Cell type-specific DNA methylation patterns in the human breast

Noga Bloushtain-Qimron[a,b], Jun Yao[a,c], Eric L. Snyder[a,c], Michail Shipitsin[a,c], Lauren L. Campbell[a,d], Sendurai A. Mani[e], Min Hu[a,c], Haiyan Chen[f,g], Vadim Ustyansky[h], Jessica E. Antosiewicz[i], Pedram Argani[b], Marc K. Halushka[j], James A. Thomson[i], Paul Pharoah[k], Angel Porgador[b], Saraswati Sukumar[j], Ramon Parsons[l], Andrea L. Richardson[m,n], Martha R. Stampfer[o], Rebecca S. Gelman[f,g], Tatiana Nikolskaya[h,p], Yuri Nikolsky[h], and Kornelia Polyak[a,c,d,q]

Departments of [a]Medical Oncology and [f]Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02115; Departments of [c]Medicine, [m]Pathology, and [d]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115; [n]Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115; [g]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115; [b]Department of Microbiology and Immunology, Faculty of Health Sciences, and the Cancer Research Center, Ben Gurion University of the Negev, Beer-Sheva 84105, Israel; [e]Whitehead Institute for Biomedical Research, Cambridge, MA 02142; [h]GeneGo, Inc., St. Joseph, MI 49085; [i]University of Wisconsin-Madison Medical School, Madison, WI 53706; [j]Johns Hopkins University School of Medicine, Baltimore, MD 21231; [k]Cancer Research UK, Cambridge CB2 0RE, United Kingdom; [l]Department of Medicine, Columbia University, New York, NY 10032; [o]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; and [p]Vavilov Institute for General Genetics, Russian Academy of Sciences, 117809 Moscow, Russia

Edited by Charles R. Cantor, Sequenom, Inc., San Diego, CA, and approved July 24, 2008 (received for review May 28, 2008)

Cellular identity and differentiation are determined by epigenetic programs. The characteristics of these programs in normal human mammary epithelium and their similarity to those in stem cells are unknown. To begin investigating these issues, we analyzed the DNA methylation and gene expression profiles of distinct subpopulations of mammary epithelial cells by using MSDK (methylation-specific digital karyotyping) and SAGE (serial analysis of gene expression). We identified discrete cell-type and differentiation state-specific DNA methylation and gene expression patterns that were maintained in a subset of breast carcinomas and correlated with clinically relevant tumor subtypes. CD44+ cells were the most hypomethylated and highly expressed several transcription factors with known stem cell function including HOXA10 and TCF3. Many of these genes were also hypomethylated in BMP4-treated compared with undifferentiated human embryonic stem (ES) cells that we analyzed by MSDK for comparison. Further highlighting the similarity of epigenetic programs of embryonic and mammary epithelial cells, genes highly expressed in CD44+ relative to more differentiated CD24+ cells were significantly enriched for Suz12 targets in ES cells. The expression of FOXC1, one of the transcription factors hypomethylated and highly expressed in CD44+ cells, induced a progenitor-like phenotype in differentiated mammary epithelial cells. These data suggest that epigenetically controlled transcription factors play a key role in regulating mammary epithelial cell phenotypes and imply similarities among epigenetic programs that define progenitor cell characteristics.

cancer | differentiation | progenitor | stem cell

Stem cells are defined as cells with both self-renewal capacity and the ability to give rise to multiple distinct differentiated cell types. Recent studies have demonstrated the existence of cells with these properties in normal human breast epithelium (1). However, the identity, molecular characteristics, and location of these cells are poorly defined. By using *in vitro* clonogenicity assays, several candidate human mammary epithelial progenitors have been identified, and numerous cell surface markers have been proposed for their enrichment, including MUC1, CD10, CD44, and ITGA6 (1–3). These cells are thought to be restricted to the basal layer of terminal end ducts, and known stem cell pathways are activated in mammosphere cultures used to enrich for putative mammary stem cells (4).

By using mouse mammary fat pad injection assays, lineage$^-$/CD24$^{-/low}$/CD44$^+$ ("CD44+") breast tumor cells were found to be more tumorigenic than more differentiated CD44$^-$/CD24$^+$ ("CD24+") cells, identifying the CD44+ cells as human breast "cancer stem cells" (5). We previously isolated these cells from primary breast tumors and similar cells from normal breast tissue and determined that their comprehensive gene expression profiles were consistent with the hypothesis that CD24+ and CD44+ cells represent differentiated luminal epithelial and progenitor-like cells, respectively (6).

Epigenetic programs, including DNA methylation and chromatin patterns, play a key role in ES cell function and differentiation (7, 8). Several genes important in pluripotency and self-renewal are hypomethylated and expressed in stem cells and silenced by methylation in differentiated cells (9), suggesting that other genes that control stem cell characteristics, such as transcription factors, may be epigenetically regulated. The identity of mammary epithelial progenitor cell-specific epigenetic programs, their relatedness to those in ES cells and breast carcinomas, and which genes they regulate are undefined. To begin investigating these issues, we analyzed the comprehensive DNA methylation profiles of four distinct cell populations from normal human breast tissue, including CD44+ and CD24+ cells, and demonstrated that epigenetically controlled transcription factors seem to help define progenitor and differentiated cell phenotypes. Furthermore, we found that genes encoding transcription factors with known stem cell function are similarly methylated in CD44+ cells from normal mammary epithelium and some breast carcinomas and in BMP4-treated ES cells. These findings imply conservation of epigenetic programs that define progenitor characteristics.

