nature genetics

The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores

Rafael A Irizarry[1,2,8], Christine Ladd-Acosta[2,3,8], Bo Wen[2,3], Zhijin Wu[4], Carolina Montano[2,3], Patrick Onyango[2,3], Hengmi Cui[2,3], Kevin Gabo[2,3], Michael Rongione[2,3], Maree Webster[5], Hong Ji[2,3], James B Potash[2,6], Sarven Sabunciyan[2,7] & Andrew P Feinberg[2,3,8]

For the past 25 years, it has been known that alterations in DNA methylation (DNAm) occur in cancer, including hypomethylation of oncogenes and hypermethylation of tumor suppressor genes. However, most studies of cancer methylation have assumed that functionally important DNAm will occur in promoters, and that most DNAm changes in cancer occur in CpG islands. Here we show that most methylation alterations in colon cancer occur not in promoters, and also not in CpG islands, but in sequences up to 2 kb distant, which we term 'CpG island shores'. CpG island shore methylation was strongly related to gene expression, and it was highly conserved in mouse, discriminating tissue types regardless of species of origin. There was a notable overlap (45–65%) of the locations of colon cancer–related methylation changes with those that distinguished normal tissues, with hypermethylation enriched closer to the associated CpG islands, and hypomethylation enriched further from the associated CpG island and resembling that of noncolon normal tissues. Thus, methylation changes in cancer are at sites that vary normally in tissue differentiation, consistent with the epigenetic progenitor model of cancer, which proposes that epigenetic alterations affecting tissue-specific differentiation are the predominant mechanism by which epigenetic changes cause cancer.

Since the discovery of altered DNA methylation in human cancer[1], the focus has largely been on specific genes of interest and regions assumed to be important functionally, such as promoters and CpG islands[2,3], and there has not been a comprehensive genome-scale understanding of the relationship between DNA methylation loss and gain in cancer and in normal differentiation. In these experiments, we focused on three key questions. First, where are DNA methylation changes that distinguish tissue types? Taking a comprehensive genome-wide approach, we examined three normal tissue types representing the three embryonic lineages—liver (endodermal), spleen (mesodermal) and brain (ectodermal)—obtained from five autopsies. A difference from previous methylation studies of tissues, besides the genome-wide design, is that here tissues were obtained from the same individual, thus controlling for potential interindividual variability. Second, where are DNAm alterations in cancer, and what is the balance between hypomethylation and hypermethylation? For this purpose, we examined 13 colorectal cancers and matched normal mucosa from the subjects. Third, what is the functional role of these methylation changes? To this end, we carried out a comparative epigenomics study of tissue methylation in the mouse, as well as gene expression analyses.

To examine DNAm on a genome-wide scale, we carried out comprehensive high-throughput array-based relative methylation (CHARM) analysis, which is a microarray-based method agnostic to preconceptions about DNAm, including location relative to genes and CpG content[4]. The resulting quantitative measurements of DNAm, denoted with M, are log ratios of intensities from total (Cy3) and McrBC-fractionated DNA (Cy5): positive and negative M values are quantitatively associated with methylated and unmethylated sites, respectively. For each sample, we analyzed ~4.6 million CpG sites across the genome using a custom-designed NimbleGen HD2 microarray, including all of the classically defined CpG islands as well as all nonrepetitive lower CpG density genomic regions of the genome. We included 4,500 control probes to standardize these M values so that unmethylated regions were associated, on average, with values of 0. CHARM is 100% specific at 90% sensitivity for known methylation marks identified by other methods (for example, in promoters) and includes the approximately half of the genome not identified by conventional region preselection[4]. The CHARM results were also extensively corroborated by quantitative bisulfite pyrosequencing analysis.

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA. [2]Center for Epigenetics, Institute for Basic Biomedical Sciences and [3]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [4]Center for Statistical Sciences, Brown University, Providence, Rhode Island, USA. [5]Stanley Laboratory of Brain Research, Uniformed Services University of Health Sciences, Bethesda, Maryland 20892, USA. [6]Departments of Psychiatry and [7]Pediatrics, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to R.A.I. (rafa@jhu.edu) or A.P.F. (afeinberg@jhu.edu).

Received 27 October 2008; accepted 7 November 2008; published online 18 January 2009; doi:10.1038/ng.298

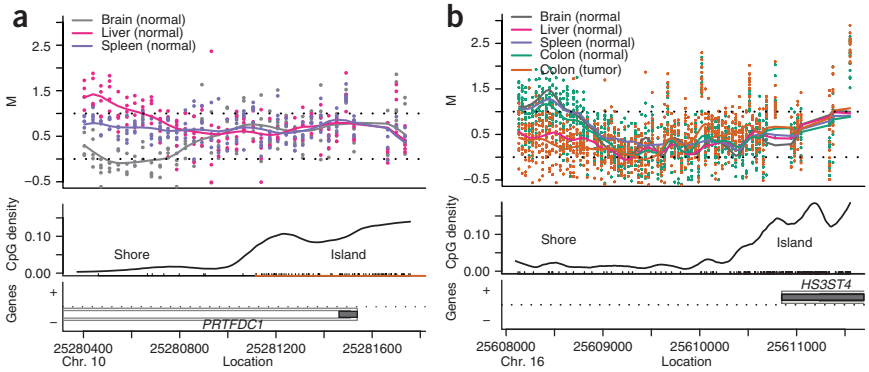

Figure 1 Most tissue-specific differential DNA methylation is located at CpG island shores. (**a**) An example of a T-DMR located at a CpG island shore in the *PRTFDC1* gene. The upper panel is a plot of M value versus genomic location for brain (gray), liver (pink) and spleen (purple). Each point represents the methylation level of an individual sample for a given probe. The curve represents averaged smoothed M values, described in detail in the Methods. Because of the scale and standardization used, M values that range from −0.5 to 0.5 represent unmethylated sites as defined by the control probes, and values from 0.5 to 1.5 represent baseline levels of methylation. The middle panel provides the location of CpG dinucleotides with black tick marks on the *x* axis. CpG density was calculated across the region using a standard density estimator and is represented by the smoothed black line. The location of the CpG island is denoted on the *x* axis as an orange line. The lower panel provides gene annotation for the genomic region. The thin outer gray line represents the transcript, the thin inner lines represent a coding region. Filled in gray boxes represent exons. On the *y* axis, plus and minus marks denote sense and antisense gene transcription, respectively. (**b**) An example of a C-DMR that is located in a CpG island shore and that overlaps a T-DMR. Liver (pink) is hypomethylated relative to brain (gray) and spleen (purple) tissues. Hypomethylation of colon tumor (orange) is observed in comparison to matched normal colon tissue (green) and overlaps the region of liver hypomethylation.

