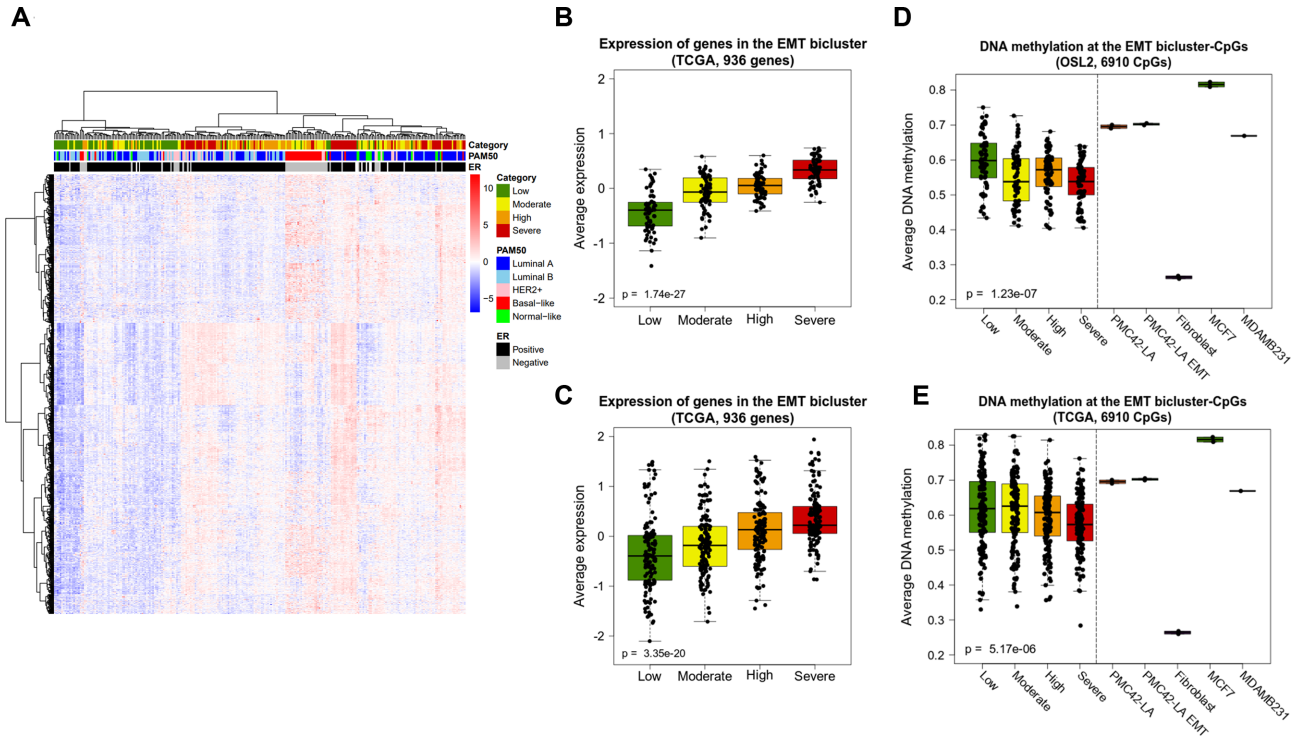

Figure 7. The EMT bicluster highlights an association between DNA methylation and fibroblast infiltration. (**A**) Heatmap showing the unsupervised clustering of the expression levels of the 936 genes contained within the EMT bicluster for 272 tumor samples from the OSL2 cohort. Rows represent genes and columns represent tumor samples annotated by histopathological features including PAM50 subtype and ER status. The tumor samples were divided into quartile groups based on the severity of fibroblast infiltration according to the relative amount of fibroblast in the tumor samples estimated by xCell. Differences in expression of the EMT bicluster genes between the quartile groups are shown for the OSL2 (**B**) and TCGA (**C**) cohorts. Each quartile group consisted of 68 tumor samples in OSL2 and 139 samples in TCGA. Boxplots showing the average DNA methylation at the 6910 CpGs contained within the EMT bicluster according to fibroblast infiltration score in (**D**) OSL2 ($n = 272$) and (**E**) TCGA ($n = 556$). Average DNA methylation values for these CpGs for in the PMC42-LA before and after EGF induced EMT. Fibroblasts, and the ER+ MCF7 and ER– MDAMB436 breast cancer cell lines are also included. Kruskal–Wallis test *P*-values are denoted in the bottom left corner.