Results and Discussion

Characterization of Distinct Mammary Epithelial Cell Populations. To characterize the cellular composition of normal human mam-

Author contributions: N.B.-Q. and K.P. designed research; N.B.-Q., J.Y., E.L.S., M.S., L.L.C., and J.E.A. performed research; M.S., S.A.M., M.H., J.E.A., P.A., M.K.H., J.A.T., P.P., S.S., R.P., A.L.R., M.R.S., and Y.N. contributed new reagents/analytic tools; N.B.-Q., J.Y., E.L.S., L.L.C., M.H., H.C., V.U., R.S.G., T.N., Y.N., and K.P. analyzed data; N.B.-Q., J.Y., E.L.S., L.L.C., H.C., A.P., R.S.G., T.N., Y.N., and K.P. wrote the paper.

Conflict of interest statement: K.P. receives research support from and is a consultant to Novartis Pharmaceuticals, Inc. K.P. also receives research support from Biogen-Idec and is a consultant to Aveo Pharmaceuticals, Inc. and GeneGo Inc. M.H., N.B.Q., and K.P. submitted a patent application on the MSDK method and findings described in this manuscript in accordance with the policies of DFCI. V.U., T.N., and Y.N. are employees of Genego Inc.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the SAGE Genie database, http://cgap.nci.nih.gov/SAGE.

[q]To whom correspondence may be addressed: Kornelia Polyak, Dana-Farber Cancer Institute, 44 Binney Street D740C, Boston, MA 02115. E-mail: kornelia_polyak@dfci.harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0805206105/DCSupplemental.

mary epithelium, we performed FACS analyses by using cell surface markers previously associated with luminal epithelial (MUC1 and CD24), myoepithelial (CD10), and progenitor (CD44) cell phenotypes [supporting information (SI) Fig. S1*A*]. Then, we designed an immuno-magnetic bead purification procedure that minimized the overlap among cell fractions (Fig. S1*B*). By using this procedure, we isolated CD44+, CD24+, MUC1+, and CD10+ cells from normal breast tissue of healthy women. Each cell fraction was isolated and characterized from multiple independent cases [see Tables S1 and S2]. The phenotypes of the enriched cell fractions were initially assessed by analyzing the expression of known differentiated and progenitor cell markers by semiquantitative RT-PCR (Fig. S1*C* and data not shown). CD24+ cells expressed luminal cell markers but were devoid of myoepithelial and progenitor cell markers, whereas CD44+ cells lacked lineage-specific genes and expressed multiple progenitor cell markers. MUC1+ cells were positive for luminal and some progenitor markers, implying that they may represent luminal lineage-committed progenitors or a mix of differentiated and progenitor cells. Similarly, CD10+ cells expressed myoepithelial and some progenitor markers.

To define the differences among the four cell types in further detail, we analyzed their gene expression profiles by using SAGE (10). Hierarchical clustering of the SAGE libraries by using tags differentially expressed among the four cell types effectively separated the samples into two major basal/myoepithelial (CD44+ and CD10+ cells) and luminal (CD24+ and MUC1+ cells) branches that further subdivided into four branches by cell type (Fig. 1*A*). Notably, known progenitor and differentiated cell markers were more abundant in the expected cell populations (Table S3). A subset of genes in the "basal/progenitor" cluster was highly expressed in CD44+ cells, and it included several genes with known developmental and stem-cell function (e.g., MSC, BRD2, and ELF1), whereas a set of genes was common between CD10+ and CD44+ cells. These results confirmed our prior study of CD24+ and CD44+ cells (6) and determined that MUC1+ and CD10+ cells are distinct subsets of luminal epithelial and myoepithelial cells, respectively.

To define the functional differences among the four cell types, we classified the genes differentially expressed between CD44+ and either one of the other three cell types by using gene ontology (GO) terms. These analyses revealed that, compared with all three more differentiated cell types, CD44+ cells were enriched for genes encoding for proteins with extracellular function and roles in development and differentiation (Fig. 1*B* and Table S4). Proteins specifically secreted by CD44+ cells included several chemokines (CCL2, CXCL2, and CXCL14), proteases (MMP2, MMP3, and MMP9), protease inhibitors (TIMP1–3), and cytokines involved in stem-cell signaling pathways such as TGFb (INHBA, BMP2, DCN, and LTBP4), WNT (SFRP and SFRP4), and Hedgehog (BGN). The enrichment of these genes in progenitor-like CD44+ cells compared with more differentiated CD24+, MUC1+, and CD10+ cells were consistent with the presumed *in vivo* function of these four cell populations.

To correlate the molecular profiles of the four cell types with their differentiation capacity, we performed colony growth assays in various culture conditions (Table S5). Colonies derived from CD10+ cells were homogenously positive for the basal/myoepithelial markers cytokeratin 14 (CK14), smooth muscle actin (SMA), and vimentin (VIM), whereas MUC1+ and CD24+ cells formed uniform colonies positive for the luminal marker cytokeratin 18 (CK18) and negative for basal/myoepithelial markers (CK14, VIM, and SMA) (Fig. 1*C*). In contrast, CD44+ cells gave rise to different types of colonies depending on the media used. In media 1, most of the colonies were mesenchymal-like, CK14+, VIM+, CK18−, and SMA−, whereas the majority of the colonies were epithelium-like and