RESULTS

Most tissue-specific DNAm occurs in 'CpG island shores'

Because CHARM is not biased for CpG island or promoter sequences, we could obtain objective data on tissue-specific methylation. We identified 16,379 tissue differential methylation regions (T-DMRs), defined as regions with M values for one tissue consistently different than that for the others at a false discovery rate (FDR) of 5% (see Methods). The median size of a T-DMR was 255 bp. Previous studies of tissue- or cancer-specific DNAm have focused on promoters and/or CpG islands, which have been defined as regions with a GC fraction greater than 0.5 and an observed-to-expected ratio of CpG greater than 0.6 (refs. 2,5). It has previously been reported that the degree of differences in DNAm of promoters in somatic cells is relatively low in conventionally defined CpG islands and higher at promoters with intermediate CpG density[6,7]. Two recent studies identified a relatively small fraction, 4–8%, of CpG islands with tissue-specific methylation[8,9]. We also found that DNAm variation is uncommon in CpG islands (**Supplementary Fig. 1** online).

The genome-wide approach of CHARM also enabled us to find an unexpected physical relationship between CpG islands and DNAm variation, namely that 76% of T-DMRs were located within 2 kb of islands in regions we now denote as 'CpG island shores'. For example, for the T-DMR in the *PRTFDC1* gene, which encodes a brain-specific phosphoribosyltransferase that is relatively hypomethylated in the brain, the spreading of M values among the tissues begins ~200 bp from the CpG island and at a point where the CpG density associated with the island has fallen to 1/10 the density in the island itself (**Fig. 1**). The association of T-DMRs with CpG island shores is not due to an arbitrary definition of CpG islands but to a true association of these DNAm differences near but not in the regions of dense CpG content (**Supplementary Data 1** online describes all T-DMR regions, and plots similar to those in **Fig. 1** for the complete set of T-DMRs, ordered by statistical significance, are available online; see URLs section in Methods). The distribution of T-DMRs by distance from the respective islands shows that DNAm variation is distributed over a ~2 kb shore, and that although CpG islands are enriched on the arrays, because of their high CpG content (33% of CHARM probes are in islands), only 6% of T-DMRs are in islands, compared to 76% in shores; an additional 18% of T-DMRs were located greater than 2 kb from the respective islands (**Fig. 2**). The localization of T-DMRs also occurred largely outside of promoters (96%), as CpG islands are localized primarily within promoters[10]. Furthermore, more than half (52%) of T-DMRs were greater than 2 kb from the nearest annotated gene. The distribution of the distance to islands remained essentially unchanged when we used FDR cutoffs of 0.01, 0.05 and 0.10 (data not shown).

We confirmed the array-based result that the differential methylation was in CpG island shores rather than in the associated islands by carrying out bisulfite pyrosequencing analysis on over 100 CpG sites in the islands and shores associated with four genes, three T-DMRs and one cancer differential methylation region. At all 101 sites, the DMR was confirmed to lie within the shore rather than the island (**Supplementary Table 1** online). For example, *PCDH9*, which encodes a brain-specific protocadherin, was relatively hypomethylated in the brain at all 6 sites examined in the CpG island shore but unmethylated in both brain and spleen at all 18 sites examined in the associated island (**Supplementary Table 1**). Differential methylation of an additional four CpG island shores was also confirmed by bisulfite pyrosequencing of 39 total CpG sites, and all showed statistically significant differences in DNAm ($P < 0.05$) (**Supplementary Table 2** online). These data verify the sensitivity of CHARM for detecting subtle differences in DNAm. Furthermore, they confirm that most normal differential methylation takes place at CpG island shores.

Similar CpG island shore hypo- and hypermethylation in cancer

We used the same comprehensive genome-wide approach to address cancer-specific DNA methylation. We focused on colorectal cancer, a paradigm for cancer epigenetics because of the availability of subject-matched normal mucosa, the cell type from which the tumors arise. We analyzed DNAm on 13 colon cancers and matched normal mucosa from the same individuals, identifying 2,707 regions showing differential methylation in cancers (C-DMRs) with an FDR of 5% (**Supplementary Data 2** online, and plots similar to those in **Fig. 1** for the complete set of C-DMRs, ordered by statistical significance, are available online; see URLs section in Methods). These C-DMRs were similarly divided between those showing hypomethylation in the cancer (compared to the normal colon) and those showing hypermethylation (1,199 (44%) and 1,508 (56%), respectively). The CHARM arrays, like other tiling arrays, do not contain repetitive sequences, so the abundance of hypomethylation is not due to

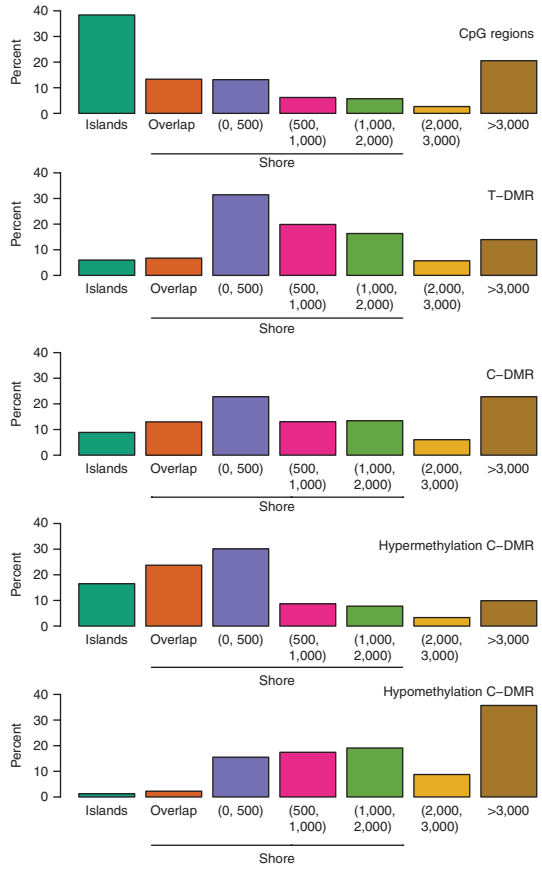

Figure 2 Distribution of distance of T-DMRs and C-DMRs from CpG islands. Islands (teal) are regions that cover or overlap more than 50% of a CpG island. Overlap (orange) are regions that overlap 0.1–50% of a CpG island. Regions denoted by (0, 500] (purple) do not overlap islands but are located $\leq$500 bp of islands. Regions denoted by (500, 1,000] (magenta) are located $>$500 and $\leq$1,000 bp from an island. Regions denoted by (1,000, 2,000] (green) are regions $>$1,000 bp and $\leq$2,000 bp from an island. Regions denoted by (2000, 3000] (yellow) are located $>$2,000 bp and $\leq$3,000 bp from an island. Regions denoted $>$3,000 (brown) are $>$3,000 bp from an island. The percentage of each class is provided for CpG regions (the CHARM arrays themselves, null hypothesis), tissue-specific differentially methylated regions (T-DMRs), cancer-specific differentially methylated regions (C-DMRs), and the latter subdivided into regions of cancer-specific hypermethylation and hypomethylation.