in UniBind are based on ChIP-seq data from breast cancer cell lines among others. While the CEBP-β TFBS have been mapped by ChIP-seq in the estrogen receptor-positive (ER+; MCF7), FOSL1 TFBS have been mapped in the ER– BT549 and FOSL1/2 in the ER+ MCF7 breast cancer cell lines by ChIP-seq respectively. The CEBP family of TFs are known to be involved in regulating proliferation, and the CEBP-β member is commonly overexpressed in ER– tumors compared to ER+ tumors and is positively associated with tumor grade (41). Several of the Fos family TFs have also been implicated in proliferation. FOSL1 binding has previously been found enriched at enhancers of triple-negative breast cancers and positively associated with proliferation in ER– and ER+ cell lines (43). Furthermore, FOSL2 overexpression has been linked to proliferation in the triple-negative MDA-MB-231 and Her2-enriched SK-BR-2 breast cancer cell lines (45). FOS has previously been shown to be an important regulator of proliferation in the MCF7 breast cancer cell line (42). Here, we show that the CpGs in close proximity to the TFBS of these TFs were less methylated in the most proliferative tumor subtypes such as the Basal-like and Her2-enriched tumors. Altogether, we speculate that demethylation of the cell cycle bicluster-CpGs leads to more frequent binding of proliferation-related TFs and looping to their associated gene, thereby causing enhanced expression. The predictive and prognostic relevance of DNA methylation levels around the genomic regions binding CEBP-β, FOSL1 and FOSL2 constitute interesting regions for further investigation. At present, there is a lack of ChIP data mapping genome-wide TF-DNA interactions. Therefore, there may be other TFs as well involved in TF binding at the specified enhancers that are not included here and might also be drivers of proliferation in breast cancer.

By characterizing several aspects of the regulatory pathways associated with proliferation in breast cancers, we were able to identify potential downstream drivers of carcinogenic signaling relating to proliferation. The identified candidate gene with the strongest and most significant negative correlation was the *PIM1* gene which belongs to the Serine/Threonine protein kinase family of proteins. PIM1 is known to be implicated in the cell cycle, and knockdown experiments in TNBC cell lines have been shown to decreased proliferation and survival (54). Another candidate was *CDKL3* which is a *CDK3* homolog belonging to the cyclin-dependent protein kinase (CDK) family of proteins. *CDKL3* is known to be implicated in cell cycle progression from G1 to the S phase (55,56). The methylation status of an emQTL-CpG located in a distal enhancer region was found linked to the expression of the *CDKL3* gene through chromatin looping defined by an experimentally defined ChIA-PET Pol2 loop. A previous study found CDKL3 upregula-

