Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer

Duncan Sproul[a,b], Colm Nestor[a,c], Jayne Culley[a,b], Jacqueline H. Dickson[a,b], J. Michael Dixon[a,d], David J. Harrison[a], Richard R. Meehan[a,c], Andrew H. Sims[a,b], and Bernard H. Ramsahoye[a,b,1]

[a]Breakthrough Breast Cancer Research Unit and [b]Edinburgh Cancer Research Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, [c]Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, and [d]Edinburgh Breast Unit, Western General Hospital, Edinburgh EH4 2XU, United Kingdom

Edited* by Rudolf Jaenisch, Whitehead Institute for Biomedical Research, Cambridge, MA, and approved February 7, 2011 (received for review September 10, 2010)

Aberrant promoter hypermethylation is frequently observed in cancer. The potential for this mechanism to contribute to tumor development depends on whether the genes affected are repressed because of their methylation. Many aberrantly methylated genes play important roles in development and are bivalently marked in ES cells, suggesting that their aberrant methylation may reflect developmental processes. We investigated this possibility by analyzing promoter methylation in 19 breast cancer cell lines and 47 primary breast tumors. In cell lines, we defined 120 genes that were significantly repressed in association with methylation (SRAM). These genes allowed the unsupervised segregation of cell lines into epithelial (EPCAM+ve) and mesenchymal (EPCAM−ve) lineages. However, the methylated genes were already repressed in normal cells of the same lineage, and >90% could not be derepressed by treatment with 5-aza-2′-deoxycytidine. The tumor suppressor genes APC and CDH1 were among those methylated in a lineage-specific fashion. As predicted by the epithelial nature of most breast tumors, SRAM genes that were methylated in epithelial cell lines were frequently aberrantly methylated in primary tumors, as were genes specifically repressed in normal epithelial cells. An SRAM gene expression signature also correctly identified the rare claudin-low and metaplastic tumors as having mesenchymal characteristics. Our findings implicate aberrant DNA methylation as a marker of cell lineage rather than tumor progression and suggest that, in most cases, it does not cause the repression with which it is associated.

Aberrant CpG island methylation occurs in cancer and is implicated in tumor progression (1), particularly when methylation of a tumor suppressor gene appears to phenocopy the equivalent genetic mutation. Examples include *MLH1* methylation in sporadic microsatellite unstable colon cancer (2) and *Rb* in retinoblastoma (3).

Several tumor suppressor genes and putative tumor suppressor genes have been reported to be methylated in breast cancer (4), but in most cases, evidence for a functional role in tumorigenesis is lacking. BRCA1, which is mutated in familial breast cancer, is reported to be methylated in ~10% of sporadic tumors. In BRCA1-associated familial tumors, the wild-type BRCA1 allele is frequently lost. One report suggested that the loss of function could occur through methylation of the remaining wild-type allele (5), but this finding has not been supported by subsequent, larger studies (6, 7).

Breast development begins in embryonic life when epidermal cells of ectodermal origin project into the mesenchyme underlying the mammary ridge and form lactiferous ducts. Mammary stem cells give rise to both the inner luminal-epithelial and the outer "basal" myoepithelial cells of the lobulo-ductal system (8). Primary breast tumors can been subdivided into many different types by histology and by molecular profiling, but most tumors are thought to be epithelial in origin, deriving either from luminal-epithelial cells or from their progenitors (9).

It is known that many genes de novo methylated in cancer have "bivalent" histone marks (combined histone H3 lysine-27 and lysine-4 trimethylation) in embryonic stem (ES) cells (10). Because bivalently marked genes frequently have a role in development, we asked whether cancer-associated aberrant methylation might reflect the particular cell lineage from which a breast tumor was derived (its ontogeny). We show that aberrant DNA methylation occurs in genes down-regulated through normal lineage commitment and that the genes affected can be used to distinguish breast tumors of epithelial and mesenchymal lineage. We propose that most aberrant methylation reflects lineage commitment rather than tumor progression.

Results

DNA Methylation Occurs Variably Across Breast Cancer Cell Lines and Is Associated with Gene Repression. We correlated promoter methylation with gene expression by analyzing 19 breast cancer cell lines on Infinium arrays and combining these results with published transcriptome data (11). Infinium arrays assay the proportion of 5-methylcytosine to total cytosine at 27,578 different CpG dinucleotides in >14,000 genes after bisulfite conversion (12). We validated the capacity of the Infinium arrays to detect changes in DNA methylation by using DNA from wild-type and DNA methyltransferase deficient HCT116 colon cancer cell lines (Fig. S1*A*). The methylation levels reported by the arrays also corresponded well to those assayed by bisulfite sequencing, both for the individual CpGs interrogated and for neighboring CpGs (Fig. S1 *B* and *C*). We restricted our analysis to probes within 200 bp of transcription start sites because we were interested in the effects of methylation on expression. As expected, genes associated with methylated promoters were less expressed than genes with unmethylated promoters (Fig. S1*D*).

To understand the factors that might be influencing methylation in the cell lines, we categorized the CpG probes into three groups depending on their consistency of methylation across the cell lines (Fig. 1*A* and *Materials and Methods*) and determined the proportion of CpG island genes with each group. Most consistently unmethylated (CU) probes (3,901 genes) were located within CpG islands, whereas consistently methylated (CM) probes (259 genes) were mostly located at non-CpG island promoters (Fig. S1*E*). Variably methylated (VM) probes (1,023 genes) were significantly more likely to be in CpG islands than

Author contributions: D.S., D.J.H., R.R.M., A.H.S., and B.H.R. designed research; D.S., C.N., J.C., and J.H.D. performed research; J.M.D. contributed new reagents/analytic tools; D.S. analyzed data; and D.S. and B.H.R. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE26990).

[1]To whom correspondence should be addressed. E-mail: Bernard.Ramsahoye@ed.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1013224108/-/DCSupplemental.

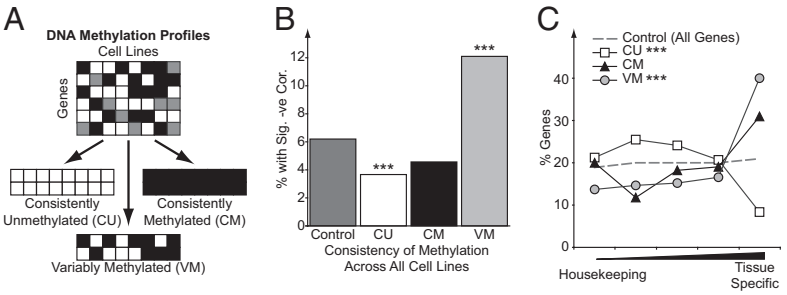

Fig. 1. VM genes have tissue-specific expression patterns. (*A*) An illustration of the general strategy used to segregate genes into sets with different methylation patterns. (*B*) The proportion of CU, CM, and VM genes, which show a significantly negative correlation between expression and methylation compared with the percentage found on the whole array (Fisher's exact tests). (*C*) The expression patterns of genes in different methyl gene sets in normal tissues as defined using a specificity score (*SI Materials and Methods*). The distributions of CU and VM genes were significantly different from the profiles of all genes (***$P < 0.001$, χ^2 test).

CM probes (51% vs. 24%). VM genes were frequently not expressed even when unmethylated, with 45% being unexpressed in all 19 cell lines (Fig. S1*F*). However, a significant proportion (12%) of the VM genes did show the expected inverse relationship between DNA methylation and expression (Fig. 1*B*).

Methylated and Variably Methylated Genes Have Tissue-Specific Expression Patterns. We functionally characterized the gene groups using Gene Ontology (GO) terms (Fig. S1*G*). CU genes were associated with metabolic or housekeeping processes, whereas CM genes were associated with more specialized, lineage-restricted terms, such as meiosis, and mast cell activation. In contrast, VM genes were significantly associated with general developmental processes.

Given that genes with different methylation patterns were associated with different functions, we examined whether they also had different patterns of expression in normal tissues by scoring them according to their degree of tissue specificity (*SI Materials and Methods*). CU genes were significantly enriched in genes showing a housekeeping expression pattern (Fig. 1*C*). In contrast, VM genes were significantly enriched for tissue-specific expression. CM genes displayed a similar pattern to VM genes but did not quite reach significance ($P = 0.065$). The tissue specificity of VM genes was also apparent when VM genes with CpG island and non-CpG island promoters were analyzed separately (Fig. S1*H*).

CpGs That Are Variably Methylated in Cell Lines Are Frequently Unmethylated in Normal Breast Tissue and Normal Mammary Epithelial Cells. We next asked whether it was the CM or VM probes that could be regarded as aberrantly methylated in cancer because they were unmethylated in normal human mammary epithelial cells (HMEC) and normal breast tissue. CU probes were nearly always unmethylated in the normal DNA samples. A high proportion of VM probes (58–66%) were also unmethylated in these normal samples, and this was a significantly greater proportion than was found for the CM probes (5–7%, $P < 2.2 \times 10^{-16}$, Fisher's exact tests; Fig. 2*A*). As there were also ~4 times more VM genes ($n = 1{,}023$) than CM genes ($n = 259$), aberrant methylation was significantly more likely to occur at VM genes. VM probes were also more likely to be unmethylated than CM probes in a panel of nine normal tissues and in human ES (hES) cells (Fig. S2 *A* and *B*).

VM Genes Are Enriched for "Bivalent" Histone Marks in hES Cells. Cancer-associated aberrant methylation frequently occurs at genes with bivalent histone marks in hES cells (histone H3K4me3 and H3K27me3; ref. 10). We noticed a striking similarity between functional terms associated with VM genes and those previously associated with bivalently marked genes in hES cells (Fig. S2*C*; ref. 13). We used data from this study to determine the histone marks associated with CU, CM, and VM genes in hES cells. A significant proportion of CU genes were marked by H3K4me3 alone ($P < 2 \times 10^{-16}$, Fisher's exact test), whereas most CM and VM genes lacked H3K4me3 and H3K27me3 (Fig. 2*B*). However, the VM group was significantly enriched for bivalent marks (16.9% of the total; $P = 7 \times 10^{-22}$, Fisher's exact test) compared with the control. This enrichment was not seen in the CM group.

Genes That Are Significantly Repressed in Association with Methylation (SRAM) Segregate Breast Cancer Cell Lines into Epithelial and Mesenchymal Lineages. As VM genes were lineage-specific, we asked whether they could be used to categorize the cell lines according to lineage. The expression levels of the 1,000 most variably expressed genes segregated the 19 breast cancer cell lines into the previously described luminal, basal A, and basal B subtypes (Fig. 3*A*; ref. 11). However, hierarchical clustering using methylation levels of the 1,023 VM genes derived different groupings (Fig. 3*B*): Two of the basal A cell lines (MDAMB468 and HCC1954) now clustered with the luminal cell lines. As not all VM genes showed a good correlation with repression (Fig. 1*B*), we repeated the analysis using the expression levels of those VM genes that were significantly repressed in association with methylation (SRAM; 120 genes; Fig. 3*C* and Dataset S1). In this analysis, all of the basal A cell lines clustered with the luminal cell lines. Similar results were observed when we used only those SRAM genes with CpG island promoters (67 genes; Fig. 3*A*).

The classification based on SRAM genes correlated well with cell morphology; the luminal group cells generally grew as tight clusters typical of epithelial cells, whereas the other group showed less cell–cell contact and were spindle-shaped (Fig. S3B). The epithelium-like cells were all exclusively positive for the epithelial marker *EPCAM* (also known as *TACSTD1* and recognized by the BerEP4 antibody; Fig. 3*D*) and, with the exception of HCC1569, all expressed cytokeratin 19 and other markers expressed by normal epithelial cells (Fig. 3*E*; ref.14). In contrast, the other group was negative for *EPCAM* expression and, with the exception of MCF10A cells, did not consistently express keratins. However, they did express genes associated with mesenchyme (Fig. 3*E*; ref. 15). These data indicated that EPCAM–ve cells were likely to be of mesenchymal lineage. Thus, the

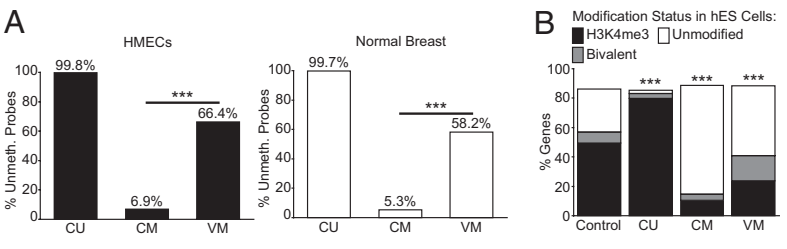

Fig. 2. VM genes are usually unmethylated in normal breast tissues and are enriched for bivalent histone marks in hES cells. (*A*) The proportions of CU, CM, and VM probes that are unmethylated in either HMECs or the normal breast are shown. Significantly more VM than CM probes are unmethylated in the normal samples (Fisher's exact tests). (*B*) The proportions of CU, CM, and VM genes that have different histone modification patterns in hES cells are shown. All three groups show a distribution that is significantly different from the control (all genes on the array, χ^2 tests). ***$P < 0.001$.

differential expression of SRAM genes classified breast cancer cell lines into those of epithelial and mesenchymal lineage.