enrichment for repetitive DNA, which has been shown to be hypomethylated in cancer[11]. This similarity in amount of hypomethylation and hypermethylation is also shown in a quantile-quantile plot, in which quantiles for the observed average difference between tumor and normal sample Ms are plotted against quantiles from a null distribution constructed with the control (M = 0) regions (**Fig. 3a**).

Although both hypomethylation and hypermethylation in cancer involved CpG island shores, there were subtle differences in the precise regions that were altered. The hypermethylation extended to include portions of the associated CpG islands in 24% of cases (termed 'overlap' in **Fig. 2**), which could account for the island hypermethylation frequently reported in cancer, even though that is not the predominant site of modification. In contrast, the hypomethylation extended to between 2 and 3 kb from the associated island in 10% of cases and was not associated with an island in 35% of cases (**Fig. 2**).

To confirm differential methylation in colon tumors, we carried out additional bisulfite pyrosequencing validation of nine C-DMRs, including five regions showing hypermethylation and four regions with hypomethylation, in an average of 50 primary cancer and normal mucosal samples per gene. For all of the genes, the pyrosequencing data matched the CHARM data (P values ranging from 10^{-4} to 10^{-17}) (**Fig. 3b–j** and **Supplementary Table 3** online). Thus, CHARM was precise in identifying both T-DMRs and C-DMRs.

Our screening process was effective at identifying known targets of altered DNAm in cancer. For example, 10 of the 25 most statistically significant C-DMRs have previously been reported to show altered DNAm in cancer, for example, *WNK2*, hypermethylated in glioblastoma[12] and *HOXA6*, hypermethylated in lymphoid malignancies[13]. However, we also identified hundreds of genes not previously described. For example, for hypermethylation, we identified genes encoding GATA-2, an important regulator of hematopoetic differentiation[14], and RARRES2, whose expression is decreased in intestinal adenomas[15]. For hypomethylation, we identified genes encoding DPP6, a biomarker for melanoma[16], MRPL36, a DNA helicase that confers susceptibility to breast cancer[17], and MEST, a known target of hypomethylation and loss of imprinting in breast cancer[18]. Note that although previous T-DMR screens have focused on CpG islands, which we show account for only 8% of T-DMRs, our screen did identify CpG island loci validated by others as well, for example, *PAX6*, *OSR1* and *HOXC12*. Thus, cancer, like normal tissues, involves changes in DNAm in CpG island shores, with comparable amounts of hypomethylation and hypermethylation but with subtle differences in the precise distribution of these alterations with respect to the associated CpG island. These differences will have important functional implications for gene expression, as discussed later.

Gene expression is linked to non–CpG-island methylation

Because the identification of CpG island shores was unexpected, we explored the functional relationship between their differential methylation and the expression of associated genes. To address tissue- and cancer-specific DNAm, we analyzed gene expression across the genome in five primary brains and livers from the same autopsy specimens and in four colon cancers and subject-matched normal mucosa; all samples were from subjects for whom we had genome-wide methylation analysis data. Methylation of T-DMRs showed a strong inverse relationship with differential gene expression, even though these DMRs were not CpG islands but rather CpG island shores. The relationship between DNAm and gene expression was greater for DMRs in which one of the two measured points had approximately no methylation ('none-to-some' methylation compared to 'some-to-more' or 'some-to-less' methylation), particularly for hypomethylation (**Fig. 4**). The significant association of gene expression with T-DMRs was true even when the DMR was 300–2,000 bp from the transcription start site, for example, average log-ratio values of 0.84 and 0.35 ($P < 10^{-37}$ and 10^{-4}) for some-to-none and some-to-more/less methylation, respectively, comparing liver to brain expression (**Fig. 4**). Moreover, when we related T-DMRs to changes in gene expression from over 242 samples, representing 20 different tissue types, we found that 5,352 of the 8,910 genes that were differentially expressed across the 20 tissues were within 2 kb of a T-DMR, much more than expected by chance ($P < 10^{-15}$). For C-DMRs as well, even though there were fewer of them than T-DMRs, there was a significant association of gene expression with DNAm: $P < 10^{-6}$ and $P < 10^{-3}$ for hypermethylation and hypomethylation, respectively; again, the relation was much more marked when one of the two measured points had no methylation (**Supplementary Fig. 2** online).