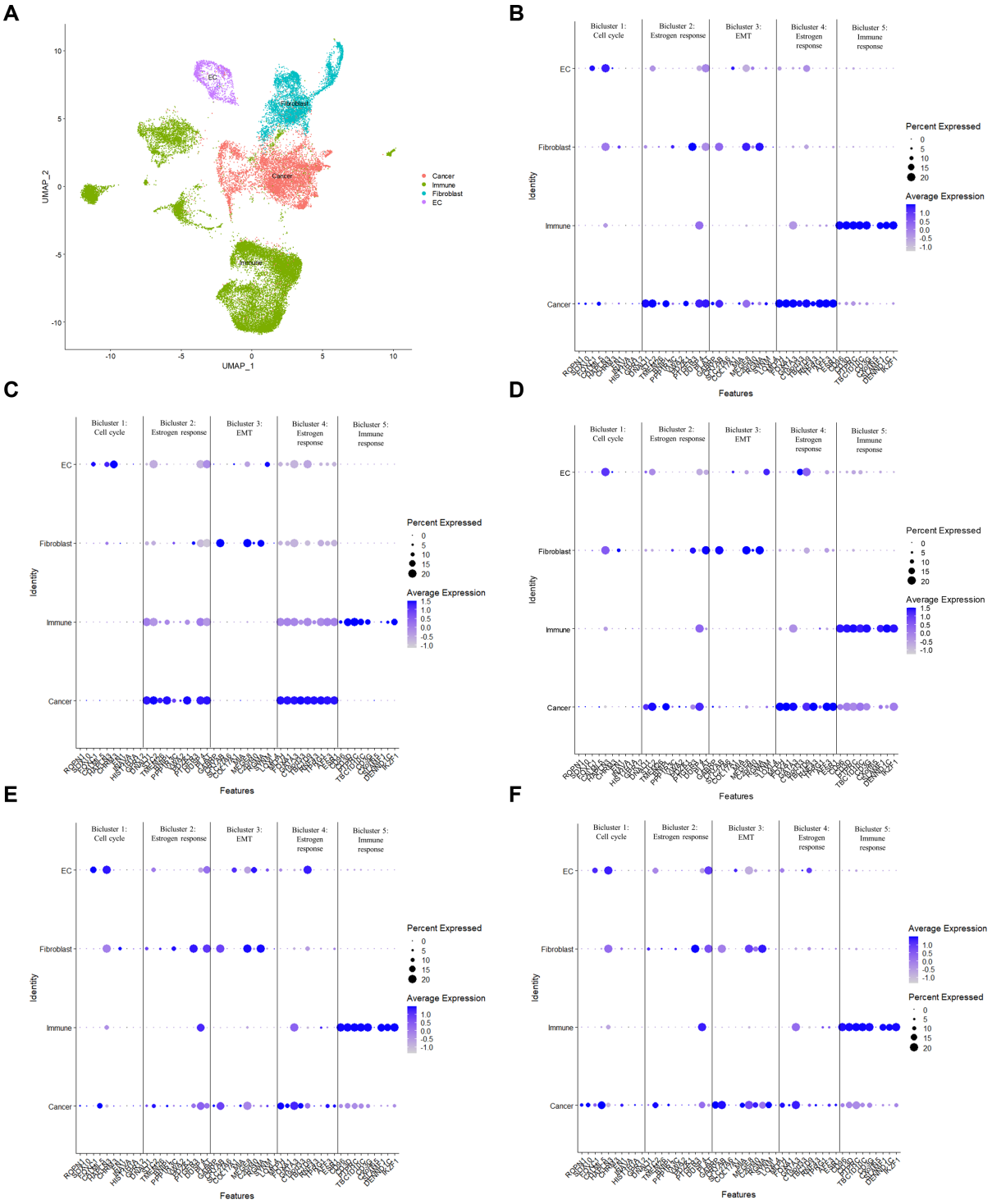

Figure 8. Cell-type-specific expression of the emQTL-bicluster genes. Combined Uniform Manifold Approximation and Projection (UMAP) plot for all 14 breast cancer samples annotated by cell type (**A**). Dot plots showing the expression of the selected genes for each bicluster for B all patients and patients with a (**C**) Luminal A, (**D**) Luminal B, (**E**) Her2-enriched and (**F**) TNBC. The size of the dot depicts the percentage of cells within each class and the intensity of the color shows the average expression level of each class.

tion to be associated with faster-growing HeLa cells derived from cervical cancer (57). However, less is known about the influence of *CDKL3* upregulation on proliferation in breast cancer. Another candidate such as MUC1, which is an oncoprotein, has been linked to proliferation in breast cancer cell lines upon siRNA knockdown experiments (58). Cyclin D1 (CCND1) is involved in the progression of several cancer types including breast, lung, esophagus, and bladder cancers. CCND1 is associated with proliferation by regulating the G1/S-phase transition (59). Knockdown of *CCND1* using siRNA has been shown to decrease proliferation rates in the MCF7 breast cancer cell line (60). Altogether, this indicates that our identified proliferation-promoting candidate genes play key roles in proliferation-related processes in breast cancer.

Previous studies have linked increased proliferation rates with prognosis in breast cancers (52,53). Here, we report the expression of the proliferation-related genes in the cell cycle bicluster to be associated with poorer prognosis within the established breast cancer subtypes, including Luminal A and Luminal B. We are thereby identifying a subgroup of patients which may benefit from more aggressive treatment, and equally importantly, we identify a subgroup of patients that may benefit from less treatment.