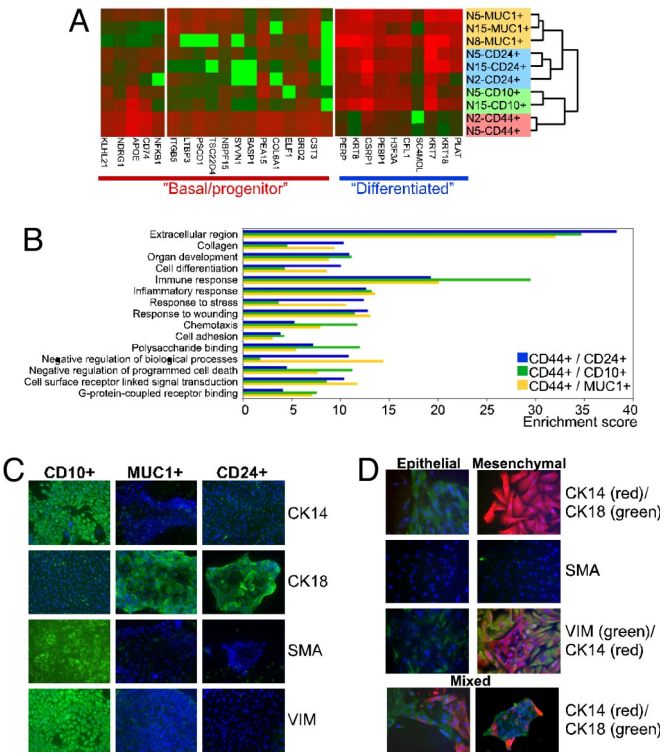

Fig. 1. Enrichment and phenotype of assorted mammary epithelial cells. (*A*) Dendrogram depicting relatedness of SAGE libraries prepared from CD44+, CD10+, MUC1+, and CD24+ cells from multiple independent cases. Selected portions of the clustering heat map are shown representing genes enriched in luminal (CD24+ and MUC1+) and basal/progenitor-like (CD44+ and CD10+) cells. Each row represents a tag and is labeled with the symbol of the gene that best matches that tag. Red and green indicate high and low expression levels, respectively. (*B*) Gene ontology enrichment analysis of genes differentially expressed between CD44+ cells and each of the other three cell types. Ontology terms for genes highly expressed in CD44+ relative to CD24+ (blue), CD10+ (green), and MUC1+ (yellow) cells were compared with those for the rest of genes expressed in CD44+ cells by using the Fisher Exact test. Significantly enriched ontology terms are shown with relative enrichment fold as negative log of P value. Immunofluorescence analyses of cell type-specific markers in representative colonies of other (*C*) and CD44+ (*D*) cells in 2D culture conditions.

CK18+ in media 3 (Fig. 1*D*). Media 1 and 3 had different additives (e.g., growth factors, hormones, and antioxidants), and media 3 was similar to the WIT medium (11). In both media, a subset of the colonies was composed of a mix of CK18+ and CK14+ cells (Fig. 1*D*), indicating that CD44+ cells can give rise to both luminal and myoepithelial cells. Thus, results of the colony growth assays correlated well with the *in vivo* expression patterns and presumed progenitor and differentiated-cell characteristics of the four cell types.

Cell Type-Specific DNA Methylation Patterns. Next, we analyzed the comprehensive DNA methylation profiles of each of the above-described cell types by using MSDK, a comprehensive DNA methylation profiling technology previously developed in our lab (12). By using a combination of methylation-sensitive (e.g., *Asc*I) and nonsensitive (e.g., *Nla*III) restriction enzymes, we derive short sequence tags from the genome. The number of these tags reflects the methylation status of each recognition site of the methylation-sensitive enzyme. Because we used AscI, our analysis is limited to its recognition sites. Nonetheless, we analyzed >5,000 unique tags corresponding to >4,000 genes from each sample, representing ≈ 25% of all human coding genes associated with CpG islands (Table S6 and Fig. S2 *A* and *B*).

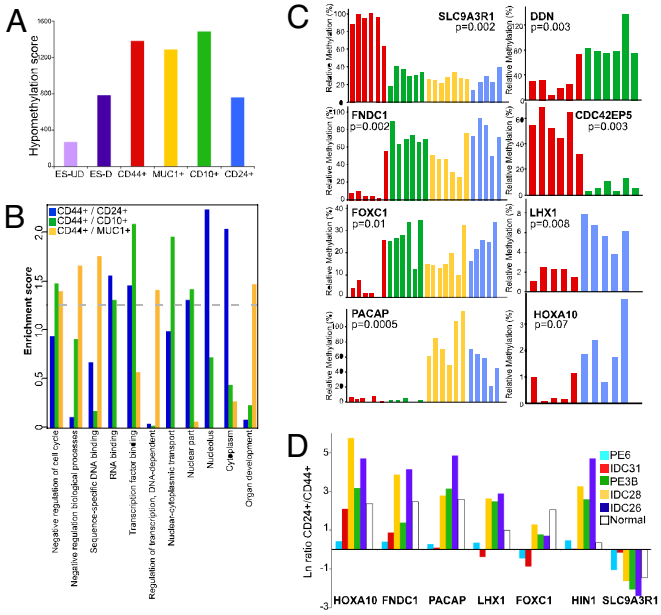

Fig. 2. Cell type-specific DNA methylation patterns and their functional significance. (*A*) Arbitrary hypomethylation score of each cell type analyzed by MSDK. (*B*) Gene ontology enrichment analysis of genes differentially methylated between CD44+ cells and each of the other three cell types. Ontology terms for genes hypomethylated in CD44+ relative to CD24+ (blue bars), CD10+ (green bars), and MUC1+ (yellow bars) cells were compared with those for the rest of genes expressed in CD44+ cells by using the Fisher Exact test. Significantly enriched ontology terms are shown with relative enrichment fold as negative log of *P* value. The dotted gray line marks statistical significance (1.3). (*C*) Validation of MSDK results by qMSP for selected genes by using bisulfite-treated DNA from CD44+ (red bars), CD10+ (green bars), MUC1+ (yellow bars), and CD24+ (blue bars) cells. Each column represents a different individual. *y*-axis indicates relative methylation levels normalized to ACTB. *P* values indicate the statistical significance of the observed differences in DNA methylation among the cell types. (*D*) qMSP analysis of the methylation of the indicated genes in CD44+ and CD24+ cells isolated from breast tumors and normal breast. The *y*-axis indicates the ln ratio of qMSP values in CD24+ and CD44+ cells. PE, pleural effusion, IDC, invasive ductal carcinoma. Colors indicate different samples.