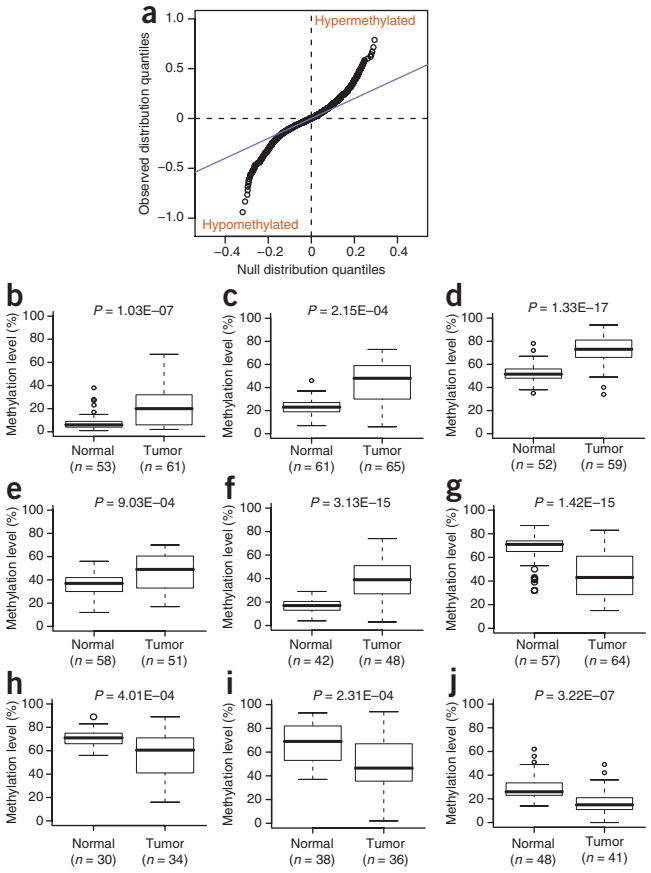

Figure 3 Similar numbers of sites of hypomethylation and hypermethylation in colon cancer. (**a**) A quantile-quantile plot shows a similar number of sites of hypomethylation and hypermethylation in colon cancer. The quantiles of the differences in M values between tumor and normal colon tissues are plotted against the quantiles of a null distribution formed using the differences seen in the control regions. Points deviating from the diagonal are not expected by chance, and a similar proportion is seen for hypomethylation and hypermethylation in cancer. (**b**–**j**) Bisulfite pyrosequencing confirms the prevalence of five hypermethylated and four hypomethylated C-DMR shores in a large set of colon tumor and normal mucosa samples. Box plots represent degree of DNA methylation as measured using bisulfite pyrosequencing. (**b**) *DLX5* (distal-less homeobox 5). (**c**) *LRFN5* (leucine-rich repeat and fibronectin type III domain containing 5). (**d**) *HOXA3* (homeobox A3). (**e**) *SLITRK1* (SLIT and NTRK like family, member 1). (**f**) *FEZF2* (FEZ family zinc finger 2). (**g**) *TMEM14A* (transmembrane protein 14A). (**h**) *ERICH1* (glutamate-rich 1). (**i**) *FAM70B* (family with sequence similarity 70, member B). (**j**) *PMEPA1* (prostate transmembrane protein, androgen induced 1). *n*, number of samples analyzed by pyrosequencing.

We validated the inverse relationship between DNAm and transcription at eight CpG island shores, two T-DMRs and six C-DMRs in tissues and colon cancers, respectively, using quantitative real-time PCR. Both of the T-DMRs were in shores, one located 844 bp upstream of the promoter and one within the gene body. Similarly, all six of the C-DMRs assayed were in shores, with five located in the gene promoter and one within the gene body (**Supplementary Table 4** online). These quantitative data provide additional support for a strong relationship between differential methylation in CpG island shores and transcription of associated genes. This functional relationship between gene expression and shore methylation applies to shores located within 2 kb of an annotated transcriptional start site but leaves open the possibility of additional regulatory function for shores located in intragenic regions or gene deserts.

Shore-linked silencing reversed by methyltransferase inhibition

The previous data, although compelling, are associative in nature. For a more functional analysis, we therefore compared DNA methylation and gene expression data from tissues studied in the current work to a rigorous analysis using hundreds of expression microarray experiments published earlier[19], which tested the effects on gene expression of 5-aza-2′-deoxycytidine (AZA), and also to double DNA methyltransferase 1 and 3B somatic cell knockout (DKO) experiments. We compared genes from the present study that had DMRs meeting an FDR < 0.05 and that showed differential expression in the tissues at $P < 0.05$ to genes that had significant P values after AZA or DKO. Of 27 DMRs that showed relative hypermethylation with gene silencing in tissues, 23 were activated by AZA (**Fig. 5a** and **Supplementary Data 3** online). Similarly, of 25 DMRs that showed relative hypermethylation with gene silencing in tissues, all 25 were activated by DKO (**Fig. 5b** and **Supplementary Data 3**). Thus, both chemical and genetic demethylation cause changes in gene expression similar to those associated with increased methylation of CpG island shores.

DMRs are associated with alternative transcription

What might be the function of differential methylation at CpG island shores? One possibility is alternative transcription. Both the T-DMRs and C-DMRs often involved alternative transcripts, as defined by cap analysis gene expression (CAGE)[20,21]: 68% and 70% of the T-DMRs

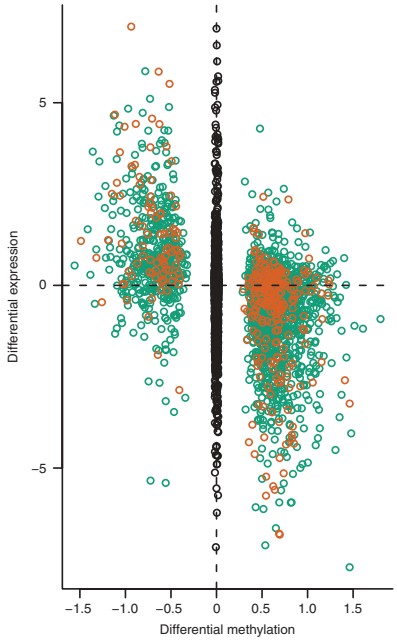

Figure 4 Gene expression is strongly correlated with T-DMRs at CpG island shores. For each brain versus liver T-DMR, we found the closest annotated gene on the Affymetrix HGU133A microarray, resulting in a total of 2,041 gene–T-DMR pairs. Plotted are log (base 2) ratios of liver to brain expression against ΔM values for liver and brain DNAm. Orange dots represent T-DMRs located within 300 bp from the corresponding gene's transcriptional start site (TSS). Green dots represent T-DMRs that are located from 300 to 2,000 bp from the TSS of an annotated gene. Black dots, in the middle, represent log ratios for all genes further than 2 kb from an annotated TSS.