SRAM Genes Undergo Lineage-Specific Repression. Heat maps of SRAM gene expression and methylation illustrate the striking patterns that differentiate epithelial and mesenchymal cell lines (Fig. 4*A*; larger heat maps are presented in Fig. S4*A*). The SRAM gene list contains *APC*, *GSTP1*, and *PYCARD* (16, 17), which have been reported to be methylated in breast cancer, and *CLDN7*, a tight junction protein expressed in epithelial cells that is methylated in some breast cancer cell lines (18). It also contains genes that have been shown to be differentially expressed in different subcompartments of the normal breast (for example, *SPARC* and *MB*; refs. 19 and 20). Indeed, 71 SRAM genes are included in published signatures of different cell populations purified from normal breast tissue (21), a highly significant enrichment ($P = 7.1 \times 10^{-16}$, Fisher's exact test).

To determine whether the SRAM genes were coordinately repressed in association with lineage in the normal breast, we interrogated the same dataset of normal cell populations (21). SRAM genes preferentially methylated in *EPCAM*+ve breast cancer cell lines had significantly lower levels of expression in cellular fractions corresponding to differentiated luminal and luminal progenitor cells (both *EPCAM*+ve, Wilcoxon test; Fig. 4*B*). In contrast, genes methylated in *EPCAM*–ve breast cancer cell lines had significantly lower levels of expression in the basal/myoepithelial cell fraction and even lower levels of expression in the mesenchymal stromal fraction (both *EPCAM*–ve, Wilcoxon test; Fig. 4*B*). A similar pattern was observed when we considered SRAM genes with CpG island and non-CpG island promoters separately (Fig. S4*B*). Thus, genes prone to methylation in cell lines of different lineages are generally already repressed in normal cells of the corresponding lineage.

Majority of Genes Methylated in Breast Cancer Cell Lines Are Not Derepressed by Demethylation. As our results suggested that genes prone to methylation might already be repressed by lineage-specific factors, we investigated the extent to which DNA methylation might be important for their repression using the demethylating agent 5-aza-2′-deoxycytidine (5-aza-dC). Treatment of three breast cancer cell lines with 5-aza-dC led to the demethylation and reexpression of *DAZL*, a gene whose expression is known to be directly controlled by DNA methylation in normal development (Fig. S4 *C* and *D*; ref. 22). The cancer testis antigen *GAGE4* was also derepressed as expected (23). We profiled gene expression levels after 5-aza-dC or mock treatment using microarrays, combining this with our methylation data to ascertain, in an unbiased manner, the proportion of methylated genes that were reactivated. Less than 10% of the silenced methylated genes were derepressed by 5-aza-dC in the three breast cancer lines, and derepression did not show a greater specificity for VM genes (Fig. 4 *C* and *D*). A similar proportion of genes with unmethylated promoters were derepressed by 5-aza-dC exposure (Fig. 4*C*). Our arrays indicated that methylated *CDH1* gene was not reexpressed by 5-aza-dC in HBL100 cells; this result was verified using quantitative RT-PCR (Fig. S4*D*).

As would be expected, 5-aza-dC treatments lead to significant but incomplete demethylation (Fig. S4*C*). To be sure that we were not missing transcription effects because of inadequate demethylation, we took advantage of the DNA methyltransferase-deficient HCT116 DKO cells where DNA methylation is reduced to 3–4% of that seen in wild-type (24, 25). We compared genes reactivated in DKO cells with those reactivated by treating wild-type HCT116 cells with 1 μM 5-aza-dC for 3 d, a dose shown to reduce global methylation to 35% of control (26). As expected, there was a significant overlap in the methylated genes dere-

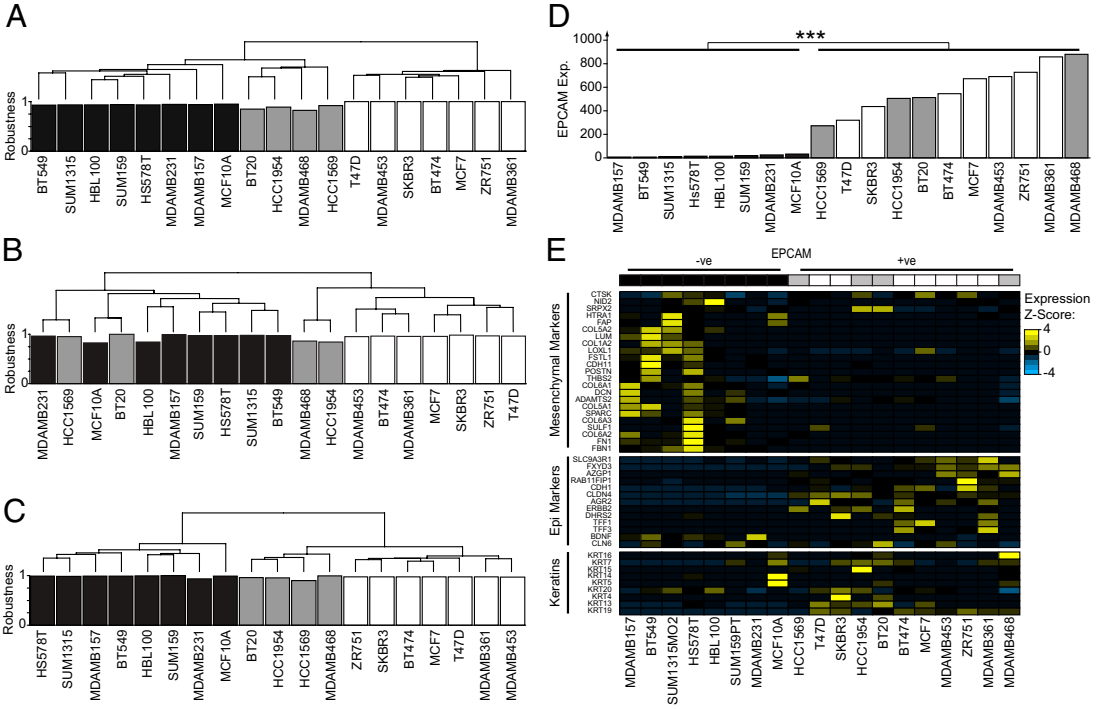

Fig. 3. SRAM gene expression segregates breast cancer cell lines into cells of epithelial and mesenchymal lineage. (*A*–*C*) Dendrograms derived from unsupervised hierarchical clustering of the cell lines based on expression of the 1,000 most variably expressed genes (*A*), percentage methylation of the 1,023 VM genes (*B*), and expression values from a subset of genes that are SRAM (120 genes; *C*). The robustness of each sample's cluster membership is shown below the dendrogram, expressed as the percentage of permutations in which that sample grouped in its cluster (consensus clustering, see supplementary methods). White, luminal A; gray, basal A; black, basal B (according to ref. 11). (*D*) The expression of *EPCAM* correlates with the two main clusters derived in *C* ($P < 2.2 \times 10^{-16}$, Wilcoxon test). The cell lines are ordered based upon their expression of *EPCAM*. Color coding as for *A*–*C*. (*E*) Markers of epithelial and mesenchymal lineages (*SI Materials and Methods*) are differentially expressed between the cell lines. The cell lines are ordered and color-coded as in *D*. Genes that were silent in all 19 cell lines were excluded from the analysis.

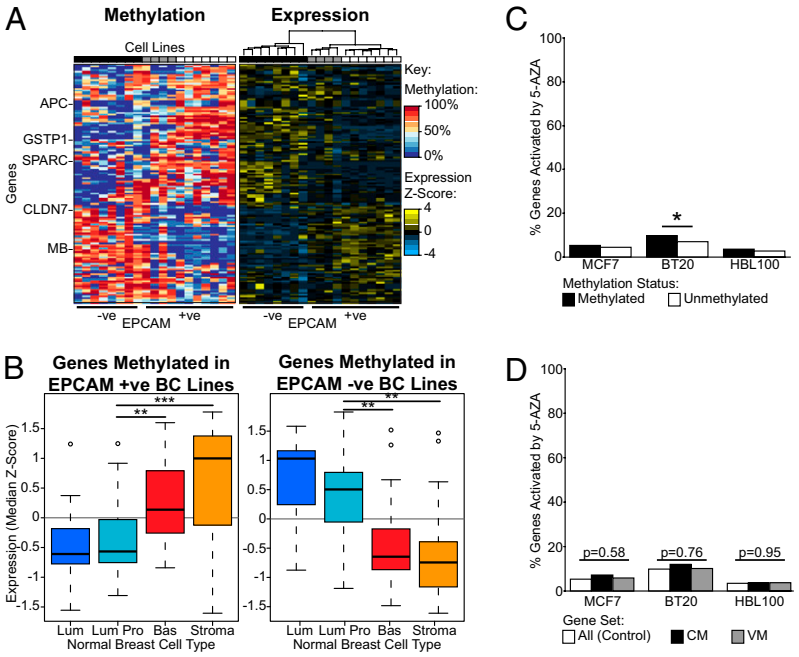

Fig. 4. In cell lines SRAM genes are repressed and methylated in a lineage-dependent manner, and most are not controlled by DNA methylation. (*A*) Heat maps showing the expression and methylation levels of SRAM genes in breast cancer cell lines (color coded as in Fig. 3) together with their *EPCAM* status. The cell lines and genes are clustered using hierarchical clustering. See Fig. S4*A* for larger heat maps. (*B*) Expression levels of differentially methylated SRAM genes in different cell types in the normal breast. Lum, luminal epithelial cells; Lum Pro, luminal epithelial progenitors (both *EPCAM*+ve); Bas, basal myoepithelial cells; Stroma, mesenchymal stromal cells (both *EPCAM*−ve). Expression values are median *z* scores, and differences between groups were tested using Wilcoxon tests. (*C*) The percentages of methylated and unmethylated genes that were reactivated by 5-aza-dC treatment in three breast cancer cell lines (Fisher's exact tests). (*D*) The percentage of CM and VM genes reactivated by 5-aza-dC compared with the percentage of all genes reactivated by 5-aza-dC. No significant differences were detected (χ^2 tests). $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

pressed by these two methods (Fig. S4*E*). However, despite the fact that more methylated genes were derepressed in the DKO cells than by 5-aza-dC treatment, this result still only represented 16.5% of methylated genes (Fig. S4*F*). These data suggest that DNA methylation at promoters is not the primary mechanism responsible for the repression of most methylated genes in cancer.

Lineage-Specific Aberrant Methylation Occurs in Primary Tumors. To examine whether lineage-specific methylation also occurred in primary tumors, we generated methylation profiles from 47 primary breast tumors. Firstly, we analyzed SRAM gene methylation in the samples. After excluding probes that were methylated in normal breast, SRAM probes that were methylated in *EPCAM*+ve cell lines were significantly more frequently methylated in primary tumors than those specific for *EPCAM*−ve cell lines (Fig. 5 *A* and *B* and Fig. S5*A*). A further analysis using all genes that showed a significant preference for methylation in *EPCAM*+ve or −ve cell lines produced a similar result (Fig. S5*B*). Furthermore, genes that were specifically repressed in normal luminal epithelial cells (compared with stroma) were also significantly more frequently methylated in primary tumors than those genes that were active (Fig. S5*C*). Within this list of genes we also found significant enrichments in genes previously reported to be frequently methylated in breast tumors (Table S1).

We then looked specifically at the methylation of important tumor suppressor genes in breast cancer (*BRCA1* and *CDH1*), as well as other genes that have been frequently reported to be methylated and that might also be important in breast cancer biology (*APC*, *GSTP1*, and *ESR1*). *GSTP1* and *APC* are both SRAM genes methylated predominantly in *EPCAM*+ve cell lines (Fig. 4*A*) and were frequently methylated in primary tumors (Fig. S5*D*). Both are expressed in luminal progenitor cells but are down-regulated in differentiated luminal cells, suggesting that their methylation could be linked to terminal differentiation. *BRCA1* displayed a similar expression pattern and was methylated to a level of >30% in 4 of the 47 (8.5%) primary tumors, a frequency consistent with previous reports (27). In contrast, *CDH1* and *ESR1*, which are both expressed in epithelial cells, were infrequently methylated (2/47 and 0/47, respectively; Fig. S5*D*). The level of *CDH1* methylation in the two tumors was also comparatively low (31% and 34%). This result is consistent with methylation rarely affecting genes that are ordinarily expressed in that lineage. In cell lines, CDH1 methylation was specific to those with low *EPCAM* expression (Fig. S5*D*).