To identify AscI sites that were statistically significantly ($P < 0.05$) differentially methylated among the four different cell types, we performed pairwise and combined comparisons of MSDK libraries (Fig. S3 and Table S7). These comparisons suggested that CD44+ cells are hypomethylated compared with each of the other three cell types.

To identify similarities between the DNA methylation profiles of ES and mammary epithelial cells, we performed MSDK analysis of undifferentiated (ES-UD) and BMP4-treated (ES-D) human embryonic stem cells (Fig. S1*D*) (13, 14). Overall, undifferentiated ES cells were hypermethylated compared with BMP-4 treated ES cells, and interestingly, several loci hypomethylated in ES-UD compared with ES-D cells were also hypomethylated in CD44+ cells relative to CD24+ cells (Table S8).

To further examine whether overall degree of DNA methylation correlates with cellular differentiation status, we calculated an arbitrary hypomethylation score for each sample defined by the abundance of MSDK tags corresponding to each AscI site present in each library. Based on this score, among mammary epithelial cells CD44+, MUC1+, and CD10+ cells were hypomethylated compared with CD24+ cells, whereas undifferentiated ES cells were more methylated compared with BMP-4-treated ES cells (Fig. 2*A*). The observed differences in the overall degree of DNA methylation were not because of cell type-specific differences in DNMT (DNA methyl-transferase) expression, because this was fairly constant among samples (data not shown). These results suggest that DNA methylation profiles may be used as markers of cellular differentiation states. However, in the mammary epithelium more differentiated cells may be more methylated, whereas in ES cells differentiation may correlate with hypomethylation.

To determine which gene categories may be regulated by DNA methylation, we classified genes differentially methylated between CD44+ and each of the other three cell types by using gene ontology (GO) terms. We found that genes hypomethylated in CD44+ cells were highly enriched for transcription-related functions and involved in the regulation of cell proliferation and differentiation (Fig. 2*B* and Table S4). In particular, several genes hypomethylated in CD44+ cells compared with CD24+ cells encode homeobox and polycomb proteins known to regulate stem cell function (Table 1). HOXA10 is a target of the MLL-AF9 oncogene and is required for the self-renewal of hematopoietic stem cells (15). Similarly, TCF7L1 (TCF3) is an upstream regulator of a transcription program involving OCT4, SOX2, and Nanog that is required for the pluripotency and self-renewal of ES cells (16). These data are consistent with the hypotheses that CD44+ cells include mammary epithelial progenitors and that their phenotype and differentiation, at least in part, are epigenetically regulated. Furthermore, our data also suggest that pluripotency and self-renewal may be regulated by an overlapping set of transcription factors in different stem cells.

To confirm the MSDK results for selected genes, we performed sequencing of bisulfite-treated genomic DNA and quantitative methylation-specific PCR (qMSP). Tested genes included FNDC1 and FOXC1 (hypomethylated in CD44+ cells compared with the other three cell types), PACAP (hypomethylated in CD44+ and CD10+ cells compared with CD24+ and MUC1+ cells), SLC9A3R1 (hypermethylated in CD44+ cells compared with the other three cell types), DDN (hypomethylated in CD44+ cells compared with CD10+ cells), DTX1 and CDC42EP5 (hypomethylated in CD10+ cells compared with CD44+ cells), LHX1, HOXA10, and SCGB3A1 (hypomethylated in CD44+ cells compared with CD24+ cells), and SOX13 (hypomethylated in CD10+ cells compared with MUC1+ and CD24+ cells). Not all genes were analyzed by both methods. The sequencing and qMSP results confirmed the cell type-specific methylation patterns demonstrated that these were consistent among samples derived from women of different ages (18–58 years old) and reproductive histories, although some variability in the degree of methylation was observed (Fig. 2*C* and Fig. S4).

Conservation of Cell Type-Specific DNA Methylation Patterns in Breast Carcinomas. Our recent analysis of gene expression profiles of CD24+ and CD44+ cells from normal and neoplastic breast tissue revealed a high degree of similarity between analogous cell types (6). To determine whether genes differentially methylated between normal CD44+ and CD24+ cells are also differentially methylated in analogous cell types from neoplastic tissue, we performed qMSP analyses of CD44+ and CD24+ cells isolated from five different breast tumors. The methylation of several genes was the same in tumor CD44+ and CD24+ cells as in their normal counterparts, although tumors were more variable than normal samples (Fig. 2*D* and Fig. S5). Interestingly, the two tumors that were the most dissimilar to normal CD24+ and CD44+ cells (IDC31 and PE6) were both HER2+, whereas all others were ER+, PR+, and HER2−. This suggests that the epigenetic profiles of progenitor-like cells in different tumor subtypes are distinct, potentially because of tumor-specific transforming events.