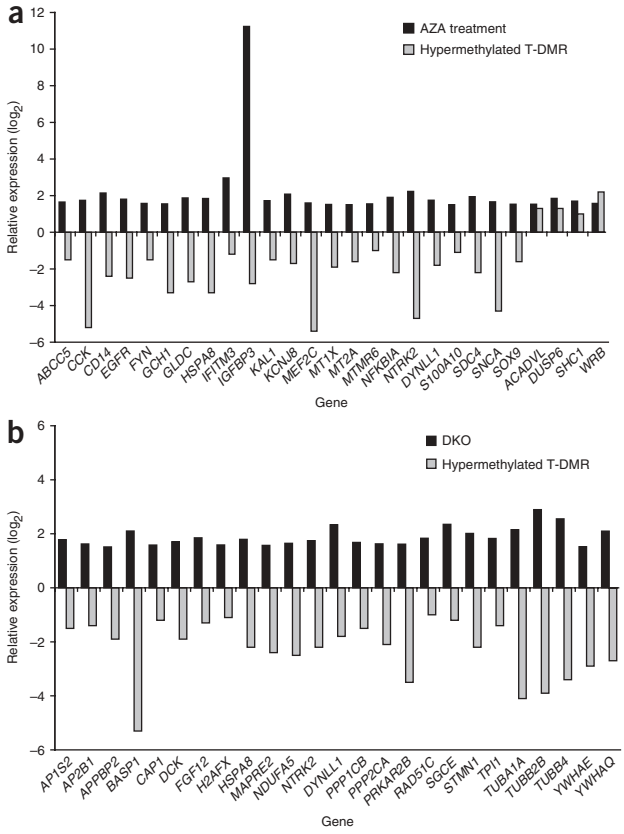

Figure 5 Genes downregulated in association with T-DMR shore hypermethylation are activated by 5-aza-2′-deoxycytidine treatment of colon cancer cell line HCT116 and knockout of DNA methyltrasferase 1 and 3b in HCT116. (**a**) Genes significantly upregulated ($P < 0.05$) after treatment of HCT116 cells with 5-aza-2′-deoxycytidine (AZA) (black) that are also associated with a relatively hypermethylated T-DMR showing a significant change in gene expression ($P < 0.05$) (gray): 23/27 genes are activated by AZA. (**b**) Genes significantly upregulated ($P < 0.05$) after knockout of DNA methyltransferases 1 and 3b (DKO) in HCT116 cells (black) that are also associated with a relatively hypermethylated T-DMRs showing a significant change in gene expression ($P < 0.05$) (gray): 25/25 genes are activated by DKO. Plotted are log (base 2) ratios of expression of AZA/untreated, DKO/HCT116 and relatively hypermethylated/hypomethylated tissue.

and C-DMRs, respectively, were not within 500 bp of an annotated transcriptional start site but were within 500 bp of an alternative transcriptional start site. By chance, we expect only 58% to have this relationship ($P < 10^{-15}$). These results suggest that DNA methylation might regulate alternative transcription in normal differentiation and cancer. We therefore carried out rapid amplification of cDNA ends (RACE) experiments in order to confirm the presence of alternative transcripts and their differential expression in cancer. We examined three colon tumor and subject-matched normal mucosa at the *PIP5K1A* locus, a C-DMR that is hypomethylated in colon tumors, and confirmed that an alternative RNA transcript is produced in colon tumors compared to their matched normal counterparts (**Supplementary Fig. 3** online). Thus, a key function for differential methylation during differentiation may be alternative transcription, and the role of altered DNAm in cancer may in part be disruption of the regulatory control of specific promoter usage.

Mouse DNAm discriminates human tissues, even far from genes

A compelling argument for the functional importance of differential DNAm of CpG island shores would be their conservation across species. One might expect DMRs near transcriptional start sites to be conserved because the genes are conserved. However, when we examined the relationship between gene-distant T-DMRs (2–10 kb away from an annotated gene) and sequence conservation using the phastCons28way table from the University of California Santa Cruz genome browser, we found that 48% of differentially methylated regions showed sequence conservation. Furthermore, 91% of DMRs were located within 1 kb of a highly conserved region ($P < 0.001$).

To address whether the DNA methylation itself is conserved across species, we created a mouse CHARM array with ~2.1 million features independently of the human array. We then isolated tissue replicates from each of three mice, corresponding to the tissues examined in the human T-DMR experiments, and then mapped these methylation data across species using the UCSC LiftOver tool. The interspecies correspondence of tissue-specific methylation was notable, and unsupervised clustering perfectly discriminated among the tissues, regardless of the species of origin (**Fig. 6**; $P < 10^{-9}$). Perfect discrimination among the tissues was found even when we limited the analysis to gene-distant DMRs (**Supplementary Fig. 4** online). Thus, DNAm itself is highly conserved across 50 Myr of evolution (approximately 51% of mapped DNAm sites were conserved). We also noticed relatively little heterogeneity in tissue-specific methylation in the mouse compared to the human (height of the cluster bars in **Fig. 6**), suggesting a genetic-epigenetic relationship, as the mice are inbred.

The location of C-DMRs overlaps that of T-DMRs

Because both C-DMRs and T-DMRs were located at CpG island shores, we then asked whether they occurred in similar locations. We

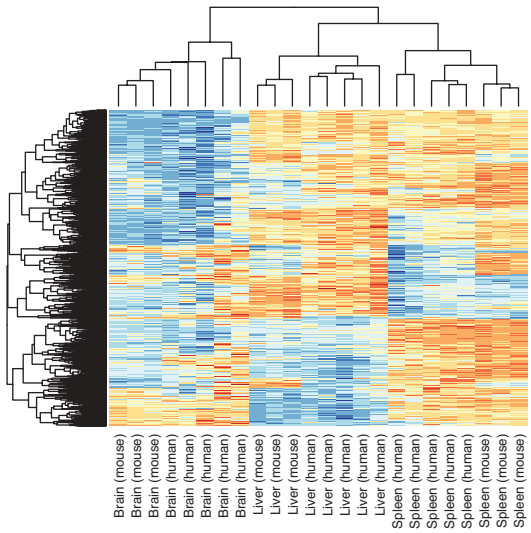

Figure 6 Clustering of human tissue samples using mouse T-DMRs results in perfect discrimination of tissues. The M values of all tissues from the 1,963 regions corresponding to mouse T-DMRs that mapped to the human genome were used for unsupervised hierarchical clustering. By definition, the mouse tissues are segregated. Notably, all of the human tissues are also completely discriminated by the regions that differ in mouse tissues. The three major branches in the dendrograms correspond perfectly to tissue type regardless of species. Columns represent individual samples, and rows represent regions corresponding to mouse T-DMRs. The heat map shows M values, with red being more methylated and blue less.