Fibroblasts, also known as cancer-associated fibroblasts (CAFs) in a tumor setting, are among the most abundant cell types of the tumor microenvironment involved in functions related to ECM remodeling (61,62). They play a key role in promoting tumorigenesis (63). An increasing number of studies have emphasized a possible link between infiltration of CAFs and epigenetic changes in tumor cells. One of the most characterized CAF-secreted factors, TGF-β, can mediate epigenetic changes through SOX4 activation, which in turn modulates the EZH2 histone methyltransferase in cancer cells (64). Moreover, aberrant DNA methylation can occur on a genome-wide scale in tumor cells treated with TGF-β (65,66). Fibroblast infiltration has been associated with treatment response and metastatic potential of cancer cells (67–70). By using the xCell (32) deconvolution tool which is based on gene expression data, we found associations between fibroblast infiltration versus expression- and methylation levels of genes and CpGs in the EMT bicluster in OSL2 and TCGA (Figure 7B–E). Lower DNA methylation at the EMT bicluster-CpGs was associated with higher fibroblast infiltration, and fibroblasts were unmethylated at these CpGs compared to tumor tissue. The emQTL analysis highlights how DNA methylation and gene expression levels may reflect infiltration levels in the tumor microenvironment. Even though the EMT bicluster is associated with fibroblast infiltration, there may be a less pronounced EMT-related signal from the tumors themselves represented in the EMT bicluster caused by fibroblast infiltration or other factors. Therefore, a more detailed study of the epigenetic effects of crosstalk between fibroblasts and tumor cells regarding the EMT bicluster would be of future interest.

In this study, we provide genome-wide evidence that DNA methylation at intergenic enhancer regions is a key regulator of proliferation in breast cancers. The CpG sites involved were proximal to TFBSs of CEBP-β, FOSL1 and FOSL2, which are TFs associated with proliferation in breast cancers. Altogether, we establish an association between DNA methylation and tumor phenotype reflecting the proliferative potential of breast cancer tumors.

DATA AVAILABILITY

R-code related to the emQTL analysis together with spectral co-clustering code in Python is available at GitHub (https://github.com/JorgenAnkill/emQTL). Level 3 gene expression and DNA methylation data from the TCGA breast cancer cohort can be found at the TCGA data portal at https://tcga-data.nci.nih.gov (21). ChromHMM segmentation data from breast cancer cell lines were obtained from Xi *et al.* (28). TF-DNA interactions in available from the UniBind database at https://unibind2018.uio.no (29).

Clinical data including PAM50 classification and mRNA expression data from the OSL2 breast cancer cohort can be obtained from GEO with accession number GSE58215 (20) and DNA methylation data ($n = 277$) is available at GEO with the accession number GSE84207 (18). The sample key to combine GSE58215 (gene expression) and GSE84207 (DNA methylation) for the OSL2 patient cohort is available upon request. Expression data from METABRIC is available from the European Genome Phenome Achieve (EGAS00000000083; https://ega-archive.org/studies/) (22). ChIA-PET data from MCF7 can be obtained from ENCODE (ENCSR000CAA; https://www.encodeproject.org/experiments/) (33) and IM-PET data from HCC1954 is available from the 4D genome data portal at https://4dgenome.research.chop.edu/Download.html (35). Illumina HumanMethylation450 BeadChip data from 17 normal healthy samples obtained by mammoplasty reductions can be obtained from GSE60185 (3). DNA methylation data from cell lines used in this study is available from GEO: PMC42-LA (GSE97853), human mammary fibroblasts (GSE74877), T-cells (GSE79144), Monocytes (GSE68456), Leukocytes (GSE69270), B-cells (GSE68456), MCF7 (GSE69188) and MDAMBA453 (GSE124368).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Cancer Online.

ACKNOWLEDGEMENTS

We would like to acknowledge Daniel Nebdal for his technical support and assistance during the project.
OSBREAC members:
Tone F Bathen (PhD), Norwegian University of Science and Technology, Norway
Elin Borgen (PhD, MD), Oslo University Hospital, Norway
Olav Engebråten (PhD, MD), Oslo University Hospital, Norway
Britt Fritzman (MD), Østfold Hospital, Norway Øystein Garred (PhD, MD), Oslo University Hospital, Norway
Jürgen Geisler (PhD, MD) Akershus University Hospital, Norway
Gry Aarum Geitvik, Oslo University Hospital, Norway
Solveig Hofvind (PhD), Cancer Registry of Norway, Norway
Rolf Kåresen (PhD, MD), Oslo University Hospital, Norway