To further test the hypothesis that the epigenetic profiles defining progenitor-like cells are distinct in different breast cancer subtypes, we analyzed the methylation of PACAP, FOXC1, SLC9A3R1, and HOXA10 in a larger set (>100 cases)

Table 1. Selected genes with known role in development differentially methylated between CD44+ and CD24+ normal mammary epithelial cells

Tag	CD24+	CD44+	Pval	Ratio	Chr	Gene	Distance	Position	Strand	Function
CACAGCCAGCCTCCCAG	0	39	0.000001	22	17	LHX1	+3696	inside	+	Homeobox gene
TATTTGCCAAGTTGTAC	0	14	0.00205972	8	7	HOXA10	−4360	upstream	–	Homeobox gene
TATTTGCCAAGTTGTAC	0	14	0.00205972	8	7	HOXA11	+627	inside	–	Homeobox gene
TCGCCGGGCGCTTGCCC	7	66	9.33E-08	5	5	PITX1	+6168	inside	–	Homeobox gene
ACAATAGCGCGATCGAG	2	14	0.0178476	4	16	IRX5	−460	upstream	+	Homeobox gene
TTAAGAGGGCCCCGGGG	0	7	0.0241671	4	14	NKX2–8	+1823	inside	–	Homeobox gene
AGCCCTCGGGTGATGAG	5	24	0.0106181	3	1	LMX1A	−747	upstream	–	Homeobox gene
CCCCGTTTTTGTGAGTG	6	22	0.0355276	2	17	HOXB9	−20615	upstream	–	Homeobox gene
CAGCCAGCTTTCTGCCC	20	56	0.0169362	2	9	LHX3	−141	upstream	–	Homeobox gene
CCCCAGGCCGGGTGTCC	9	37	0.0070473	2	17	CBX8	−16725	upstream	–	Polycomb protein
ACCCGCACCATCCCGGG	46	140	5.96E-06	2	17	CBX4	−4595	upstream	–	Polycomb protein
CACCAAACCTAGAAGGC	10	33	0.0383201	2	2	GLI2	−56233	upstream	+	Shh pathway
ACCCTGAAAGCCTAGCC	3	24	0.00179963	4	21	ITGB2	−10800	upstream	–	Stem cell marker
TGGTTTACCTTGGCATA	0	13	0.009773	7	6	FOXF2	−6378	upstream	+	Development
GTCCTTGTTCCCATAGG	0	35	0.000002	19	6	FOXC1	−5061	upstream	+	Development
CCCCCGCGACGCGGCGG	0	20	0.000800427	11	1	SOX13	−576	upstream	+	Development
CACTCCACGTTTATAGA	0	7	0.0241671	4	4	SMAD1	−783	upstream	+	BMP signaling
CCCCAGCTCGGCGGCGG	44	113	0.00118262	1	2	TCF7L1	+854	inside	+	WNT pathway

MSDK tag sequence, raw tag numbers in the indicated libraries (CD24+ and CD44+), *p*-value of the statistical significance of the observed difference in tag numbers between libraries, ratio of normalized tag numbers, chromosomal location, gene symbol, distance to predicted transcription start site, position of the *Asc*I site relative to the gene, DNA strand, and gene function are indicated.

of invasive breast carcinomas from sporadic cases and from germ line BRCA1 and BRCA2 mutation carriers. These four genes were selected because they were the most consistently differentially methylated between CD44+ and CD24+ cells in the normal breast. Methylation of PACAP, FOXC1, and HOXA10 was statistically significantly associated with PR, ER, and HER2 status and correlated with luminal (ER+/PR+), HER2+, and basal-like (ER−/PR−/HER2−) breast tumor subtypes (Table S9). Thus, the methylation pattern of luminal tumors resembled that of normal CD24+ cells, whereas HER2+ and basal-like tumors were more hypomethylated and similar to CD44+ cells.

We previously demonstrated that a CD44+ breast cancer cell gene expression signature correlates with shorter distant metastasis-free survival (6). To investigate whether this association is also true for cell type-specifically methylated genes, we analyzed the expression of five such genes in two independent cohorts of patients with clinical follow-up and microarray data (Fig. S6*A*). In both datasets, patients with CD44+ cell-like tumors had statistically significantly shorter distant metastasis-free survival than patients with CD24+ cell-like tumors (Fig. S6*A*).

To determine whether the epigenetic profiles of cancer cells are maintained during tumor progression, we analyzed the methylation profiles of matched primary tumors and distant metastases located in different organs in four independent patients. The methylation of HOXA10, FOXC1, and LHX1 (hypermethylated in CD24+ cells) was higher in distant metastases compared with primary tumors, whereas the methylation of PACAP and SLC9A3R1 showed more variability (Fig. S6*B* and data not shown). These results suggest that distant metastases are enriched for hypermethylated CD24+ breast cancer cells, confirming our prior immunohistochemical analyses (6). However, because almost all of the patients analyzed had ER+/PR+/HER2− primary invasive breast carcinomas, our findings could be specific for this tumor subtype.

Associations Between DNA Methylation and Gene Expression Patterns. To determine the effect of methylation differences on gene expression patterns, we performed quantitative RT-PCR (qRT-PCR) of the same cells that were analyzed by qMSP (Fig. 3*A*). Methylation of CpG islands in the promoter area or within the gene in general led to decreased expression (FNDC1, DDN, LHX1, HOXA10, FOXC1, and SOX13), whereas methylation at certain upstream (PACAP) or downstream (CDC42EP5) sites was associated with higher expression. Increased DNA methylation may lead to increased expression because of the inhibition of binding of silencers such as CTCF and BORIS (17). Consistent with this hypothesis, the differentially methylated regions of PACAP and CDC42EP5 contain predicted CTCF/BORIS binding sites.