Our results demonstrate that primary tumors have epithelial-specific methylation patterns. However, recent reports have suggested that certain rare tumor types, claudin-low and metaplastic tumors, might have mesenchymal characteristics (28). We tested whether an expression signature composed of SRAM genes could distinguish these tumors in that dataset. As predicted, most tumor subtypes had a high *EPCAM*+ve score, but claudin-low and metaplastic tumors more closely resembled the SRAM expression profile of *EPCAM*−ve cell lines (Fig. 5*C* and Fig. S5*E*). Our signature was also predictive of *EPCAM* expression in tumors, as had been the case for the cell lines (Fig. 5*D*). The expression levels of a larger panel of marker genes further supported a mesenchymal origin for claudin-low tumors and metaplastic tumors, although the latter also expressed some epithelial markers (Fig. S5*F*).

Discussion

The methylation of CpG island promoters is a normal developmental process that is essential for repression of some genes, such as those on the inactive X chromosome, imprinted genes, and some tissue-specific genes (29). In cancer, many additional promoters are both repressed and methylated. It is often argued that methylation could also be instrumental in their repression. However, our data suggest a model whereby in breast cancer aberrant methylation occurs at genes that are already repressed through normal lineage commitment and methylation is generally not required for their repression (Fig. 5*E*). Lineage-specific aberrant methylation has not been previously reported but can be found in datasets of breast cancer methylation patterns from a number of other studies (Table S1).

The finding that most cancer-associated aberrant methylation occurs in genes that are already down-regulated has been alluded to previously (30), and this phenomenon also occurs in normal cultured neural cells (31). However, the literature contains many examples of methylated genes being derepressed by 5-aza-dC in cell lines, which has been central to the argument that aberrant methylation causes tumor progression by silencing genes. Indeed, one study assumed that in HCT116 cells, all methylated genes are repressed because of methylation and used the amount of deregulation induced by 5-aza-dC to estimate the size of the methylome (5% of all genes; ref. 32). However, by using an unbiased approach and directly measuring the proportions of methylated and unmethylated genes that are actually derepressed by 5-aza-dC, our results challenge this view. We find that

GENETICS

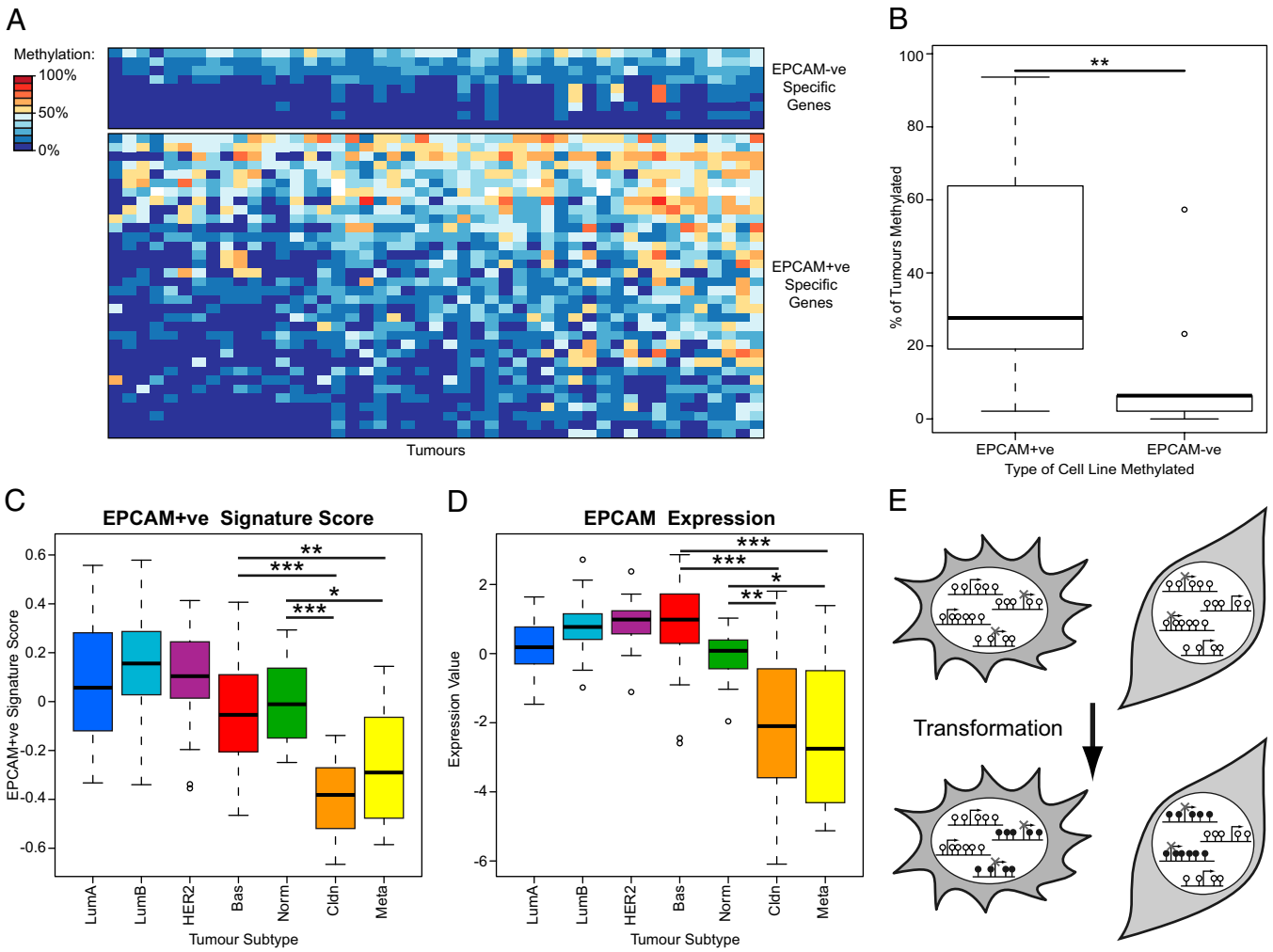

Fig. 5. Lineage-specific aberrant methylation occurs in primary breast tumors. (*A*) Heat map indicating methylation frequency of differentially methylated SRAM genes in 47 primary breast tumors. Only genes that are unmethylated in the normal breast are shown. Genes and samples are ordered by their frequency of methylation. A larger version of the heat map is in Fig. S5*A*. (*B*) Genes methylated in *EPCAM*+ve cell lines are more frequently methylated in primary tumors. The frequency of methylation in tumors of the groups of genes shown in *A* was compared. Significance was assessed using a Wilcoxon test. (*C*) Boxplot of *EPCAM*+ve SRAM expression signature scores by tumor type for a series of breast tumors. Claudin-low (Cldn) and metaplastic tumors (Meta) have scores that are significantly lower than all other subtypes (Wilcoxon tests). A plot using an *EPCAM*−ve signature is in Fig. S5*E*. (*D*) Boxplot of *EPCAM* expression by tumor subtype. Claudin-low and metaplastic tumors have significantly lower *EPCAM* expression than the other subtypes (Wilcoxon tests). (*E*) Model showing that normal lineage commitment leads to the repression of genes in a lineage-specific manner. Lineage-repressed genes are prone to hypermethylation upon transformation. **P* < 0.05; ***P* < 0.01; ****P* < 0.001, Wilcoxon tests.

5-aza-dC derepresses <10% of all methylated genes and that 25.5% of all genes are methylated in HCT116 cells. Repressive histone marks may remain after treatment with 5-aza-dC (33), indicating that DNA methylation may be one of many epigenetic mechanisms involved in repression.

We found *CDH1* and *ESR1* to be rarely methylated in primary tumors, and as these genes are expressed in epithelial cells, this finding would be predicted by our model. Some studies using the nonquantitative methylation-specific PCR (MSP) technique have reported higher methylation frequencies of *CDH1* and *ESR1* (16, 34), including 72% in the case of *CDH1* (35). However, MSP may be prone to detecting low-level methylation at some genes. A study using a quantitative version of MSP agrees with our finding of infrequent methylation at *CDH1* and *ESR1* (36). We found genuine *BRCA1* methylation in 8.5% of tumors, consistent with a previous report (27). *BRCA1* is down-regulated during terminal epithelial differentiation, which could make it susceptible to methylation. We also note that disproportionately frequent *BRCA1* methylation is observed in metaplastic carcinomas (63% of a series of 27 tumors; ref. 27). As these tumors appear to be mesenchymal and *BRCA1* is repressed in normal mesenchymal cells, this finding is consistent with our model that methylation affects lineage-repressed genes.

The detection of *CDH1* methylation by sensitive nonquantitative techniques (MSP) could be due to the presence of stromal cells, contaminating blood cells (37), or tumor cells that have undergone epithelial-to-mesenchymal transition (EMT). The existence of EMT is breast cancer is contentious, but it would be predicted to down-regulate *CDH1* and induce metastasis. EMT might also lead to *CDH1* methylation under our model. However, as we did not detect significant *CDH1* methylation, our data do not support extensive EMT in most breast tumors. A previous study also found no differences in *CDH1* expression between primary tumors and their metastases (38). Whether EMT is responsible for the mixed epithelial and mesenchymal characteristics in rare metaplastic tumors (Fig. S5*E*) remains to be determined.

Although we observed that aberrantly methylated genes were significantly enriched for those that are bivalently marked in ES cells (10), most of the affected genes lacked these marks. It is possible, therefore, that in cancer, bivalent genes are prone to methylation because they are lineage-specific and repressed,

rather than because of a direct interaction of the polycomb and DNA methylation machineries.

In summary, our data indicate that aberrant methylation is a marker of lineage restriction in cancer. Although we cannot claim that this finding applies to every aberrantly methylated gene, our unbiased approach clearly demonstrates that normal developmental repression influences whether genes become aberrantly methylated in cancer. Our findings force a reappraisal of the likely efficacy of DNA demethylating agents in cancer therapy.

Materials and Methods

A brief summary of methods used is given below. For full details, see *SI Materials and Methods* and Tables S2–S4.

Breast Cancer Cell Lines and Samples. Breast cancer cell lines were obtained from Cancer Research UK or ATCC. Wild-type and DKO (Dnmt1$^{-/-}$, Dnmt3b$^{-/-}$) HCT116 cells were kind gifts from B. Vogelstein (24). HMECs were a gift from E. Katz at the Edinburgh Breakthrough Breast Cancer Research Unit. SHEF-6 hES cell DNA was a gift from D. Hay (MRC Centre for Regenerative Medicine). DNA from normal breast, fetal and adult brain, testis, liver, placenta, spleen, blood, and colon were from Biochain, After approval by our ethical board, 47 fresh frozen unselected tumor samples were obtained through the Experimental Cancer Medicine Centre in Edinburgh.

5-aza-dC Treatment. Cell lines were exposed to 1 μM 5-aza-dC, refreshed every 24 h, for a total of 72 h.

Microarrays. We used Illumina Infinium Human Methylation27 beadarrays and Illumina human HT12 Expression beadarrays. All data have been submitted to the Gene Expression Omnibus (GEO) database (accession no. GSE26990).

Bioinformatic Analyses. Bioinformatic analyses were performed using R (version 2.9.2). Ensembl 54 (NCBI36) gene annotations were used throughout. In cell lines unmethylated genes were defined as those with ≤30% methylation and methylated ones as ≥70%. In tumor samples methylated genes were defined as those with >30% methylation due to the heterogeneity of the samples. We defined groups of genes with different methylation patterns as follows: CU, unmethylated in all cell lines; CM, methylated in all cell lines or all but one cell line; VM, methylated in at least four and unmethylated in at least four cell lines. Only CpGs within 200 bp of transcription start sites were considered in our analyses. SRAM genes were VM genes that had significantly lower expression when methylated (one-sided Wilcoxon test). The specificity of a gene expression pattern was measured using a method based on information theory (*SI Materials and Methods*). Datasets were downloaded from data repositories or individual papers as appropriate.

ACKNOWLEDGMENTS. We thank Dr. T. I. Simpson for the use of his consensus clustering algorithm, L. Renshaw for assistance in the collection of clinical samples, and Dr. J.S. Thomas for assistance with breast cancer pathology. This work was supported by Breakthrough Breast Cancer and the Medical Research Council. Central services utilized in the study are funded by Cancer Research UK and the Wellcome Trust.