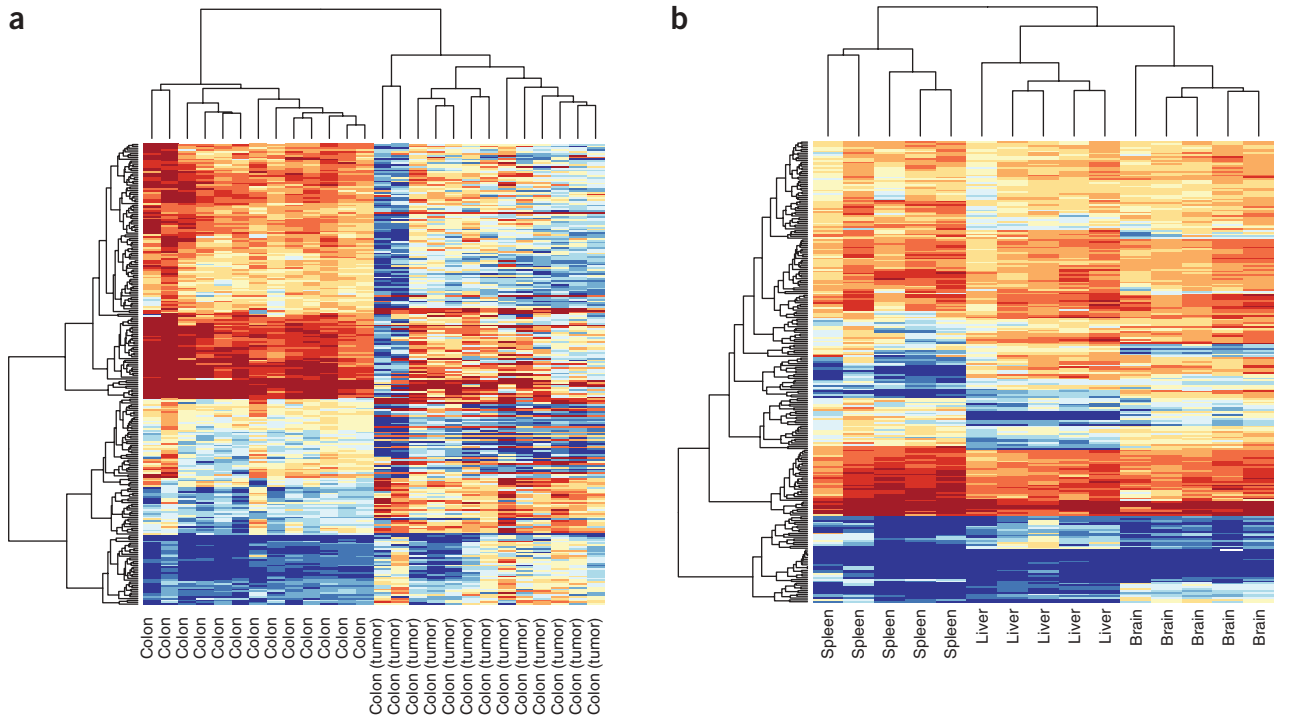

Figure 7 Clustering of normal tissue samples using C-DMRs results in perfect discrimination of tissues. The M values of all tissues from the 2,707 regions corresponding to C-DMRs were used for unsupervised hierarchical clustering. (**a**) By definition, the colon tumors and matched normal mucosa are segregated. The two major branches in the dendrograms correspond perfectly to tissue type. (**b**) Notably, all of the normal brains, spleens and livers are also completely discriminated by the regions that differ in colon cancer. The three major branches in the dendrograms correspond perfectly to tissue type. Columns represent individual samples, and rows represent regions corresponding to C-DMRs. The heat map shows M values, with red being more methylated and blue less.

focused on DMRs in which the methylation difference was from no methylation to some methylation, that is, those DMRs for which the gene expression data above showed a strong relationship between 'none-to-some' methylation and gene silencing. Notably, we found that 52% of the C-DMRs overlapped a T-DMR, compared to only 22% expected by chance ($P < 10^{-14}$), when using an FDR of 5% for defining T-DMRs. Although these data are significant, the definition of a T-DMR based on FDR of 5% is conservative. We therefore also asked directly whether C-DMRs are enriched for tissue variation in DNAm by computing an averaged F-statistic (comparison of cross-tissue to within-tissue variation) at each C-DMR. The cross-tissue variation in normal tissues was significant at 64% of the C-DMRs, compared to 20% of randomly selected CpG regions on the array matched for size ($P < 10^{-143}$). When we defined DMRs using an FDR of 5%, 1,229 of 2,707 C-DMRs overlapped a T-DMR, of which 265, 448 and 185 are brain-, liver- and spleen-specific, respectively, and 331 show variation among all of the tissues (**Supplementary Data 4** online). The colon C-DMRs were highly enriched for overlap with liver T-DMRs ($P < 10^{-15}$), and liver is embryologically closest to colon of the autopsy tissues studied. For example, the C-DMR located in the CpG island shore upstream of the *HS3ST4* (heparan sulfate D-glucosaminyl 3-O-sulfotransferase 4) gene is hypomethylated in colon cancer compared to normal colon and coincides with a T-DMR that distinguishes liver from other tissues (**Fig. 1b**). The correspondence between C-DMRs and T-DMRs was so marked that when we carried out unsupervised clustering of the normal brain, liver and spleen, using the M values from the C-DMRs, there was perfect discrimination of the tissues (**Fig. 7**).

Most tissue-specific methylation difference more commonly involves hypomethylation, although this varies by tissue type with 50% of liver, 62% of spleen, and 79% of brain DMRs representing hypomethylation, and cancer-specific methylation differences slightly more frequently involve hypermethylation (56%:44%). For both T-DMRs and C-DMRs, when there was differential methylation, it was common that at least one of the tissues was completely unmethylated (68% and 37%, respectively). Furthermore, hypomethylated C-DMRs were twice as likely to resemble another tissue type, such as liver, than were hypermethylated C-DMRs (82% versus 61%, $P < 10^{-31}$), even though hypermethylated C-DMRs overlapped T-DMRs 1.5-fold more frequently than did hypomethylated C-DMRs (54% versus 35%, $P < 10^{-21}$).

To further explore the relationship between differentiation and type of methylation change, we carried out Gene Ontology (GO) analysis for both hypomethylated and hypermethylated C-DMRs in the cancers (see Methods). The GO analysis showed enrichment for development and pluripotency-associated genes for both hyper- and hypomethylated C-DMRs ($P < 0.01$) (**Supplementary Table 5** online). Hypomethylated C-DMRs were also enriched for genes associated with differentiated cellular functions for lineages other than the colon ($P < 0.01$) (**Supplementary Table 5**). Thus, cancer-specific DNA methylation predominantly involves the same sites that show normal DNAm variation among tissues, particularly at genes associated with development.