To obtain additional evidence to support our hypothesis that CD44+ cells include cells with progenitor properties and to further explore similarities of epigenetic programs of embryonic and putative adult stem cells, we analyzed whether genes differentially expressed between CD44+ and CD24+ cells are enriched for Suz12 targets. Suz12 is a member of PRC2 (Polycomb-Repressive Complex 2) associated with H3K27-containing nucleosomes, and it is essential for mouse embryonic development and ES cell differentiation (8, 18). Targets of Suz12 have been identified based on genome-wide ChIP studies in ES cells, and many of them encode for proteins required for pluripotency and self-renewal (19, 20). When we compared genes differentially expressed between CD44+ and CD24+ cells to Suz12 targets in ES cells, we found a statistically significant enrichment for Suz12 targets in genes highly expressed in CD44+ cells (Fig. 3*B* and Table S10). Similar observations were made when analyzing genes differentially expressed between CD44+ cells and the other two (CD10+ and MUC1+) cell types. Thus, CD44+ cells include cells with stem cell characteristics that seem to be defined by the same genes regardless of tissue type.

To further strengthen the association between the genes we identified as hypomethylated and expressed in CD44+ cells and mammary epithelial progenitors, we performed immunohistochemical analysis of four such genes (HOXA10, SOX13, HOXA11, and MSC) in normal human breast tissue. Occasional cells localized in the basal layer of the ducts and alveoli expressed these genes, consistent with the presumed location of mammary epithelial progenitors (Fig. 3*C*). Double immunohistochemical analyses demonstrated that these cells were also positive for CD44v6, an epithelial-specific isoform of the marker used for their enrichment (Fig. 3*D*). These results support our hypothesis

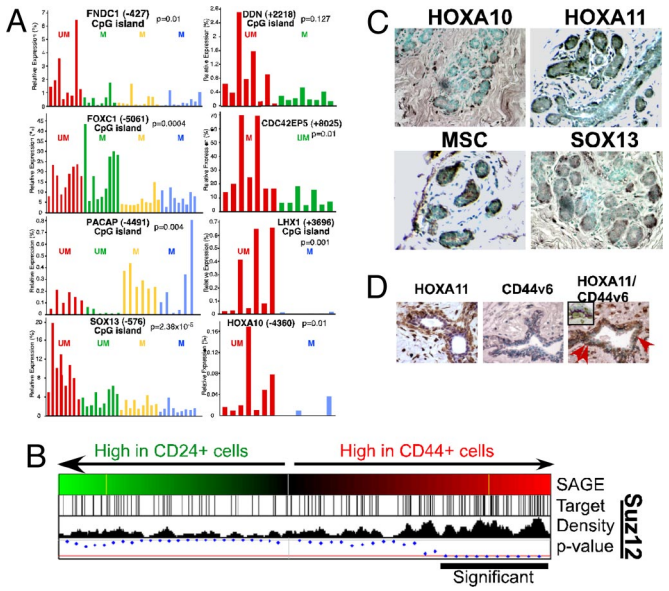

Fig. 3. Expression of differentially methylated genes. (*A*) qRT-PCR analysis of the indicated genes in the same set of cells as used for methylation studies. Colors denote cell types as described in Fig. 2. Relative expression levels normalized to RPL19 are indicated on the *y*-axis. ''U'' (unmethylated) and ''M'' (methylated) indicate the methylation status of the gene in each of the cell types, and *p*-values indicate the statistical significance of the observed difference in expression levels among the cell types analyzed. Numbers next to gene names indicate *Asc*I site location relative to transcription start site. ''CpG island'' indicates that the *Asc*I site analyzed is within a predicted CpG island. (*B*) Relationships between genes highly expressed in CD44+ compared with CD24+ cells and Suz12 targets. SAGE expression ratios are represented by shades of red and green depicting up and down-regulated genes, respectively. Yellow bars mark threshold ratios (2-fold difference). Genes are ordered according to their fold difference. Suz12 target genes are indicated by black bars with right-hand panels showing their density and corresponding *p*-values along the gene list. (*C*) Immunohistochemical analysis of the expression of the indicated genes in normal human breast tissue. (*D*) Dual immunohistochemical staining for HOXA11 and CD44v6 expression to demonstrate coexpression of the two markers in the same cells (red arrows). Inset shows higher magnification image.

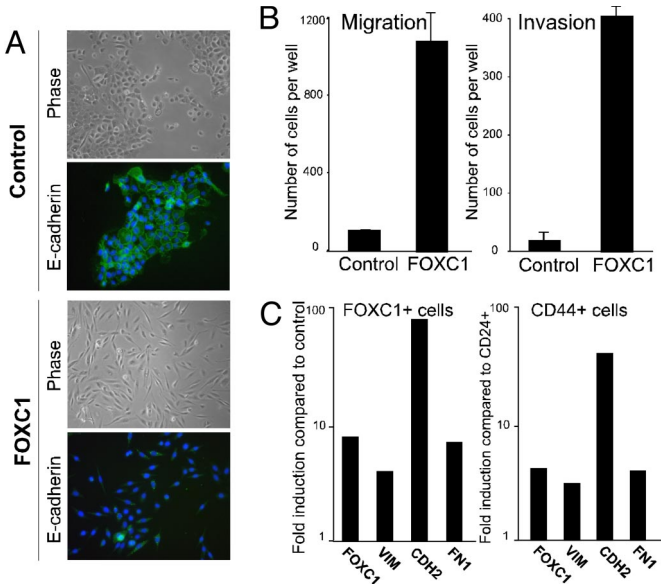

Fig. 4. The role of FOXC1. (*A*) Morphology of MCF12A mammary epithelial cells and expression of E-cadherin after infection with control or FOXC1-expressing retrovirus. (*B*) FOXC1 expression increases cell migration and invasion in MCF12A cells. (*C*) qRT-PCR analysis of the indicated genes in FOXC1-expressing MCF12A cells and in normal CD44+ progenitor-like cells. The *y*-axis indicates fold induction of the indicated genes in FOXC1+ and CD44+ cells compared with control and CD24+ cells, respectively.