1. Jones PA, Baylin SB (2002) The fundamental role of epigenetic events in cancer. *Nat Rev Genet* 3:415–428.
2. Herman JG, et al. (1998) Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proc Natl Acad Sci USA* 95:6870–6875.
3. Ohtani-Fujita N, et al. (1997) Hypermethylation in the retinoblastoma gene is associated with unilateral, sporadic retinoblastoma. *Cancer Genet Cytogenet* 98: 43–49.
4. Widschwendter M, Jones PA (2002) DNA methylation and breast carcinogenesis. *Oncogene* 21:5462–5482.
5. Esteller M, et al. (2001) DNA methylation patterns in hereditary human cancers mimic sporadic tumorigenesis. *Hum Mol Genet* 10:3001–3007.
6. Dworkin AM, Spearman AD, Tseng SY, Sweet K, Toland AE (2009) Methylation not a frequent "second hit" in tumors with germline BRCA mutations. *Fam Cancer* 8: 339–346.
7. Tung N, et al. (2010) Prevalence and predictors of loss of wild type BRCA1 in estrogen receptor positive and negative BRCA1-associated breast cancers. *Breast Cancer Res* 12: R95.
8. Shackleton M, et al. (2006) Generation of a functional mammary gland from a single stem cell. *Nature* 439:84–88.
9. Gusterson B (2009) Do 'basal-like' breast cancers really exist? *Nat Rev Cancer* 9: 128–134.
10. Ohm JE, et al. (2007) A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat Genet* 39: 237–242.
11. Neve RM, et al. (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 10:515–527.
12. Laird PW (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet* 11:191–203.
13. Zhao XD, et al. (2007) Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell* 1:286–298.
14. Allinen M, et al. (2004) Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell* 6:17–32.
15. Herschkowitz JI, et al. (2007) Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol* 8:R76.
16. Esteller M, Corn PG, Baylin SB, Herman JG (2001) A gene hypermethylation profile of human cancer. *Cancer Res* 61:3225–3229.
17. Conway KE, et al. (2000) TMS1, a novel proapoptotic caspase recruitment domain protein, is a target of methylation-induced gene silencing in human breast cancers. *Cancer Res* 60:6236–6242.
18. Kominsky SL, et al. (2003) Loss of the tight junction protein claudin-7 correlates with histological grade in both ductal carcinoma in situ and invasive ductal carcinoma of the breast. *Oncogene* 22:2021–2033.
19. Jones C, et al. (2004) Expression profiling of purified normal human luminal and myoepithelial breast cells: Identification of novel prognostic markers for breast cancer. *Cancer Res* 64:3037–3045.
20. Kristiansen G, et al. (2010) Endogenous myoglobin in human breast cancer is a hallmark of luminal cancer phenotype. *Br J Cancer* 102:1736–1745.
21. Lim E, et al. (2009) Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med* 15:907–913.
22. Maatouk DM, et al. (2006) DNA methylation is a primary mechanism for silencing postmigratory primordial germ cell genes in both germ cell and somatic cell lineages. *Development* 133:3411–3418.
23. Kumagai T, et al. (2009) Epigenetic regulation and molecular characterization of C/EBPalpha in pancreatic cancer cells. *Int J Cancer* 124:827–833.
24. Rhee I, et al. (2002) DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature* 416:552–556.
25. Egger G, et al. (2006) Identification of DNMT1 (DNA methyltransferase 1) hypomorphs in somatic knockouts suggests an essential role for DNMT1 in cell survival. *Proc Natl Acad Sci USA* 103:14080–14085.
26. Patel K, et al. (2010) Targeting of 5-aza-2′-deoxycytidine residues by chromatin-associated DNMT1 induces proteasomal degradation of the free enzyme. *Nucleic Acids Res* 38:4313–4324.
27. Turner NC, et al. (2007) BRCA1 dysfunction in sporadic basal-like breast cancer. *Oncogene* 26:2126–2132.
28. Hennessy BT, et al. (2009) Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. *Cancer Res* 69:4116–4124.
29. Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes Dev* 16:6–21.
30. Keshet I, et al. (2006) Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat Genet* 38:149–153.
31. Meissner A, et al. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454:766–770.
32. Schuebel KE, et al. (2007) Comparing the DNA hypermethylome with gene mutations in human colorectal cancer. *PLoS Genet* 3:1709–1723.
33. McGarvey KM, et al. (2006) Silenced tumor suppressor genes reactivated by DNA demethylation do not return to a fully euchromatic chromatin state. *Cancer Res* 66: 3541–3549.
34. Parrella P, et al. (2004) Nonrandom distribution of aberrant promoter methylation of cancer-related genes in sporadic breast tumors. *Clin Cancer Res* 10:5349–5354.
35. Caldeira JR, et al. (2006) CDH1 promoter hypermethylation and E-cadherin protein expression in infiltrating breast cancer. *BMC Cancer* 6:48.
36. Suijkerbuijk KP, et al. (2008) Methylation is less abundant in BRCA1-associated compared with sporadic breast cancer. *Ann Oncol* 19:1870–1874.
37. Lombaerts M, et al. (2004) Infiltrating leukocytes confound the detection of E-cadherin promoter methylation in tumors. *Biochem Biophys Res Commun* 319: 697–704.
38. Kowalski PJ, Rubin MA, Kleer CG (2003) E-cadherin expression in primary carcinomas of the breast and its distant metastases. *Breast Cancer Res* 5:R217–R222.

Supporting Information

Sproul et al. 10.1073/pnas.1013224108

SI Materials and Methods

DNA Methylation Profiling. 500 ng of DNA was bisulfite-converted (EZ DNA Methylation kit, Zymo Research), amplified, and hybridized to Illumina HumanMethylation27 Beadarrays following standard Illumina protocols. Array processing was performed at the Wellcome Trust Clinical Research Facility in Edinburgh. DNA methylation data has been submitted to the Gene Expression Omnibus (GEO) database (accession no. GSE26990).

Verification of Illumina Infinium Arrays by Bisulfite Sequencing. 500 ng of DNA was bisulfite-converted (EZ DNA Methylation kit, Zymo Research) and then subjected to two rounds of PCR amplification (35 cycles each, except DAZL primers, which were 38 cycles each) using nested primers (see Table S2 for primer sequences and annealing temperatures). One-tenth of the reaction from the first round was used in the second-round reaction. PCR products were cloned into pGEM T-Easy (Promega) and sequenced from the SP6 primer. Sequencing was analyzed with the BiQ Analyzer (1).

Expression Microarray Analysis. Total RNA was isolated from cell lines using TRIzol (Invitrogen). RNA integrity (RIN) was assessed using an Agilent 2100 Bioanalyzer (samples used had RIN score of ≥9.0). RNA was amplified and biotinylated using an Illumina TotalPrep RNA Amplification Kit and subsequently hybridized to Illumina human HT12 Expression BeadChips. Array processing was performed at the Wellcome Trust Clinical Research Facility in Edinburgh. Gene expression data have been submitted to the GEO database (accession no. GSE26990).

Quantitative RT-PCR Validation of Expression Microarray Analysis. cDNA was prepared from 400 ng of total RNA using random priming (Promega) and the SuperScriptII system (Invitrogen). Quantitative RT-PCR reactions were prepared using SYBR Green PCR Master Mix (Roche) and run under standard conditions on a LightCycler 480. Primers and conditions are shown in Table S3.

Preprocessing of Methylation Data and Gene Group Definition. Methylation data were exported from Illumina's Genome Studio, and beta values were converted to percentage methylation by multiplying by 100. The detection *P* value was used to filter out undetected probes from the analysis, flagging them as not available (NA) values (threshold 0.01). The profiles of BT474s, MCF7s, MDA-MB-468s, and MDA-MB-231s used here represent the median profile of multiple biological replicates (2, 7, 2, and 3, respectively).

In cell lines and normal tissues, probes were defined as unmethylated when they had ≤30% methylation, partially methylated when they had >30% and <70% methylation, and methylated when they had ≥70% methylation. In tumors, high methylation values were rarely observed due to heterogeneous mix of cell types in each sample. So we defined methylated probes as those that were not unmethylated (i.e., >30% methylation). Probes were defined as aberrantly methylated if they were unmethylated in the normal breast sample (i.e., ≤30% methylation).

We defined groups of genes with different methylation patterns as follows: consistently unmethylated (CU), unmethylated in all cell lines; consistently methylated (CM), methylated in all cell lines or all but one cell line; variably methylated (VM), methylated in at least four and unmethylated in at least four cell lines. Only CpGs within 200 bp of transcription start sites were considered in our analyses. Probes were mapped to genes, and any genes with an ambiguous status (e.g., found in both the CM and CU lists) were removed from analysis.

Analysis of Gene Expression Data. Raw expression values were background subtracted and normalized (average normalization) by using Illumina Genomestudio. We defined genes that were "off" and methylated as those for which all probes were undetected on Illumina expression arrays (detection $P \geq 0.05$) and median percent methylation values for Infinium probes within 200 bp of the TSS were ≥70%. The "off" and unmethylated genes had median percent methylation values of ≤30%. Reactivated genes were defined as those that had at least one probe detected in the 5-aza-dC sample.

Analysis of Specificity of Gene Expression. The specificity of a gene's expression pattern was measured by using a method based on information theory (2). A low score indicates that a gene is uniformly expressed, and a high score indicates that it is expressed specifically in one tissue. Specificities were calculated for all genes in the genome, and then genes were divided into five equal groups based on their ranking. The distributions of gene sets with different methylation patterns in breast cancer cell lines were compared with those of all genes on the array using a χ^2 test.

Relating Gene Expression to Methylation. To define SRAM genes, we performed a one-sided Mann–Whitney test on each of the VM genes. Genes were selected for which the expression values in methylated cell lines (methylation ≥70%) was significantly lower than in unmethylated cell lines (≤30%) using a cutoff of $P < 0.05$. This approach is similar to one that has been successfully applied to analysis of array comparative genomic hybridization data (3).

Definition and Application of EPCAM−ve and EPCAM+ve SRAM Expression Signatures. Signatures were defined as the mean expression (as a *z* score) for each of the SRAM genes in *EPCAM* +ve or −ve cell lines. These were applied to tumor samples by calculating the Spearman rank correlation (Rho) between signatures and scaled expression values for the individual tumors (z-scores). A total of 69 SRAM genes were present in the tumor dataset. High scores mean that SRAM genes that were relatively highly expressed in *EPCAM*+ve vs. *EPCAM*−ve cell lines were highly expressed in that particular tumor relative to the other 243 and vice versa.

Expression Panel of Epithelial and Mesenchymal Markers. Mesenchymal markers were taken from Herschkowitz et al. (4), epithelial markers from Allinen et al. (5), and keratins from Malzahn et al. (6).

Genome Annotation. All platforms used in this study were annotated to Ensembl 54 gene IDs (NCBI36). The annotated position of each CpG assayed on the Infinium arrays was mapped to the closest Ensembl gene based on transcriptional start site (TSS) location. CpGs that ambiguously mapped to more than one gene ID were removed from the analysis. Illumina expression probes were directly mapped to the Ensembl 54 annotation by using BLAST with the Ensembl cDNA and ncRNA sequence sets (ungapped alignment). Probes were mapped if they matched at least one transcript with no more than 2 mismatches and did not match transcripts from another gene with <10 mismatches. CpG island locations were taken from those biologically defined in a recent study (7).

Public Datasets. Expression data for the breast cancer cell lines was from Neve et al. (8). Raw data were downloaded from Array Express (E-TABM-157) and processed by using the RMA algorithm (Bioconductor *affy* package) and an updated annotation (U133A, Ensembl gene CDF Version 11; ref. 9). Probe set calls used to define silent genes were similarly generated, but by using the MAS5 algorithm. Raw gene expression data for normal tissues was from Ge et al. (10) and was similarly processed (GEO accession no. GSE2361). Processed data describing the gene expression patterns of cells from the normal breast was from Lim et al. (ref. 11; GEO accession no. GSE16997). Illumina expression probes were mapped to the Ensembl annotation as above. Mean expression values were calculated for genes with multiple probes. Gene expression signatures for different cellular fractions in the normal breast were taken from the supplemental materials of the same study, and probes were mapped to Ensembl genes as above. Processed expression data from breast tumors was from Hennessy et al. (ref. 12; GEO accession no. GSE10885). Refseq IDs for array probes were taken from GEO and mapped to Ensembl gene IDs by using Ensembl Biomart. Where multiple probes mapped to a single gene ID, mean values were calculated, and those mapping to none or multiple IDs were discarded. Clinical annotation (breast cancer subtypes) was taken from the annotation included with the GEO series.