Next, we examined the magnitude of differential methylation and variation in C-DMRs and T-DMRs. The ΔM values for tissue and cancer DMRs differed markedly from nonmethylated controls or randomly selected regions (the latter have an average value

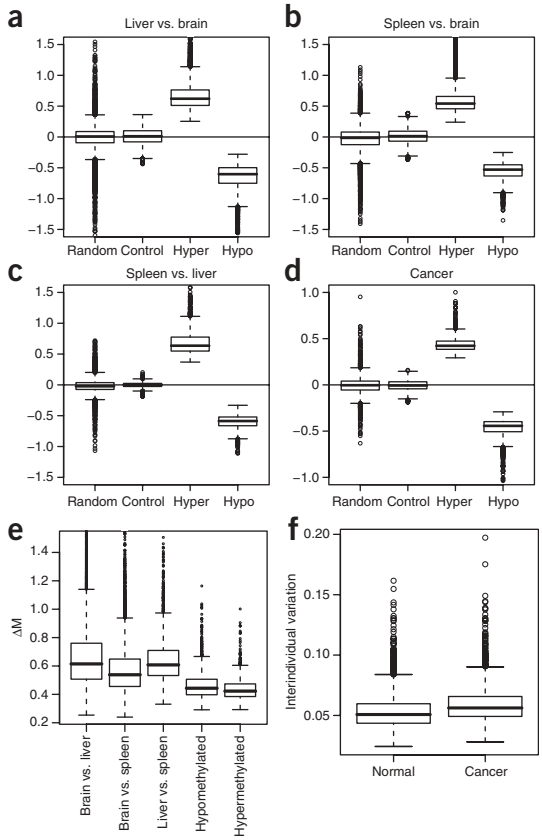

Figure 8 Magnitude of differential methylation and variation in C-DMRs and T-DMRs. (**a**–**d**) Box plots of average ΔM values over all DMRs, compared to randomly chosen regions and unmethylated control regions, matched for length. (**a**) Liver versus brain. (**b**) Spleen versus brain. (**c**) Spleen versus liver. (**d**) Colon cancer versus normal colonic mucosa. (**e**) Differences in DNA methylation are greater in magnitude among normal tissues than are differences between colon tumors and matched normal mucosa. For all DMRs we computed the average ΔM. We then stratified these values into T-DMRs, hypermethylated C-DMRs and hypomethylated C-DMRs. T-DMRs were further stratified according to brain versus liver, brain versus spleen and liver versus spleen pairwise comparisons. The box plots represent absolute values of the ΔMs. (**f**) Interindividual variation in M is larger among colon tumors than matched normal mucosa. For each C-DMR we computed the average interindividual s.d. of the M values. The box plots represent these values for normal colon mucosa and colon tumors.

comparable to controls but with significant tails, as by definition they may contain DMRs themselves) (**Fig. 8a–d**). The ΔM values for normal tissues were comparable across the tissues, but the ΔM values between normal and cancer tissues were on average approximately half the ΔM between normal tissue pairs (**Fig. 8e**), which is logical given that the cancers are compared with their tissue of origin. Another difference between cancer and normal tissues was an increase in the interindividual variation in M among the colon cancers, which was on average ~50% greater than the interindividual variation among the normal colons (**Fig. 8f**), a result which may help to explain tumor cell heterogeneity. Given the strong interindividual variability we found in cancer, we identified 205/2,707 C-DMRs that were consistently differentially methylated between the colon tumor and matched normal mucosa from all 13 individuals examined (**Supplementary Data 4**). These regions, heavily over-represented for development and morphogenesis genes, provide a smaller, more focused set of regions for biomarker discovery and carcinogenesis studies.

DISCUSSION

We have carried out a genome-wide analysis of DNA methylation addressing variation among normal tissue types, variation between cancer and normal, and variation between human and mouse, revealing several surprising relationships among these three types of epigenetic variation, supported by extensive bisulfite pyrosequencing and functional analysis. First, most tissue-specific DNAm occurs not at CpG islands but at CpG island shores. The identification of these regions opens the door to functional studies, such as those investigating the mechanism of targeting DNAm to these regions and the role of differential methylation of shores. Supporting a functional role for shores, gene expression was closely linked to T-DMR and C-DMR methylation, particularly for switches from 'none' to 'some' methylation. The relationship between shore methylation and gene expression was confirmed by 5-aza-2′-deoxycytidine and DNA methyltransferase knockout experiments altering expression of the same genes. Another mechanism for shores supported by this study is regulation of alternative transcripts, supported by mapping and RACE experiments.

Although 76% of T-DMRs were in CpG island shores, at least for the three tissues examined here, 24% were not adjacent to conventionally defined CpG islands. However, many of these regions were nevertheless shores of CpG-enriched sequences (for an example, see **Supplementary Fig. 5** online). We are currently developing a novel algorithm for CpG island definition based on hidden Markov modeling that will likely increase the fraction of T-DMRs in CpG island shores. The 'CpG clusters' recently identified[22] are not CpG island shores (only 4% of shore DMRs map to them), although the shores of these clusters, like the shores of CpG islands, are enriched for DMRs. Note that the variation in DNAm is still not within the dense CpG regions as defined by any of these definitions but in CpG shores.

The second key finding of the study is that T-DMRs are highly conserved between human and mouse, and the methylation itself is sufficiently conserved to completely discriminate tissue types regardless of species of origin. This was true even for T-DMRs located >2 kb from transcriptional start sites. The incorporation of epigenetic data, such as DNAm, in evolutionary studies as done here, should greatly enhance the identification of conserved elements that regulate differentiation. We also found greater DNAm heterogeneity in human than in mouse (at least in an inbred strain), even for DMRs located >2 kb from a gene promoter. This result suggests that the conservation of DNAm between human and mouse may have a strong genetic basis, consistent with a greater degree of tissue DNAm homogeneity in the inbred mouse strain.

The third key finding of the study is that most cancer-related changes in DNAm, that is, C-DMRs, at least for colon cancer, correspond to T-DMRs, and that these changes are similarly divided between hypomethylation and hypermethylation and also involve CpG island shores. Thus, epigenetic changes in cancer largely involve the same DMRs as epigenetic changes in normal differentiation. These results have important implications for studies such as the Cancer Genome Atlas, in that most altered DNA methylation in cancer does not involve CpG islands, and thus these studies would benefit from analysis of CpG island shores. Similarly, high-throughput sequencing efforts based on reduced representation analysis of CpG islands *per se* are unlikely to identify most DNAm variation in normal tissues or in cancer.

Finally, GO annotation analysis suggests that DNAm changes in cancer reflect development and pluripotency-associated genes, and differentiated cellular functions for lineages other than the colon. These data are consistent with the epigenetic progenitor model of cancer[23], which proposes that epigenetic alterations affecting tissue-specific differentiation are the predominant mechanism by which epigenetic changes cause cancer. The genes identified in this analysis will themselves be of considerable interest for further study, as will be the potential regulatory regions that did not lie in close proximity to annotated genes.