that genes encoding transcription factors are candidates for epigenetically controlled regulators of the mammary epithelial progenitor cell phenotype because their methylation and expression patterns were consistently CD44+ cell -specific among samples and their *in vivo* expression correlated with the putative location of mammary epithelial progenitors.

Systemic Network Analysis of Cell Type-Specific Gene Expression and DNA Methylation Patterns. To further analyze signaling pathways and gene interaction networks that may play an important role in determining stem and differentiated mammary epithelial cell phenotypes, MSDK and SAGE data of the four distinct cell types from normal breast were subjected to functional analysis by using Metacore (21). First, we determined that genes differentially methylated or expressed in the four cell types were mapped to the same pathways and cellular processes (Figs. S7–S9 and Table S11 and Table S12). Next, we analyzed both methylation and expression data for individual cell type-specific differences in signaling pathways and networks (Figs. S10 and S11). Genes highly expressed in CD24+ cells were enriched for insulin-regulated pathways, mitochondrial metabolism, and apoptosis, whereas genes highly expressed in CD44+ cells were enriched for ECM (extracellular matrix) and cytoskeleton remodeling, integrin-mediated cell adhesion, immune response processes, and IL-4-mediated signaling (Table S13). The top-scoring ANR network for CD24+ cells included several DNA damage checkpoint genes (e.g., p53, Chk1, Chk2, and ATM), whereas MYC, AR, and TGF-β/SMAD signaling predominated in CD44+ cells (Figs. S10 and S11). These results were consistent with the progenitor-like and luminal epithelial phenotypes of CD44+ and CD24+ cells, respectively.

FOXC1, a Candidate Regulator of Mammary Epithelial Progenitor Cell Function. To further test the hypothesis that the transcription factors we identified as hypomethylated and highly expressed in CD44+ cells compared with the other three cell types may play a role in progenitor cell function, we determined the effect of the expression of FOXC1 in differentiated mammary epithelial cells. FOXC1 was a top candidate progenitor cell phenotype regulator because the FOXC1 interaction network was enriched for a number of other genes differentially methylated between CD24+ and CD44+ cells (e.g., IRX5, ROR-alpha, and BAI1) and included several key pathways regulating progenitor cell function, including FGF, TGF-β, Notch, and WNT signaling (Fig. S12 *A* and *B*). Genes with the most extensive network in a particular cell have been shown to be essential regulators of cellular phenotypes (22). FOXC1 was also the most hypomethylated and highly expressed in CD44+ cells among all cell types analyzed. Furthermore, FOXC1 has been shown to play an essential role in development, and its expression is regulated by Hedgehog and TGF-β, both of which are important regulators of stem cell function (23). Stable expression of FOXC1 in MCF12A cells resulted in the conversion of the differentiated epithelial phenotype to a CD44+ cell-like mesenchymal phenotype as determined by morphologic changes, increased cell migration and invasion, and altered gene expression patterns (Fig. 4 *A–C*). Further studies are required to define the function of FOXC1 in the mammary gland, but this finding indicates that at least some of the cell type-specifically methylated genes we identified, particularly transcription factors, may play a role in the regulation of mammary epithelial progenitor and differentiated cell phenotypes.

In summary, by using a combination of approaches, we identified candidate regulators of human mammary epithelial progenitors, markers that can be used to purify different mammary epithelial cell populations, and culture conditions for four cell types. Furthermore, we found cell type-specific DNA methylation programs that are maintained in breast carcinomas and correlate with tumor subtypes and clinical outcome. These data are valuable for the further characterization of the mammary epithelial cell hierarchy and understanding the regulatory pathways determining their phenotypes.

Materials and Methods

Detailed description of the procedures is in *SI Text*.

Tissue Samples and Cell Culture. Tissue specimens were collected without patient identifiers by using protocols approved by the Institutional Review Boards. Fresh samples were processed and distinct cell populations were isolated as described in ref. 6. Details of tissue samples and culture conditions tested are described in Tables S1 and Table S5, respectively.

Generation and Analyses of SAGE and MSDK Data. SAGE and MSDK libraries were generated and analyzed as described previously by using the most recent human genome sequence information (6, 12).

Pathway Map and Network Analyses By Using Metacore. Analysis was conducted in accordance with the Metacore analytical suite version 4.2 (GeneGo, Inc. St. Joseph, MI) manual and has been described previously (6, 21, 24).

Bisulfite Sequencing, qMSP, and RT-PCR. Genomic DNA was bisulfite-treated and purified; qMSP and RT-PCR amplifications and sequencing were performed as previously described (12, 25). A list of all primers used is available from the authors on request. Statistical significance of the differences in qMSP values among the various cell types and associations between qMSP and expression levels were calculated by using the Kruskal-Wallis test.