Data on the histone modification status of genes in human ES (hES) cells was from the supplementary data of Zhao et al. (13). Locations of blocks of histone modifications were updated to NCBI36 and assigned to genes if they were within 1 kb of a TSS.

DNA methylation data for breast tumors on the Goldengate array were from Holm et al. (14). Processed methylation data were taken from GEO (accession no. GSE22210), and Goldengate probes were mapped to gene IDs in the same manner as the Infinium probes (see above). We determined the frequency of methylation of genes by using the median level of probes within 200 bp of TSS. Frequently aberrantly methylated genes were defined as those unmethylated (≤30% methylation) in all of the normal samples from the study and methylated (>30% methylation) in at least 20% of the tumors (≥38 tumors). We determined frequently methylated genes from Pubmeth (www.pubmeth.org; ref. 15) by searching by cancer type for "all breast cancer." We excluded any genes with <100 samples analyzed and <20% methylated. The search was conducted on July 29, 2010. Genes commonly methylated in Hill et al. (16) were taken from figure 1B in that study. Only those genes methylated in ≥20% of tumors were used. The total sizes of the three lists of frequently methylated genes were: Pubmeth, 35 genes, of which 34 were in the Lim et al. dataset (11); Holm et al. (14), 78 genes, of which 74 were present in the Lim et al. dataset (11); and Hill et al. (16), 10 genes, of which 9 were present in the Lim et al. dataset (11).

Analysis of Gene Ontology (GO) Terms/Enrichments. To analyze functional terms, Ensembl Biomart was used to map gene identifiers to Gene Ontology biological process terms (Ensembl 54). Enrichment of specific terms in each gene list was then assessed by using Fisher's exact test compared with all genes present on the Infinium array. Terms that were associated with <10 genes on the Infinium arrays were excluded from the analysis.

Consensus Clustering. Hierarchical clustering was performed in R by using the Euclidian distance and the Ward algorithm. Consensus clustering was performed to estimate the robustness of each sample (17); 500 iterations of the clustering were used to estimate robustness in each case. The consensus clustering algorithm was implemented by T.I. Simpson (University of Edinburgh; ref. 18).

Methylation Status of Common Tumor Suppressor Genes. We defined the methylation status of common tumor suppressor genes by using the median methylation of probes within 200 bp of their TSS. The numbers of probes found at each gene and their locations relative to the TSS are shown in Table S4.

1. Bock C, et al. (2005) BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics* 21:4067–4068.
2. Martinez O, Reyes-Valdes MH (2008) Defining diversity, specialization, and gene specificity in transcriptomes through information theory. *Proc Natl Acad Sci USA* 105: 9709–9714.
3. Turner N, et al. (2010) Integrative molecular profiling of triple negative breast cancers identifies amplicon drivers and potential therapeutic targets. *Oncogene* 29:2013–2023.
4. Herschkowitz JI, et al. (2007) Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol* 8:R76.
5. Allinen M, et al. (2004) Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell* 6:17–32.
6. Malzahn K, Mitze M, Thoenes M, Moll R (1998) Biological and prognostic significance of stratified epithelial cytokeratins in infiltrating ductal breast carcinomas. *Virchows Arch* 433:119–129.
7. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr AR, James KD Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet* 6:e1001134.
8. Neve RM, et al. (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 10:515–527.
9. Dai M, et al. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 33:e175.
10. Ge X, et al. (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics* 86:127–141.
11. Lim E, et al. (2009) Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med* 15:907–913.
12. Hennessy BT, et al. (2009) Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. *Cancer Res* 69:4116–4124.
13. Zhao XD, et al. (2007) Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell* 1:286–298.
14. Holm K, et al. (2010) Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. *Breast Cancer Res* 12:R36.
15. Ongenaert M, et al. (2008) PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res* 36:D842–D846.
16. Hill VK, et al. (2010) Identification of 5 novel genes methylated in breast and other epithelial cancers. *Mol Cancer* 9:51.
17. Monti S, Tamayo P, Mesirov J, Golub T (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach Learn* 52:91–118.
18. Simpson TI, Armstrong JD, Jarman AP (2010) Merged consensus clustering to assess and improve class discovery with microarray data. *BMC Bioinformatics* 11:590.

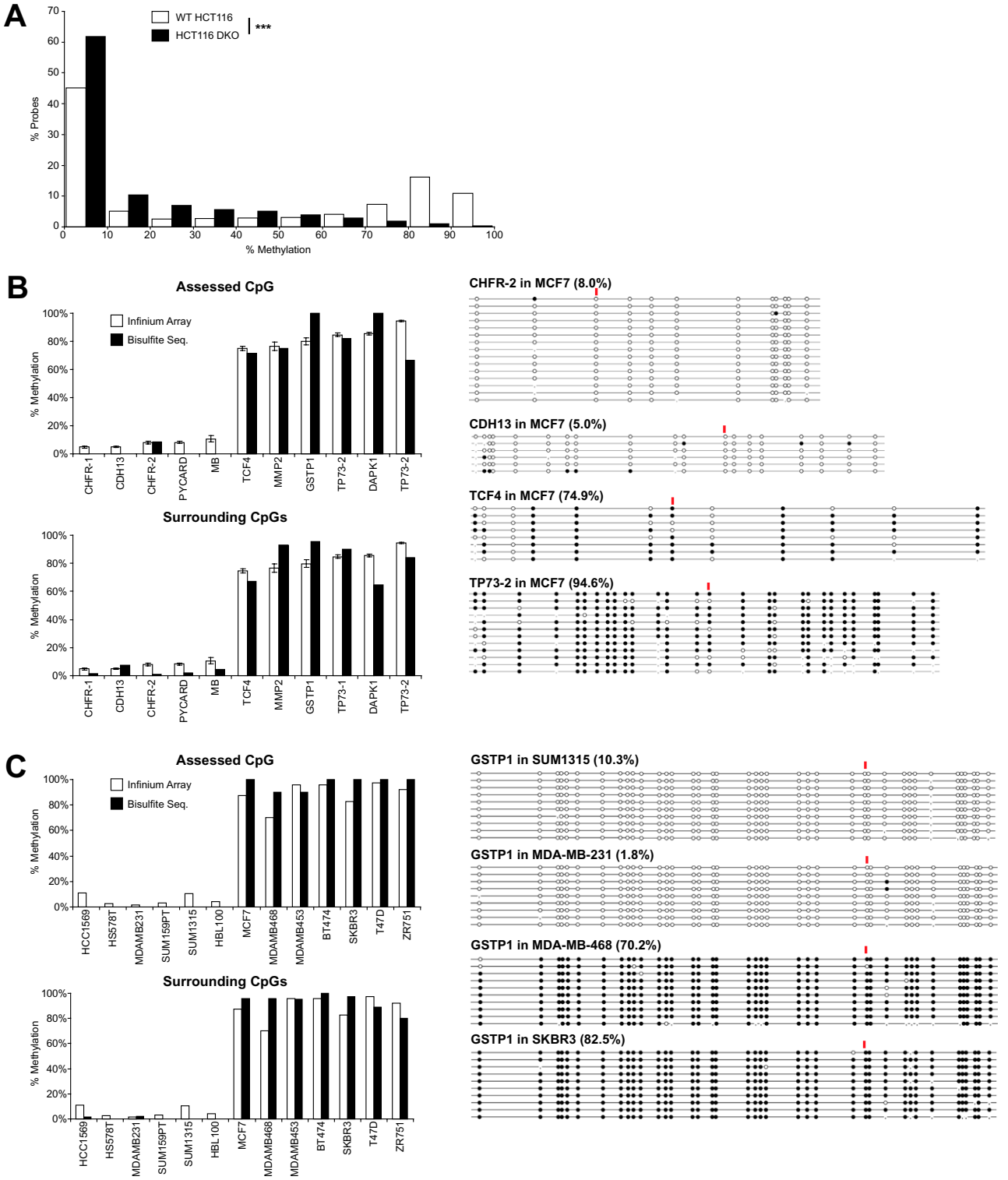

Fig. S1. (Continued)

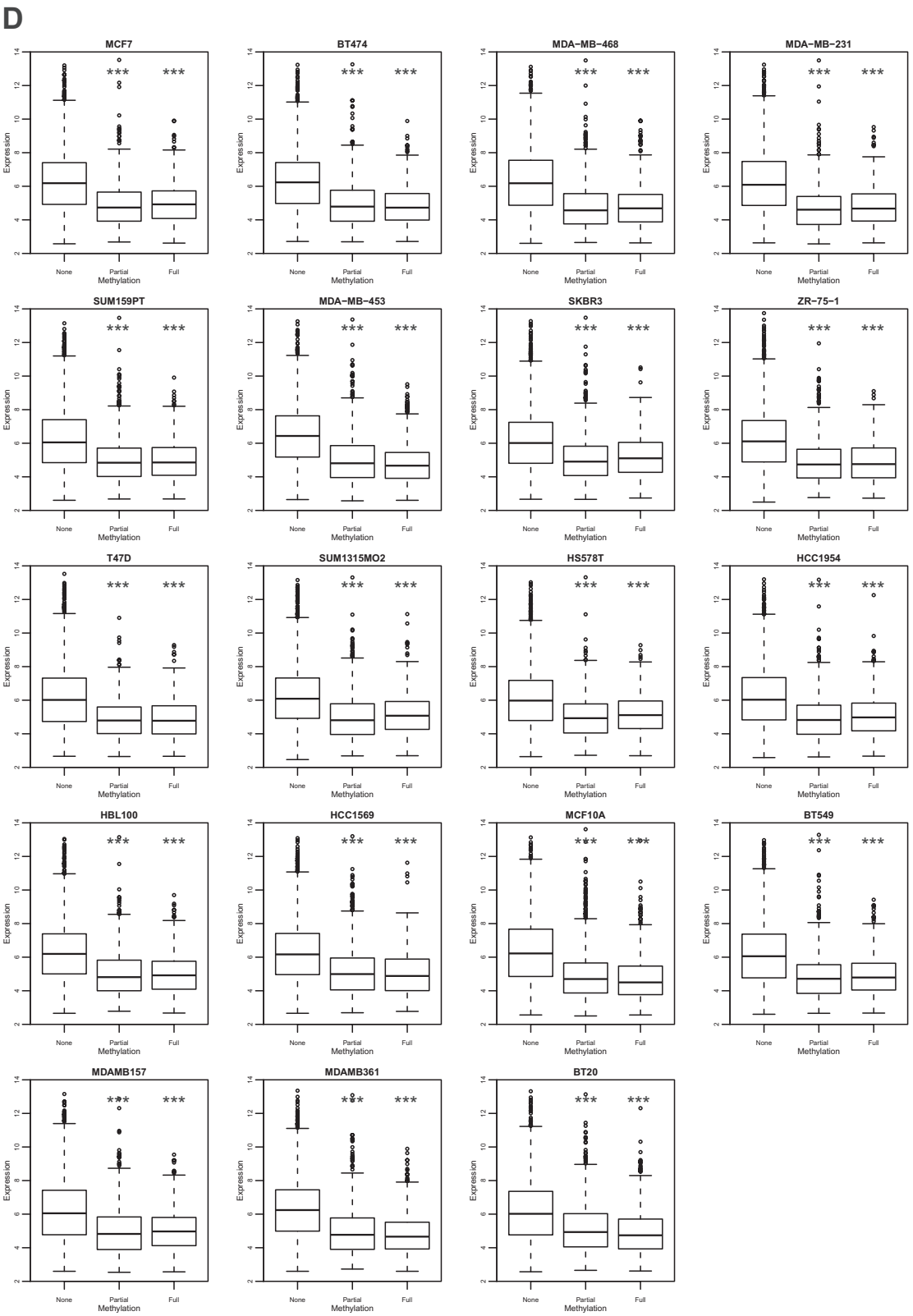

Fig. S1. (Continued)