Anita Langerød (PhD), Oslo University Hospital, Norway
Ole Christian Lingjærde (PhD), University of Oslo, Norway
Gunhild Mari Mælandsmo (PhD), Oslo University Hospital, Norway
Bjørn Naume (PhD, MD), Oslo University Hospital, Norway
Hege G Russnes (PhD, MD), Oslo University Hospital, Norway
Torill Sauer (PhD, MD), Akershus University Hospital, Norway
Helle Kristine Skjerven (MD), Vestre Viken Hospital Trust, Oslo University Hospital, Norway
Therese Sørlie (PhD), Oslo University Hospital, Norway

FUNDING

South-Eastern Norway Regional Health Authority [2020031 and 2017065 to T.F.]; S.M.B. and M.R.A. were postdoctoral fellows of the Norwegian Cancer Society [711164 to V.N.K.].
Conflict of interest statement. None declared.

REFERENCES

1. van Hoesel,A.Q., Sato,Y., Elashoff,D.A., Turner,R.R., Giuliano,A.E., Shamonki,J.M., Kuppen,P.J.K., van de Velde,C.J.H. and Hoon,D.S.B. (2013) Assessment of DNA methylation status in early stages of breast cancer development. *Br. J. Cancer*, **108**, 2033–2038.
2. Jovanovic,J., Ronneberg,J.A., Tost,J. and Kristensen,V. (2010) The epigenetics of breast cancer. *Mol. Oncol.*, **4**, 242–254.
3. Fleischer,T., Frigessi,A., Johnson,K.C., Edvardsen,H., Touleimat,N., Klajic,J., Riis,M.L.H., Haakensen,V.D., Wärnberg,F., Naume,B. *et al.* (2014) Genome-wide DNA methylation profiles in progression to in situand invasive carcinoma of the breast with impact on gene transcription and prognosis. *Genome Biol.*, **15**, 435.
4. Kulis,M. and Esteller,M. (2010) DNA methylation and cancer. *Adv. Genet.*, **70**, 27–56.
5. Kamalakaran,S., Varadan,V., Giercksky Russnes,H.E., Levy,D., Kendall,J., Janevski,A., Riggs,M., Banerjee,N., Synnestvedt,M., Schlichting,E. *et al.* (2011) DNA methylation patterns in luminal breast cancers differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables. *Mol. Oncol.*, **5**, 77–92.
6. Bulger,M. and Groudine,M. (2011) Functional and mechanistic diversity of distal transcription enhancers. *Cell*, **144**, 327–339.
7. Shlyueva,D., Stampfel,G. and Stark,A. (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, **15**, 272–286.
8. Yu,X. and Buck,M.J. (2020) Pioneer factors and their in vitro identification methods. *Mol. Genet. Genomics*, **295**, 825–835.
9. Sur,I. and Taipale,J. (2016) The role of enhancers in cancer. *Nat. Rev. Cancer*, **16**, 483–493.
10. Ordoñez,R., Martínez-Calle,N., Agirre,X. and Prosper,F. (2019) DNA methylation of enhancer elements in myeloid neoplasms: think outside the promoters?*Cancers*, **11**, 1424.
11. Héberlé,É. and Bardet,A.F. (2019) Sensitivity of transcription factors to DNA methylation. *Essays Biochem.*, **63**, 727–741.
12. Lea,A.J., Vockley,C.M., Johnston,R.A., Del Carpio,C.A., Barreiro,L.B., Reddy,T.E. and Tung,J. (2018) Genome-wide quantification of the effects of DNA methylation on human gene regulation. *elife*, **7**, e37513.
13. Tong,Y., Sun,J., Wong,C.F., Kang,Q., Ru,B., Wong,C.N., Chan,A.S., Leung,S.Y. and Zhang,J. (2018) MICMIC: identification of DNA methylation of distal regulatory regions with causal effects on tumorigenesis. *Genome Biol.*, **19**, 73.
14. Agirre,X., Castellano,G., Pascual,M., Heath,S., Kulis,M., Segura,V., Bergmann,A., Esteve,A., Merkel,A., Raineri,E. *et al.* (2015) Whole-epigenome analysis in multiple myeloma reveals DNA hypermethylation of B cell-specific enhancers. *Genome Res.*, **25**, 478–487.