Immunohistochemistry. Immunohistochemical analysis for single markers was performed essentially as described in ref. 26 by using rabbit polyclonal HOXA10 (gift of Dr. Honami Naora, MD Anderson Cancer Center, Houston, TX), SOX13 (Sigma), HOX11 (Abnova), and MSC (Santa Cruz) antibodies.

DNA Constructs and FOXC1 Expression. Human FOXC1 cDNA was cloned into pWZL-blasto retroviral construct and used for retrovirus generation by using standard procedures. MCF-12A cells were infected with control or FOXC1-expressing retrovirus, and cells were selected in blasticidin-containing medium.

ACKNOWLEDGMENTS. We greatly appreciate the help of Fiona Blows with the acquisition of human tissue samples and Dyane Bailey for assistance with immunohistochemistry. We thank Dr. Honami Naora (MD Anderson Cancer Center, Houston, TX) for her generous gift of the HOXA10 antibody; the Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, Canada, and the Cancer Genome Anatomy Project for MSDK and SAGE library sequencing, respectively. This work was supported by the National Institutes of Health Grants CA89393, CA94074, and CA116235 (to K.P.), U.S. Department of Defense Grant W81XWH-07-1-0294 (to K.P.), and American Cancer Society Grant RSG-05-154-01-MGO (to K.P.); Susan Komen Foundation fellowships PDF042234 (to M.H.) and PDF0707996 (to M.S.); Fannie E. Rippel Foundation fellowship (S.A.M.); Harvard/Dana-Farber SPORE in Breast Cancer Developmental Research (E.L.S.); and the Office of Energy Research, Office of Health and Environmental Research, U.S. Department of Energy, Contract DE-AC03-76SF00098 (to M.R.S.); and National Institutes of Health Grant CA112970 (to M.R.S.).

1. Stingl J, Raouf A, Eirew P, Eaves CJ (2006) Deciphering the mammary epithelial cell hierarchy. *Cell Cycle* 5:1519–1522.
2. Ashkenazi R, Jackson TL, Dontu G, Wicha MS (2007) Breast cancer stem cells-research opportunities utilizing mathematical modeling. *Stem cell reviews* 3:176–182.
3. Villadsen R, *et al.* (2007) Evidence for a stem cell hierarchy in the adult human breast. *J Cell Biol* 177:87–101.
4. Dontu G, *et al.* (2003) In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes Dev* 17:1253–1270.
5. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF (2003) Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci USA* 100:3983–3988.
6. Shipitsin M, *et al.* (2007) Molecular definition of breast tumor heterogeneity. *Cancer Cell* 11:259–273.
7. Jackson M, *et al.* (2004) Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells. *Mol Cell Biol* 24:8862–8871.
8. Pasini D, Bracken AP, Hansen JB, Capillo M, Helin K (2007) The polycomb group protein Suz12 is required for embryonic stem cell differentiation. *Mol Cell Biol* 27:3769–3779.
9. Wernig M, *et al.* (2007) In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* 448:318–324.
10. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. *Science* 270:484–487.
11. Ince TA, *et al.* (2007) Transformation of different human breast epithelial cell types leads to distinct tumor phenotypes. *Cancer Cell* 12:160–170.
12. Hu M, *et al.* (2005) Distinct epigenetic changes in the stromal cells of breast cancers. *Nat Genet* 37:899–905.
13. Ludwig TE, *et al.* (2006) Derivation of human embryonic stem cells in defined conditions. *Nat Biotechnol* 24:185–187.
14. Xu RH, *et al.* (2002) BMP4 initiates human embryonic stem cell differentiation to trophoblast. *Nat Biotechnol* 20:1261–1264.
15. Krivtsov AV, *et al.* (2006) Transformation from committed progenitor to leukemia stem cell initiated by MLL-AF9. *Nature* 442:818–822.
16. Cole MF, Johnstone SE, Newman JJ, Kagey MH, Young RA (2008) Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. *Genes Dev* 22:746–755.
17. Klenova EM, Morse HC, III, Ohlsson R, Lobanenkov VV (2002) The novel BORIS + CTCF gene family is uniquely involved in the epigenetics of normal biology and cancer. *Semin Cancer Biol* 12:399–414.
18. Pasini D, Bracken AP, Jensen MR, Lazzerini Denchi E, Helin K (2004) Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. *EMBO J* 23:4061–4071.
19. Lee TI, *et al.* (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* 125:301–313.
20. Bracken AP, Dietrich N, Pasini D, Hansen KH, Helin K (2006) Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes Dev* 20:1123–1136.
21. Nikolsky Y, Ekins S, Nikolskaya T, Bugrim A (2005) A novel method for generation of signature networks as biomarkers from complex high throughput data. *Toxicol Lett* 158:20–29.
22. He X, Zhang J (2006) Why do hubs tend to be essential in protein networks? *PLoS genetics* 2:e88.
23. Myatt SS, Lam EW (2007) The emerging roles of forkhead box (Fox) proteins in cancer. *Nat Rev Cancer* 7:847–859.
24. Wood LD, *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science* 318:1108–1113.
25. Herman JG, Graff JR, Myohanen S, Nelkin BD, Baylin SB (1996) Methylation-specific PCR: A novel PCR assay for methylation status of CpG islands. *Proc Natl Acad Sci USA* 93:9821–9826.
26. Shurin GV, *et al.* (2005) Loss of new chemokine CXCL14 in tumor tissue is associated with low infiltration by dendritic cells (DC), whereas restoration of human CXCL14 expression in tumor cells causes attraction of DC both in vitro and in vivo. *J Immunol* 174:5490–5498.