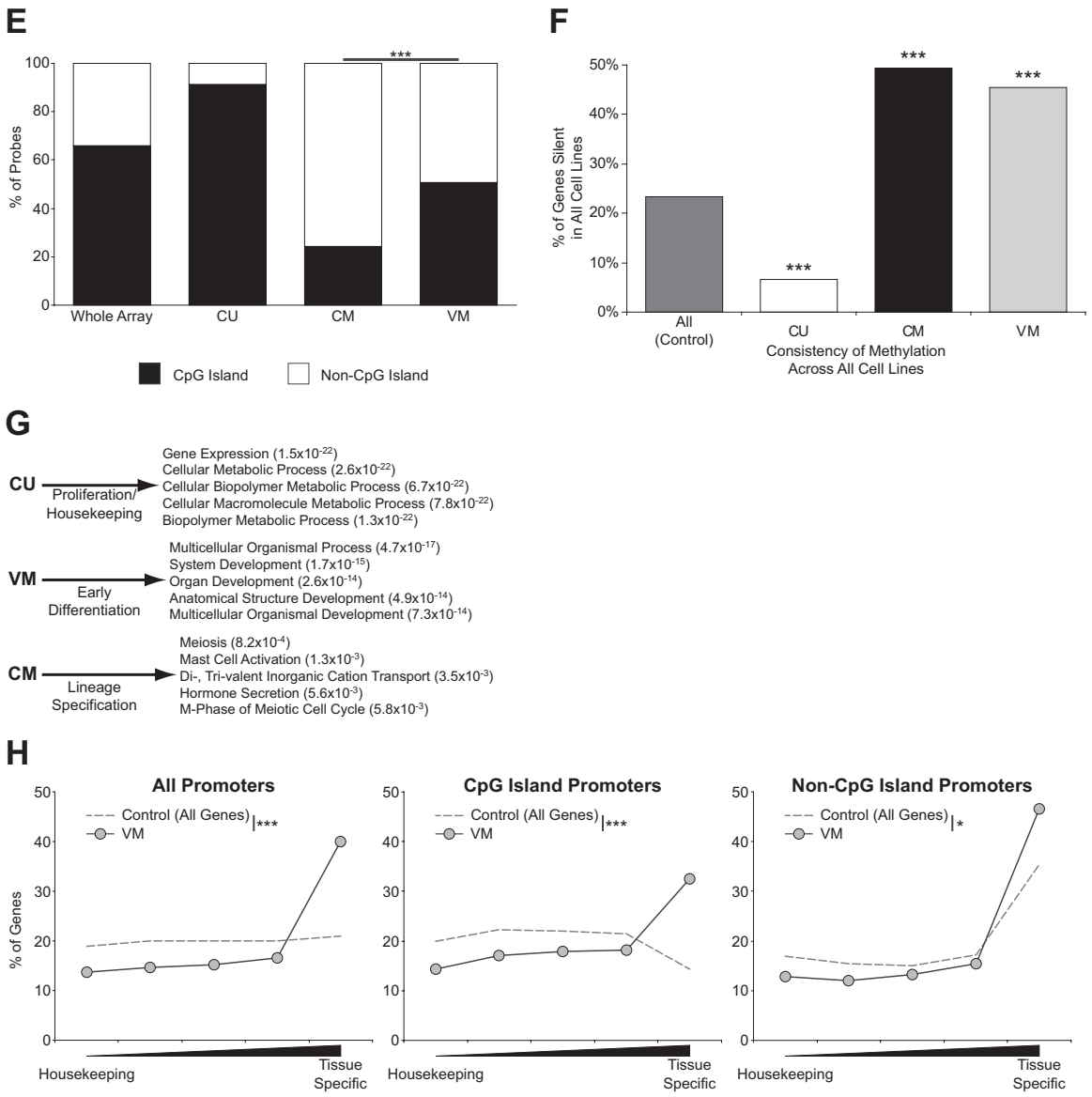

Fig. S1. (*A*) Illumina Infinium arrays can distinguish cell lines with different levels of DNA methylation. Shown is a histogram of the percentage methylation reported by all probes on the array for HCT116 colon cancer cells and a modified version of the cell line in which the DNA methyltransferases had been genetically knocked out (HCT116 DKOs). The distributions are significantly different ($P < 2.2 \times 10^{-16}$, paired Wilcoxon test), and probes reporting a high percentage of methylation are very rare in the HCT116 DKO cells. (*B*) Illumina Infinium arrays reliably report methylation levels at different genes in the same cell line. We compared the methylation levels reported for individual CpGs by the Infinium arrays to either the same CpG or the mean level of the CpG and its 10 nearest neighbors by bisulfite sequencing at a selection of genes in MCF7 cells. Multiple CpGs are shown for some genes, and the error bars represent the SEM for five replicates of MCF7 cells. Sequencing diagrams are shown for multiple clones for some of the genes with CpGs represented by circles. Filled circles were unconverted by bisulfite treatment and are methylated. Open circles were converted and therefore unmethylated. Missing circles mean sequencing was of a low quality across that CpG in that particular clone. CpGs assessed by the Infinium array are highlighted, and the reported methylation level for that CpG on the array is shown beside the gene name. (*C*) Illumina Infinium arrays reliably report methylation levels across different cell lines. As for *B*, but the region surrounding the transcription start site of GSTP1 was assessed in multiple cell lines. (*D*) Expression is inversely related to methylation in breast cancer cell lines. The methylation levels of genes were defined as none (≤30% methylation), partial (>30% and <70% methylation), or full (≥70% methylation) for each cell line based on the median of all probes within 200 bp of transcriptional start sites. The expression levels of these groups of genes were then compared (from Neve et al.; ref. 1). $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, using Wilcoxon tests. For this analysis only CpGs that were within 200 bp of a TSS were considered, and only genes present in the gene expression data were used (6,050 genes). The methylation value for each gene was the median of all probes within 200 bp of its TSS. (*E*) VM genes include CpG island and non-CpG island genes. We used a set of biologically assayed CpG islands to define CpGs on the Infinium arrays as CpG island or non-CpG island (2). We then calculated the percentage of CU, CM, and VM probes that were within CpG island genes. CM and VM probes were compared by using a Fisher's exact test. (*F*) CM and VM genes are enriched in genes silent in all breast cancer cell lines. We used expression data from the cell lines to define silent genes as those for which expression levels were called as "absent" or not significant above background in all 19 cell lines (from Neve et al.; ref. 1). The proportion of genes that were silent in each of our gene groups was then calculated and plotted. Significance was assessed by comparing the proportions of silent genes in the gene groups with the proportion of silent genes on the entire array using Fisher's exact tests. (*G*) Different methylation patterns in breast cancer cell lines are associated with different functional groups of genes. The top five biological process GO terms significantly enriched in each of our gene sets is shown. Significance was assessed using Fisher's exact tests, and *P* values are shown in brackets adjacent to each term. (*H*) VM CpG island genes are expressed in a tissue-specific manner. Shown are the expression patterns of VM genes in normal tissues (as in Fig. 1*C*) compared with all genes on the array. The patterns for all promoters are compared with those for just CpG island or non-CpG island promoters. Significance was assessed using χ^2 tests ($*P <$ 0.05, $**P < 0.01$, and $***P < 0.001$).

1. Neve RM, et al. (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 10:515–527.
2. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr AR, James KD Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet* 6:e1001134.

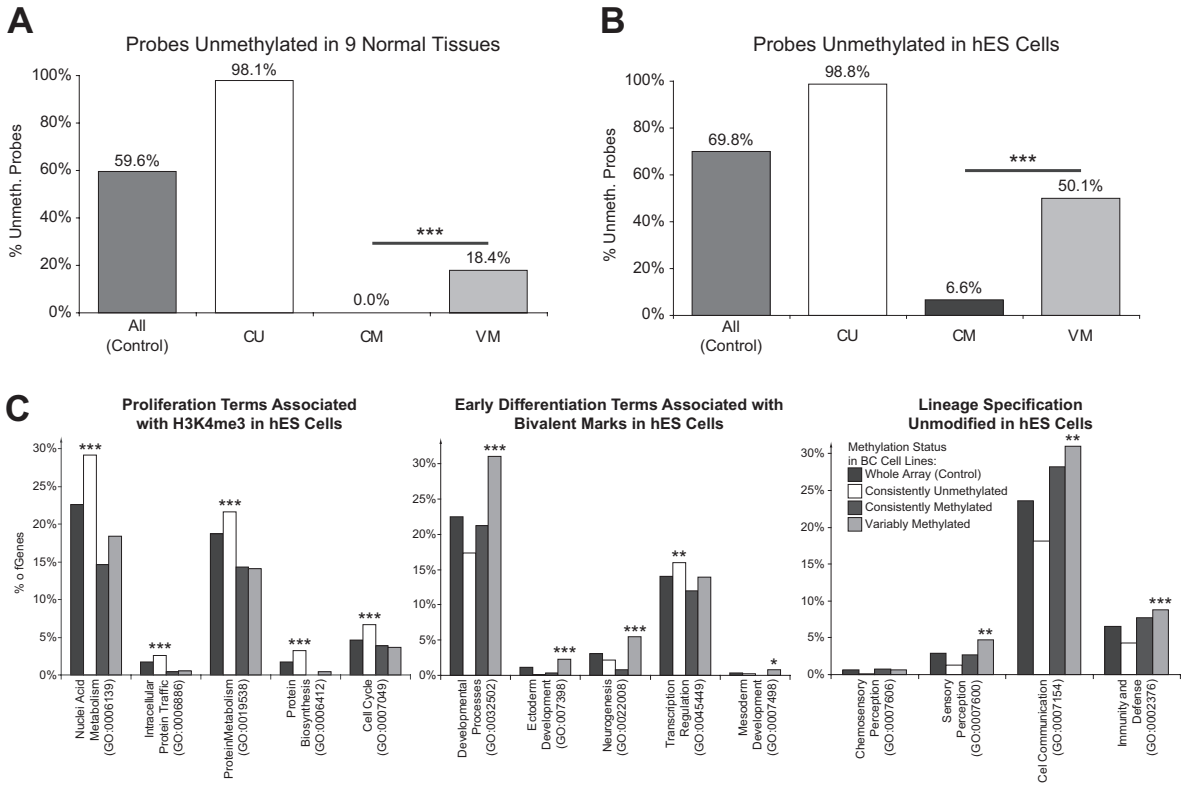

Fig. S2. (*A*) VM genes are more frequently unmethylated than CM genes in normal tissues. We defined probes that were unmethylated in all nine of a panel of nine normal tissues and analyzed their proportions in the CU, CM, and VM probe sets. CM and VM probes were compared by using Fisher's exact tests. (*B*) VM genes are more frequently unmethylated than CM genes in hES cells. The proportions of CU, CM, and VM probes that are unmethylated in hES cells are shown. CM and VM probes were compared by using Fisher's exact tests. (*C*) Functional terms associated with genes with bivalent histone marks in hES cells are also enriched in genes that show a variable methylation pattern in breast cancer cell lines. We identified the closest GO terms to the functional terms reported as being associated with different histone marks in hES cells in Zhao et al. (1). We then analyzed their enrichment in the sets of genes associated with different DNA methylation patterns in breast cancer cell lines. The term reported by Zhao et al. (1) is shown, and the closest GO term ID we identified is indicated in brackets below. *P* values for enriched terms were calculated by comparing the number of times a term appeared in each gene list to the number of times it appeared on the whole array using a Fisher's exact test. The term "segment specification" ("early differentiation" in Zhao et al.; ref. 1), which matches GO:0007379, was excluded from the analysis, as only three genes were found on the Infinium methylation array that mapped to this term. The term "cell communication" ("lineage specification" in Zhao et al.; ref. 1), which matches GO:0007154, was excluded from the analysis, as no genes were found on the Infinium methylation array that mapped to this term. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

1. Zhao XD, et al. (2007) Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell* 1: 286–298.

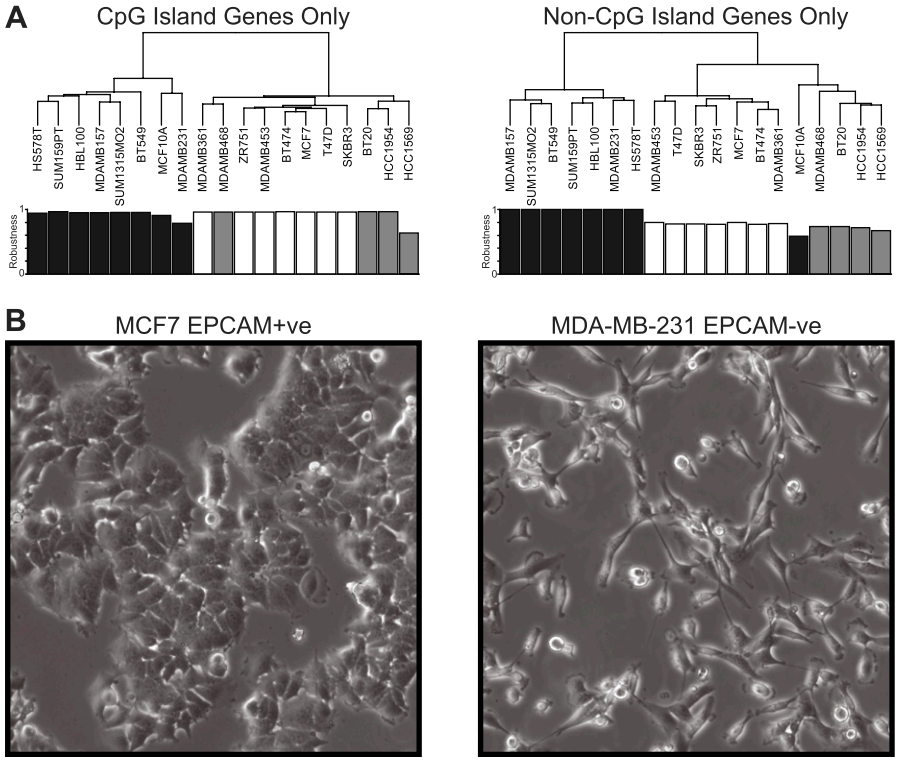

Fig. S3. (*A*) The relationships of cell lines are similar using only SRAM genes with CpG island promoter genes. Fig. 3*C* is redrawn using only SRAM genes with CpG island promoters or SRAM genes with non-CpG island promoters. Cell lines are colored as Fig. 3 (White, luminal A; gray, basal A; and black, basal B). Robustness was calculated by using consensus clustering. (*B*) Cell lines with different methylation profiles show different morphologies. Shown are phase contrast images of two representative cell lines from each of the groups we observed: MCF7, an *EPCAM*+ve cell line with an epithelial morphology, and MDA-MB-231, an *EPCAM*−ve cell line with a fibroblast/spindle-cell like morphology.