15. Aran,D., Sabato,S. and Hellman,A. (2013) DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.*, **14**, R21.
16. Kulis,M., Queirós,A.C., Beekman,R. and Martín-Subero,J.I. (2013) Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer. *Biochim. Biophys. Acta*, **1829**, 1161–1174.
17. Wiench,M., John,S., Baek,S., Johnson,T.A., Sung,M.H., Escobar,T., Simmons,C.A., Pearce,K.H., Biddie,S.C., Sabo,P.J. *et al.* (2011) DNA methylation status predicts cell type-specific enhancer activity. *EMBO J.*, **30**, 3028–3039.
18. Fleischer,T., Tekpli,X., Mathelier,A., Wang,S., Nebdal,D., Dhakal,H.P., Sahlberg,K.K., Schlichting,E., Sauer,T., Geisler,J. *et al.* (2017) DNA methylation at enhancers identifies distinct breast cancer lineages. *Nat. Commun.*, **8**, 1379.
19. Aure,M.R., Vitelli,V., Jernström,S., Kumar,S., Krohn,M., Due,E.U., Haukaas,T.H., Leivonen,S.-K., Vollan,H.K.M., Lüders,T. *et al.* (2017) Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome. *Breast Cancer Res.*, **19**, 44.
20. Aure,M.R., Jernström,S., Krohn,M., Vollan,H.K., Due,E.U., Rødland,E., Kåresen,R., Ram,P., Lu,Y., Mills,G.B. *et al.* (2015) Integrated analysis reveals microRNA networks coordinately expressed with key proteins in breast cancer. *Genome Medicine*, **7**, 21.
21. Koboldt,D.C., Fulton,R.S., McLellan,M.D., Schmidt,H., Kalicki-Veizer,J., McMichael,J.F., Fulton,L.L., Dooling,D.J., Ding,L., Mardis,E.R. *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.
22. Curtis,C., Shah,S.P., Chin,S.F., Turashvili,G., Rueda,O.M., Dunning,M.J., Speed,D., Lynch,A.G., Samarajiwa,S., Yuan,Y. *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.
23. Wickham,H. (2009) In: *ggplot2: Elegant Graphics for Data Analysis*. Springer, NY.
24. Conway,J.R., Lex,A. and Gehlenborg,N. (2017) UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, **33**, 2938–2940.
25. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R. and Dubourg,V. (2011) Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
26. Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 93–103.
27. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Nat. Acad. Sci. U.S.A.*, **102**, 15545–15550.
28. Xi,Y., Shi,J., Li,W., Tanaka,K., Allton,K.L., Richardson,D., Li,J., Franco,H.L., Nagari,A., Malladi,V.S. *et al.* (2018) Histone modification profiling in breast cancer cell lines highlights commonalities and differences among subtypes. *BMC Genomics*, **19**, 150.
29. Gheorghe,M., Sandve,G.K., Khan,A., Chèneby,J., Ballester,B. and Mathelier,A. (2019) A map of direct TF-DNA interactions in the human genome. *Nucleic. Acids. Res.*, **47**, e21.
30. Qian,J., Olbrecht,S., Boeckx,B., Vos,H., Laoui,D., Etlioglu,E., Wauters,E., Pomella,V., Verbandt,S., Busschaert,P. *et al.* (2020) A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Res.*, **30**, 745–762.
31. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck,W.M. 3rd, Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.
32. Aran,D., Hu,Z. and Butte,A.J. (2017) xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.*, **18**, 220.
33. Li,G., Ruan,X., Auerbach,R.K., Sandhu,K.S., Zheng,M., Wang,P., Poh,H.M., Goh,Y., Lim,J., Zhang,J. *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**, 84–98.
34. He,B., Chen,C., Teng,L. and Tan,K. (2014) Global view of enhancer-promoter interactome in human cells. *Proc. Nat. Acad. Sci. U.S.A.*, **111**, E2191–E2199.