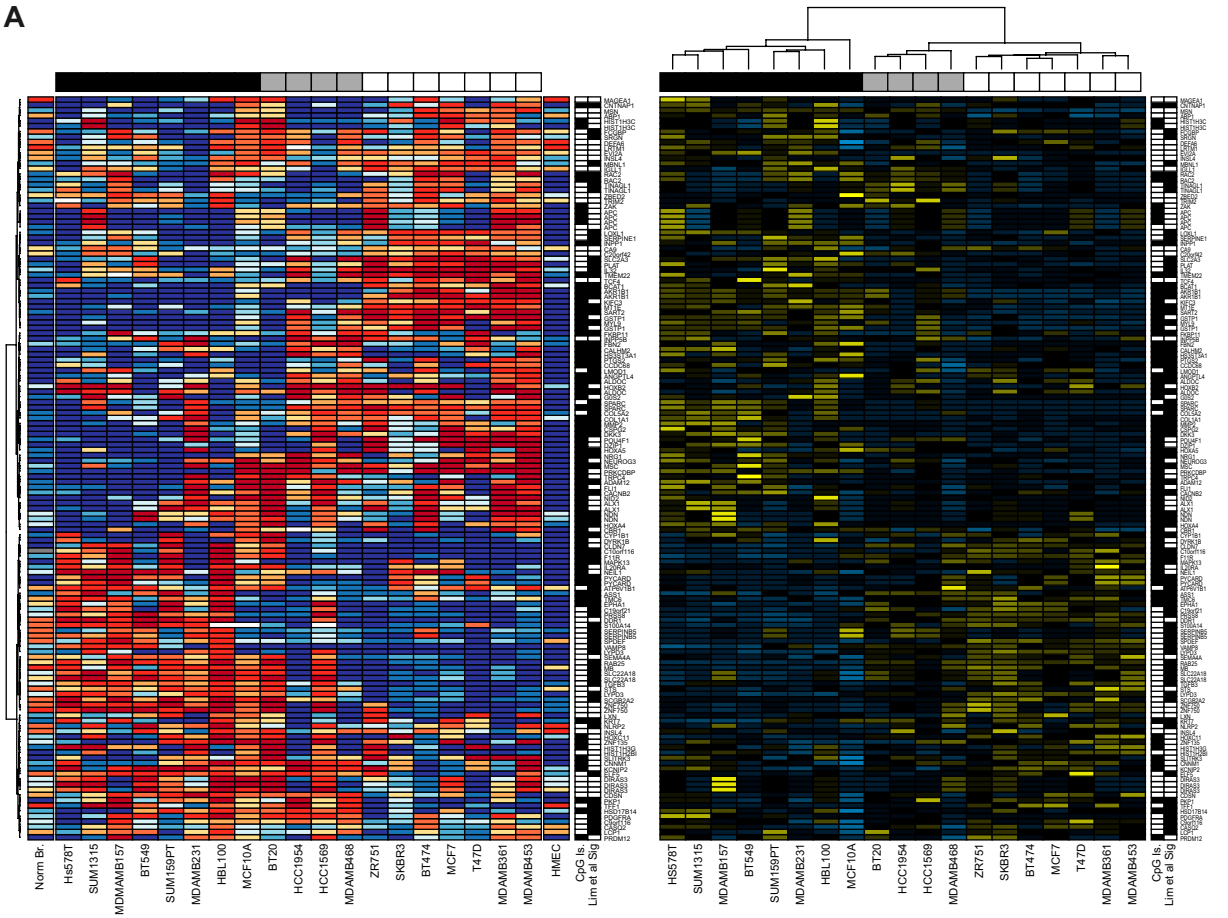

Fig. S4. (Continued)

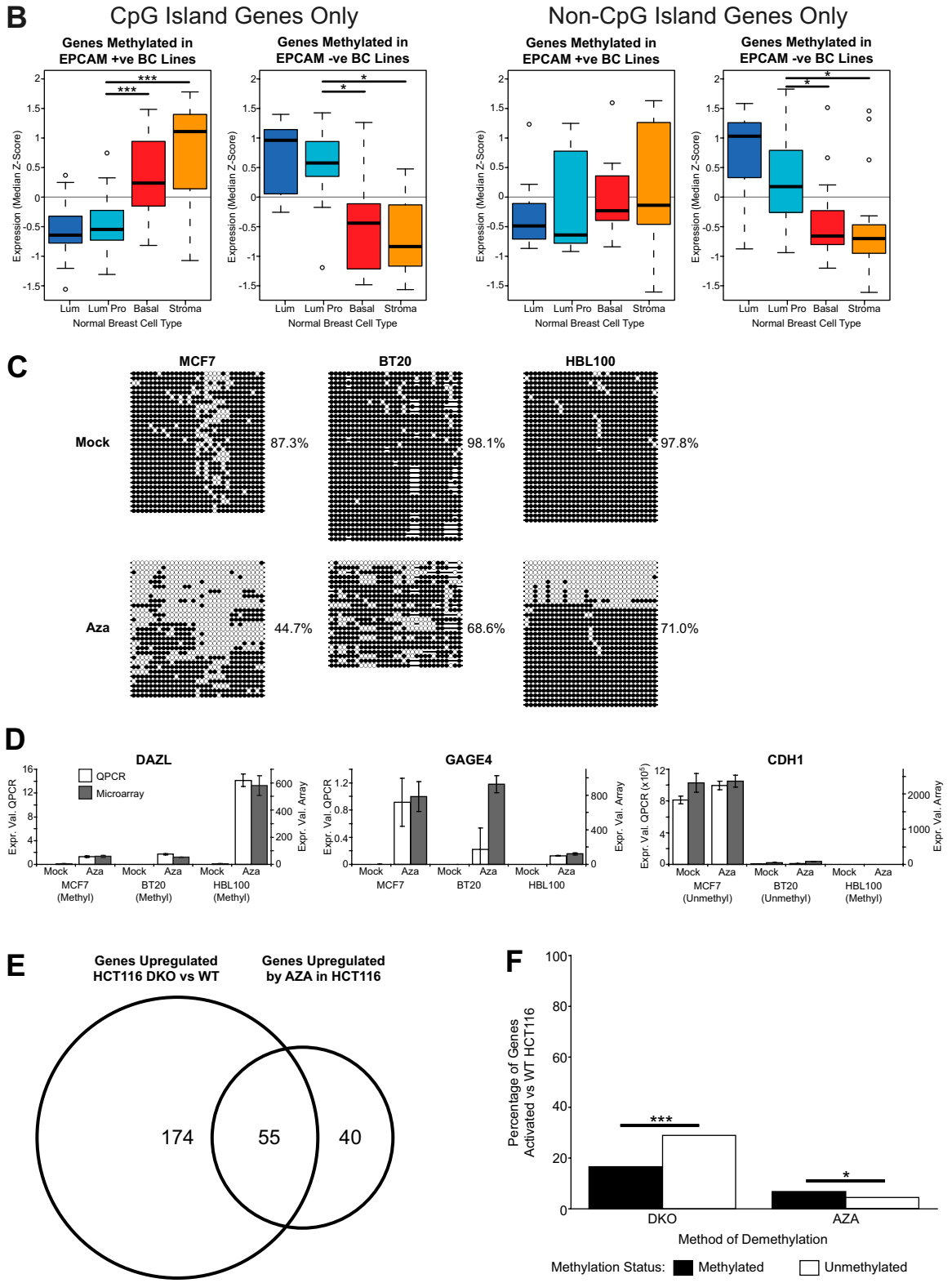

Fig. S4. This figure is an enlarged version of Fig. 4*A* with additional information. (*A*) Methylation and expression levels of SRAM genes are associated with *EPCAM* status. Shown are the results of unsupervised hierarchical clustering of cell lines using SRAM genes. Cell lines are clustered based on expression values, and genes are clustered based on methylation values. The expression values and methylation values of the SRAM genes are color coded on the figure. Expression values (*Right*) are *z* scores, and methylation data (*Left*) are given as percentage methylation by probe. The cell lines are also color-coded by type (top bar: white, luminal A; gray, basal A; black, basal B). Multiple probes are shown for some genes, illustrating their concordance. Also shown is methylation data for the same genes in normal breast tissue (*Left*, furthest left bar) and HMEC cells (*Left*, third bar from right). The rightmost sidebars on both panels show genes that are differentially expressed (indicated in black) in different lineages from normal breast tissue in Lim et al. (1). Adjacent to this sidebar is a further bar that

Legend continued on following page

indicates probes that are located within CpG islands in black (according to Illingworth et al.; ref. 2). (*B*) SRAM genes with CpG promoters that are preferentially methylated in *EPCAM*+ve or −ve cell lines are also repressed in the corresponding normal lineages. Shown are reproductions of Fig. 4*B* but using either SRAM genes that have CpG island promoters or those without CpG island promoters. Significance was assessed by using Wilcoxon tests. (*C*) 5-aza-2′-deoxycytidine (5-aza-dC) demethylates breast cancer cell lines. Bisulfite sequencing of the promoter of the *DAZL* gene is shown after 5-aza-dC treatment or with mock treatment in the three breast cancer cell lines assayed for reactivation of genes by 5-aza-dC (Fig. 4 *C* and *D*). Multiple clones are shown in each panel with methylated CpGs indicated by filled circles and unmethylated ones by open circles. Missing circles mean sequencing was of a low quality across that CpG in that particular clone. The percentage of methylated CpGs is also shown. (*D*) Illumina expression arrays reliably detect the reexpression of methylated genes. Shown are expression levels of individual genes in three breast cancer cell lines after either 5-aza-dC or mock treatment as assayed by quantitative RT-PCR (QPCR) or by Illumina expression arrays. Where known, methylation status at that gene is indicated below each cell line. DAZL is methylated in all cell lines (see *C*), and its expression is controlled by methylation in normal development (3). *GAGE4* is a cancer testis antigen and thus would be expected to be reexpressed upon 5-aza-dC treatment (4). Due to segmental duplications, we were unable to determine its methylation status in the cell lines analyzed. Both of these genes were detected as derepressed after 5-aza-dC treatment by QPCR and array analysis. In contrast *CDH1* is methylated in HBL100 cells but not the other cell lines and is not detected as being reactivated by 5-aza-dC treatment in these cells by either platform. (*E*) 5-aza-dC activates a subset of genes controlled by DNA methylation. We examined how many genes that were methylated and activated by 5-aza-dC were also activated by knocking out the DNA-methyltransferases. The two gene lists overlapped significantly (Fisher's exact test: $P < 2.2 \times 10^{-16}$). (*F*) Genetic ablation of DNA methylation reactivates only a small proportion of methylated genes. We compared the number of genes reactivated by 5-aza-dC in wild-type HCT116 cells with the number reactivated by genetic deficiency of DNA methyltransferases in HCT116 DKO cells. For both comparisons the numbers of methylated and unmethylated genes are plotted. The specificity of activation for methylated genes was tested using Fisher's exact tests. In the DKO cells, significantly more unmethylated genes were activated, and with 5-aza-dC treatment significantly more methylated genes were activated. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

1. Lim E, et al. (2009) Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med* 15:907–913.
2. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr AR, James KD Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet* 6:e1001134.
3. Maatouk DM, et al. (2006) DNA methylation is a primary mechanism for silencing postmigratory primordial germ cell genes in both germ cell and somatic cell lineages. *Development* 133:3411–3418.
4. Kumagai T, et al. (2009) Epigenetic regulation and molecular characterization of C/EBPalpha in pancreatic cancer cells. *Int J Cancer* 124:827–833.

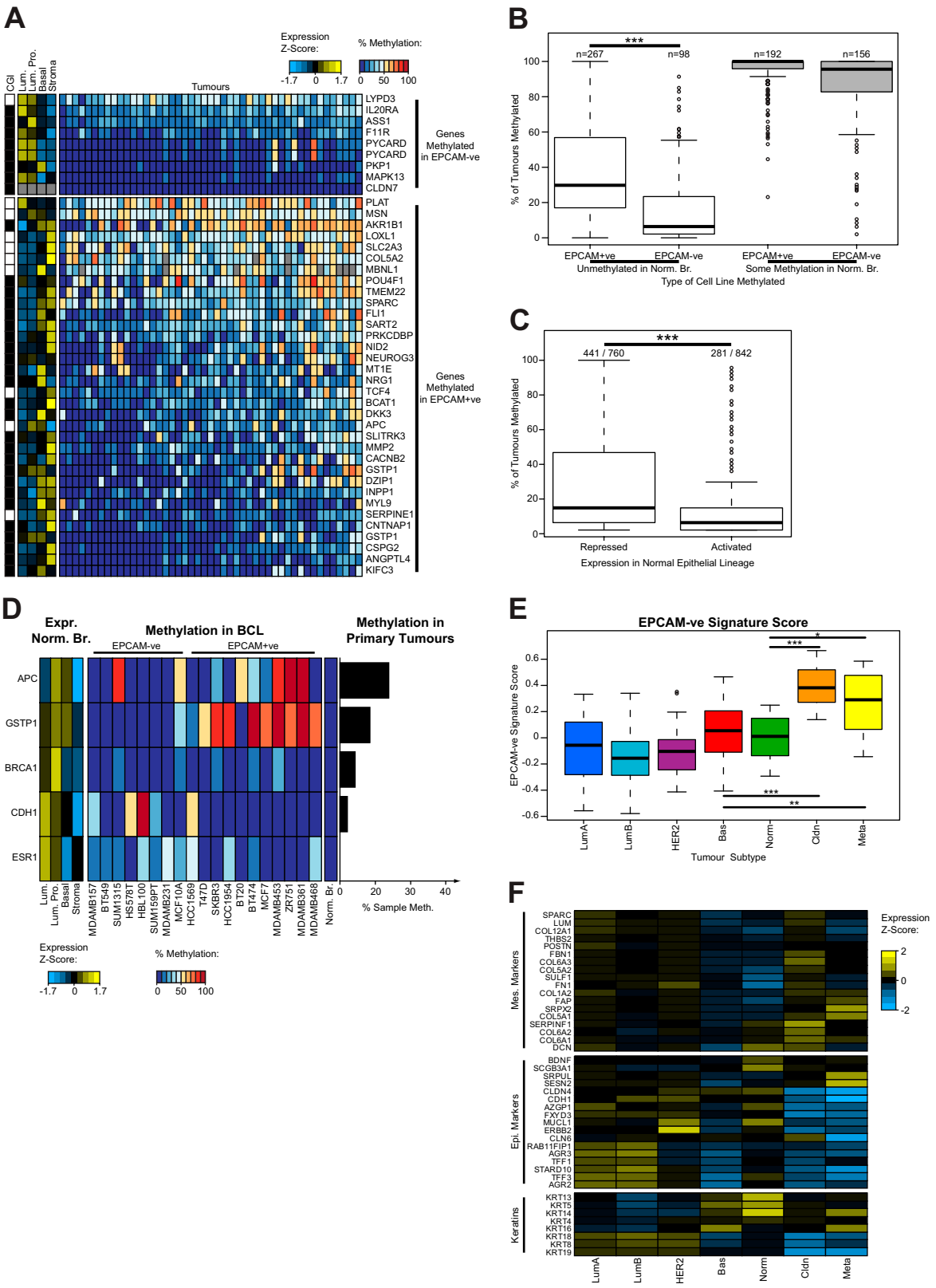

Fig. S5. (*A*) Primary tumors show differential methylation of SRAM genes preferentially methylated in *EPCAM*+ve and −ve cell lines. The heat map shows the methylation frequency of differentially methylated SRAM genes in 47 primary breast tumors. Only genes that are unmethylated in the normal breast are shown. Genes and samples are ordered by their frequency of methylation. Multiple probes are shown for some genes. This is a larger version of Fig. 5*A* and includes additional data. The sidebars indicate which probes are in CpG islands (as defined in Illingworth et al.; ref. 1) and the expression of each gene in different cell types in the normal breast (2). Expression is shown as median *z* scores. Missing data are indicated in gray. Lum, luminal epithelial cells; Lum. Pro.,

Legend continued on following page

luminal epithelial progenitors; Basal, basal myoepithelial cells; Stroma, stromal fraction. (*B*) VM genes that are preferentially methylated in *EPCAM*+ve cell lines are more frequently aberrantly methylated in primary tumors than those preferentially methylated in *EPCAM*−ve cell lines. We defined a broad set of probes variably methylated in cell lines (methylated in at least one cell line and unmethylated in at least one cell line) that showed a significant difference in their level of methylation between *EPCAM*+ve and −ve cell lines. The frequencies of methylation of these probes in primary tumors are shown separately for those genes unmethylated in normal breast tissue and for those methylated in normal breast tissue. The number of probes in each group is indicated. Significance was assessed using a Wilcoxon test. (*C*) Genes specifically repressed in the normal luminal epithelial lineage are frequently methylated in breast tumors. We defined genes specifically repressed or activated in the luminal epithelial lineage as those with significantly lower or higher expression in both luminal cells and luminal progenitors compared with stromal cells (Welch's *t* tests, $P < 0.05$; from Lim et al.; ref. 2). Genes from these two sets that were unmethylated in the normal breast were then analyzed for their methylation frequency in primary breast tumors. Shown is a boxplot of the frequency of methylation for those genes in each set methylated in one or more tumors. Indicated above each set is the number of genes in that set (denominator) and the number methylated in one or more tumors (numerator). The distributions were compared by using a Wilcoxon test. (*D*) Known breast cancer tumor suppressor genes and putative tumor suppressor genes show different expression patterns in normal luminal progenitor cells and differentiated luminal cells and have different rates of methylation in tumors and tumor cell lines. Shown are the expression patterns of the indicated genes in different lineages from the normal breast (from Lim et al.; ref. 2) derived from the median *z* scores, along with their methylation patterns in breast cancer cell lines and their frequency of methylation in primary tumors. Methylation was assayed as the median percentage methylation of probes within 200 bp of each gene's TSS. Cell lines are ordered by their *EPCAM* expression level and genes by their frequency of methylation in primary tumors. (*E*) SRAM gene expression in claudin-low and metaplastic tumors resembles that in *EPCAM*−ve cell lines. We used an SRAM gene signature to score tumors based on their similarity to *EPCAM*−ve breast cancer cell lines, and the scores were compared between tumor subtypes. Claudin-low and metaplastic tumors have scores that are significantly higher than all other subtypes (Wilcoxon tests). Tumor subtypes: LumA, luminal A; LumB, luminal B; HER2; Bas, basal; Norm, normal like; Cldn, claudin-low; Meta, metaplastic. (*F*) Claudin-low and metaplastic tumors express mesenchymal markers. Shown are the mean expression levels (as *z* scores) by tumor type for different sets of marker genes. Mesenchymal (Mes.) markers were taken from Herschkowitz et al. (3); epithelial (Epi.) markers were from Allinen et al. (4), and keratins were from Malzahn et al. (5) (as in Fig. 3*E*). Mesenchymal-specific genes are more highly expressed in claudin-low and metaplastic tumors, whereas epithelial and cytokeratin genes are more highly expressed in all other tumor types. Tumor subtypes are labeled as in *E*. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

1. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr AR, James KD, et al. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet* 6: e1001134.
2. Lim E, et al. (2009) Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med* 15:907–913.
3. Herschkowitz JI, et al. (2007) Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol* 8:R76.
4. Allinen M, et al. (2004) Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell* 6:17–32.
5. Malzahn K, Mitze M, Thoenes M, Moll R (1998) Biological and prognostic significance of stratified epithelial cytokeratins in infiltrating ductal breast carcinomas. *Virchows Arch* 433: 119–129.

Table S1. Genes frequently methylated in primary breast tumors are significantly enriched in a set of genes specifically repressed in the luminal epithelial lineage

Study	Technique	No. of frequently methylated epithelial repressed genes (≥20% tumors)	Significance
Pubmeth (1)	Mainly MSP	10	0.02
Holm et al. (2)	Illumina Golden Gate	30	0.005
Hill et al. (3)	MIRA with COBRA verification	5	0.005

We assessed whether a set of genes repressed in luminal epithelial cells (see Fig. S5*C*) was enriched in genes reported to be frequently methylated in breast cancer. Gene lists drawn from three different sources were all significantly enriched in this set of genes (using Fisher's exact tests compared with all genes assessed).

1. Ongenaert M, et al. (2008) PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res* 36:D842–D846.
2. Holm K, et al. (2010) Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. *Breast Cancer Res* 12:R36.
3. Hill VK, et al. (2010) Identification of 5 novel genes methylated in breast and other epithelial cancers. *Mol Cancer* 9:51.

Table S2. Bisulfite PCR primers

Primer	Round	Forward primer	Reverse primer	Annealing temp., °C
MB	1	GTATTTAGTGTATATTAGGG	CAACCCTAAAACAAAATCAC	44
MB	2	GTAGGAGATATTTTTATAAG	CTAAACAAACTCAATCCAAA	42
MMP2	1	GAGAGAGGTAAGTGGGGTGA	CCTAATTAAAACCTACTCC	45
MMP2	2	GTAGAGGTTAGGAGTAGTAG	ATAACCTAAAATTTACCC	39
TCF4	1	GAATTGTAAGTTTAGTAAAG	CAATTATACTATTCTATAAC	39
TCF4	2	GGGTAGGTTAGGATGTATTT	AAATATACAATTCAAATTTC	37
CDH13	1	GTAGAGAAAAGTTTAAGTTTTG	TTATCCACCCACTTACAAAC	44
CDH13	2	AGTTGTTTGTTAATTTTTAG	AACTCACTCCAAATCCCAAC	37
CHFR	1	GGTTATTTTTGATTTTGATTAGG	CACTTTCAAAAAATACCCTCTAAC	44
CHFR	2	TTATGTTATGTTGGGGTAGAAGGG	CACCCTACCCACAAACAACC	52
DAPK1	1	GTTTTTGGAGGTGGGAAAGTTG	TAATAATAAAATAACAACCCC	41
DAPK1	2	ATGTGTGTAGAGAAAGGGGAG	ACACCCTTTATTAAAACTAAAC	44
GSTP1	1	TTGTTTGTTTATTTTTTAGG	AATTAACCCCATACTAAAAAC	37
GSTP1	2	ATTTGGGAAAGAGGGAAAGG	AACTCTAAACCCCATCCCC	48
PYCARD	1	GGTTTTAGAGTTTGGAAGG	TCAACTTCTACCTAAAAACC	44
PYCARD	2	GGAAGGATATGGGTTAAGTG	ACATAAACCTACAAAAAATAACC	44
TP73	1	AGTTAGTTGATAGAATTAAG	TCACCCCAACTAACAACAAAC	40
TP73	2	ATTAAGGGAGATGGGAAAAG	CCCTACACTACAACAAAATC	46
DAZL	Both	GAAGAGAAAAGGAAAATTAAGAG	CCTTCCTAAAACTAAAACA	50

Table S3. Quantitative RT-PCR primers

Gene	Forward primer	Reverse primer	Annealing temp, °C
CDH	GACCAAGTGACCACCTTAGA	CTCCGAAGAAACAGCAAGAGC	57
DAZL	ACACTGAAACTTATATGCAGCCC	CGGAGGTACAACATAGCTCCTTT	57
GAGE4	ACACCTGAAGAAGGGGAACC	TTCACCTCCTCTGGATTTGG	57

Table S4. Positions of CpG probes relative to the TSS for the indicated genes

Gene	Ensembl ID	No. of probes	Probe locations relative to TSS
APC	ENSG00000134982	5	−151, −82, −14, 102, 185
GSTP1	ENSG00000084207	2	−20, −10
BRCA1	ENSG00000012048	6	−82, −29, 24, 72, 85, 146
CDH1	ENSG00000039068	2	5, 8
ESR1	ENSG00000091831	1	57

Other Supporting Information Files

Dataset S1 (XLS)