35. Teng,L., He,B., Wang,J. and Tan,K. (2015) 4DGenome: a comprehensive database of chromatin interactions. *Bioinformatics*, **31**, 2560–2564.
36. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
37. Hahne,F. and Ivanek,R. (2016) Visualizing genomic data using gviz and bioconductor. *Methods Mol. Biol.*, **1418**, 335–351.
38. Lawrence,M., Huber,W., Pagès,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M.T. and Carey,V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.
39. Harmston,N., Ing-Simmons,E., Perry,M., Barešić,A. and Lenhard,B. (2015) GenomicInteractions: an R/Bioconductor package for manipulating and investigating chromatin interaction data. *BMC Genomics*, **16**, 963.
40. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
41. Zahnow,C.A. (2009) CCAAT/enhancer-binding protein beta: its role in breast cancer and associations with receptor tyrosine kinases. *Expert Rev. Mol. Med.*, **11**, e12.
42. Lu,C., Shen,Q., DuPré,E., Kim,H., Hilsenbeck,S. and Brown,P.H. (2005) cFos is critical for MCF-7 breast cancer cell growth. *Oncogene*, **24**, 6516–6524.
43. Belguise,K., Kersual,N., Galtier,F. and Chalbos,D. (2005) FRA-1 expression level regulates proliferation and invasiveness of breast cancer cells. *Oncogene*, **24**, 1434–1444.
44. Franco,H.L., Nagari,A., Malladi,V.S., Li,W., Xi,Y., Richardson,D., Allton,K.L., Tanaka,K., Li,J., Murakami,S. *et al.* (2018) Enhancer transcription reveals subtype-specific gene expression programs controlling breast cancer pathogenesis. *Genome Res.*, **28**, 159–170.
45. He,J., Mai,J., Li,Y., Chen,L., Xu,H., Zhu,X. and Pan,Q. (2017) miR-597 inhibits breast cancer cell proliferation, migration and invasion through FOSL2. *Oncol. Rep.*, **37**, 2672–2678.
46. Ahmed,S.T., Ahmed,A.M., Musa,D.H., Sulayvani,F.K., Al-Khyatt,M. and Pity,I.S. (2018) Proliferative index (Ki67) for prediction in breast duct carcinomas. *Asian Pac. J. Cancer Prev.*, **19**, 955–959.
47. Walker,C., Mojares,E. and Del Rio Hernandez,A. (2018) Role of extracellular matrix in development and cancer progression. *Int. J. Mol. Sci.*, **19**, 3028.
48. Lamar,J.M., Stern,P., Liu,H., Schindler,J.W., Jiang,Z.-G. and Hynes,R.O. (2012) The Hippo pathway target, YAP, promotes metastasis through its TEAD-interaction domain. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E2441.
49. Bakiri,L., Macho-Maschler,S., Custic,I., Niemiec,J., Guío-Carrión,A., Hasenfuss,S.C., Eger,A., Müller,M., Beug,H. and Wagner,E.F. (2015) Fra-1/AP-1 induces EMT in mammary epithelial cells by modulating Zeb1/2 and TGFβ expression. *Cell Death Differ.*, **22**, 336–350.
50. Lee,H.K., Lee,D.S. and Park,J.C. (2015) Nuclear factor I-C regulates E-cadherin via control of KLF4 in breast cancer. *BMC Cancer*, **15**, 113.
51. Xu,Y., Qin,L., Sun,T., Wu,H., He,T., Yang,Z., Mo,Q., Liao,L. and Xu,J. (2017) Twist1 promotes breast cancer invasion and metastasis by silencing Foxa1 expression. *Oncogene*, **36**, 1157–1166.
52. van Diest,P.J., van der Wall,E. and Baak,J.P.A. (2004) Prognostic value of proliferation in invasive breast cancer: a review. *J. Clin. Pathol.*, **57**, 675–681.
53. Beresford,M.J., Wilson,G.D. and Makris,A. (2006) Measuring proliferation in breast cancer: practicalities and applications. *Breast Cancer Res.*, **8**, 216.
54. Brasó-Maristany,F., Filosto,S., Catchpole,S., Marlow,R., Quist,J., Francesch-Domenech,E., Plumb,D.A., Zakka,L., Gazinska,P., Liccardi,G. *et al.* (2016) PIM1 kinase regulates cell death, tumor growth and chemotherapy response in triple-negative breast cancer. *Nat. Med.*, **22**, 1303–1313.
55. Malumbres,M. (2014) Cyclin-dependent kinases. *Genome Biol.*, **15**, 122.
56. Zheng,K., He,Z., Kitazato,K. and Wang,Y. (2019) Selective autophagy regulates cell cycle in cancer therapy. *Theranostics*, **9**, 104–125.
57. Jaluria,P., Betenbaugh,M., Konstantopoulos,K. and Shiloach,J. (2007) Enhancement of cell proliferation in various mammalian cell lines by gene insertion of a cyclin-dependent kinase homolog. *BMC Biotech.*, **7**, 71.
58. Hattrup,C.L. and Gendler,S.J. (2006) MUC1 alters oncogenic events and transcription in human breast cancer cells. *Breast Cancer Res.*, **8**, R37.
59. Hydbring,P., Malumbres,M. and Sicinski,P. (2016) Non-canonical functions of cell cycle cyclins and cyclin-dependent kinases. *Nature reviews. Mol. Cell Biol.*, **17**, 280–292.
60. Grillo,M., Bott,M.J., Khandke,N., McGinnis,J.P., Miranda,M., Meyyappan,M., Rosfjord,E.C. and Rabindran,S.K. (2006) Validation of cyclin D1/CDK4 as an anticancer drug target in MCF-7 breast cancer cells: Effect of regulated overexpression of cyclin D1 and siRNA-mediated inhibition of endogenous cyclin D1 and CDK4 expression. *Breast Cancer Res. Treat.*, **95**, 185–194.
61. Sappino,A.P., Skalli,O., Jackson,B., Schürch,W. and Gabbiani,G. (1988) Smooth-muscle differentiation in stromal cells of malignant and non-malignant breast tissues. *Int. J. Cancer*, **41**, 707–712.
62. Shiga,K., Hara,M., Nagasaki,T., Sato,T., Takahashi,H. and Takeyama,H. (2015) Cancer-Associated fibroblasts: Their Characteristics and Their Roles in Tumor Growth. *Cancers*, **7**, 2443–2458.
63. Erdogan,B., Ao,M., White,L.M., Means,A.L., Brewer,B.M., Yang,L., Washington,M.K., Shi,C., Franco,O.E., Weaver,A.M. *et al.* (2017) Cancer-associated fibroblasts promote directional cancer cell migration by aligning fibronectin. *J. Cell Biol.*, **216**, 3799–3816.
64. Tiwari,N., Tiwari,V.K., Waldmeier,L., Balwierz,P.J., Arnold,P., Pachkov,M., Meyer-Schaller,N., Schübeler,D., van Nimwegen,E. and Christofori,G. (2013) Sox4 is a master regulator of epithelial-mesenchymal transition by controlling Ezh2 expression and epigenetic reprogramming. *Cancer Cell*, **23**, 768–783.
65. Martin,M., Ancey,P.B., Cros,M.P., Durand,G., Le Calvez-Kelm,F., Hernandez-Vargas,H. and Herceg,Z. (2014) Dynamic imbalance between cancer cell subpopulations induced by transforming growth factor beta (TGF-β) is associated with a DNA methylome switch. *BMC Genomics*, **15**, 435.
66. Cardenas,H., Vieth,E., Lee,J., Segar,M., Liu,Y., Nephew,K.P. and Matei,D. (2014) TGF-β induces global changes in DNA methylation during the epithelial-to-mesenchymal transition in ovarian cancer cells. *Epigenetics*, **9**, 1461–1472.
67. Brechbuhl,H.M., Finlay-Schultz,J., Yamamoto,T.M., Gillen,A.E., Cittelly,D.M., Tan,A.C., Sams,S.B., Pillai,M.M., Elias,A.D., Robinson,W.A. *et al.* (2017) Fibroblast subtypes regulate responsiveness of luminal breast cancer to estrogen. *Clin. Cancer Res.*, **23**, 1710–1721.
68. Hwang,R.F., Moore,T., Arumugam,T., Ramachandran,V., Amos,K.D., Rivera,A., Ji,B., Evans,D.B. and Logsdon,C.D. (2008) Cancer-associated stromal fibroblasts promote pancreatic tumor progression. *Cancer Res.*, **68**, 918–926.
69. Müerköster,S., Wegehenkel,K., Arlt,A., Witt,M., Sipos,B., Kruse,M.L., Sebens,T., Klöppel,G., Kalthoff,H., Fölsch,U.R. *et al.* (2004) Tumor stroma interactions induce chemoresistance in pancreatic ductal carcinoma cells involving increased secretion and paracrine effects of nitric oxide and interleukin-1beta. *Cancer Res.*, **64**, 1331–1337.
70. Marsh,T., Pietras,K. and McAllister,S.S. (2013) Fibroblasts as architects of cancer pathogenesis. *Biochim. Biophys. Acta*, **1832**, 1070–1078.