METHODS

Samples. We obtained snap-frozen colon tumors and dissected normal mucosa from the same subjects courtesy of B. Vogelstein (Johns Hopkins University School of Medicine). Human postmortem brain, liver and spleen tissues, from the same individual, were donated by The Stanley Medical Research Institute brain collection.

Genomic DNA isolation and McrBC fractionation. We carried out genomic DNA isolation using the MasterPure DNA purification kit (Epicentre) as recommended by the manufacturer. For each sample, 5 μg of genomic DNA was digested, fractionated, labeled and hybridized to a CHARM microarray.

CHARM microarray design. CHARM microarrays were prepared as previously described[4] and additionally included a set of 4,500 probes, totaling 148,500 base pairs across 30 genomic regions as controls. As these control probes represent genomic regions without CpG sites and hence can not be methylated, we used them to normalize and standardize array data. We standardized the observed M values so that the average in the control regions was 0. Therefore, M values of 0 for other probes on the array are associated with no methylation.

CHARM DNA methylation analysis. We conducted McrBC fractionation followed by CHARM array hybridization for all human tissue samples as previously described[4]. For each probe, we computed average M values across the five samples in each tissue type. Differential methylation was quantified for each pairwise tissue comparison by the difference of averaged M values (ΔM). Replicates were used to estimate probe-specific s.d., which provided standard errors (s.e.m.) for ΔM. We calculated z scores (ΔM/s.e.m.(:ΔM)) and grouped contiguous statistically significant values into regions. Because millions of z scores are examined, statistical confidence calculation needed to account for multiple comparisons. We therefore computed false discovery rates (FDR) and reported a list with an FDR of 5%[24,25]. Statistical significance of the regions was assessed as described below. C-DMRs were determined using the same procedure described above with the following exception: because we observed greater heterogeneity in the cancer samples (**Fig. 8f**), we did not divide ΔM by the standard errors, as this would penalize regions of highly variable M values. For all expression microarray analysis, we used RMA for processing[26] and then averaged the samples in each tissue, and computed the difference (equivalent to average log ratio). Mouse T-DMRs were determined using the same statistical procedures as described above for the T-DMRs and were then mapped to the human genome using the UCSC liftOver tool. To correct for possible 'array' effects, we standardized each T-DMR by subtracting the mean of M across all samples within a species and divided by s.d. across all samples within a species. A list of all mouse T-DMRs is provided in **Supplementary Data 5** online. Overlap of C-DMRs with T-DMRs was determined by adding the number of regions.

Statistical significance of DMRs. Contiguous regions composed of probes with z scores associated with P values smaller than 0.001 were grouped into regions. We used the area of each region (length multiplied by ΔM) to define statistical significance[27]. We used a permutation test to form a null distribution for these areas and the empirical Bayes approach described[28]. We tested for effects of fragment length on M values observed using CHARM by computing the expected DNA fragment size based on McrBC recognition sites. Next, we stratified the ΔM values for each probe, from the colon tumor and normal mucosa comparison, by fragment size. The results show no relationship between fragment size and ΔM (data not shown).

Bisulfite pyrosequencing. Isolation of genomic DNA for all bisulfite pyrosequencing validation was done using the MasterPure DNA purification kit (Epicentre) as recommended by the manufacturer. For validation of shore regions, 1 μg of genomic DNA from each sample was bisulfite-treated using an EpiTect kit (Qiagen) according to the manufacturer's specifications. We PCR-amplified converted genomic DNA using unbiased nested primers and carried out quantitative pyrosequencing using a PSQ HS96 (Biotage) to determine percentage methylation at each CpG site. For bisulfite pyrosequencing of C-DMRs in 34–65 colon tumor and 30–61 normal mucosa samples, 500 ng of genomic DNA was bisulfite-treated using the EZ-96 DNA Methylation Gold kit (Zymo Research) as specified by the manufacturer. Converted genomic DNA was PCR-amplified using unbiased nested primers followed by pyrosequencing using a PSQ HS96 (Biotage). Bisulfite pyrosequencing was done as previously described[29]. We determined percentage methylation at each CpG site using the Q-CpG methylation software (Biotage). **Supplementary Table 6** online provides the genomic location of CpG sites measured in the CpG shore and associated CpG island bisulfite pyrosequencing assays. Genomic coordinates for all CpG sites measured in the set of ~50 colon tumor and normal samples are provided in **Supplementary Methods** online. Primer sequences and annealing temperatures for all bisulfite pyrosequencing reactions are provided in **Supplementary Table 7** online.

Total RNA isolation. We isolated total RNA for Affymetrix microarray analysis from all human tissues using the RNeasy Mini kit (Qiagen) as specified by the manufacturer. All samples were DNase treated using the on-column DNase digestion kit (Qiagen) as recommended. We measured total RNA concentration and determined RNA quality by using an RNA 6000 Nano Lab chip kit and running the chip on a 2100 Bioanalyzer (Agilent).

Affymetrix microarray expression analysis. Genome-wide transcriptional analysis was done on a total of five liver and five brain samples from the same individuals using Affymetrix U133A GeneChip microarrays. We obtained the raw microarray gene expression data for the brain and liver tissue from The Stanley Medical Research Institute (SMRI) online genomics database (see URLs section below)[30]. The five individuals selected were unaffected controls from the SMRI Array collection. We also carried out genome-wide transcriptional profiling on four colon tumor and four matched normal mucosa using Affymetrix U133 Plus 2.0 microarrays. One microgram of high-quality total RNA was amplified, labeled and hybridized according to the manufacturer's (Affymetrix) specifications and data was normalized as previously described[26,31].

Quantitative real-time PCR. We prepared total RNA for quantitative real-time PCR using the Trizol method (Invitrogen) for *FZD3*, *RBM38*, *NDN* and *SEMA3C* and we used the RNeasy mini kit (Qiagen) to isolate total RNA for *ZNF804A*, *CHRM2* and *NQO1*. We prepared cDNA for quantitative real-time PCR using the QuantiTect RT kit (QIAGEN). We used TaqMan assays (Applied Biosystems) to determine relative gene expression and analyzed experiments on a 7900HT detection system. Taqman assay identification numbers are provided in the **Supplementary Methods**. We used human *ACTB* as an endogenous control. Relative expression differences were calculated using the ΔΔCt method[32].

5′ RACE PCR. 5′ RACE experiments were done using a second generation RACE kit (Roche Applied Science) as specified by the manufacturer's protocol. RACE PCR products were directly sequenced with 3100 Genetic Analyzer (AB Applied Biosystems). Gene specific primer sequences are provided in **Supplementary Methods**.

GO annotation. We analyzed GO annotation using the Bioconductor[33] Gostats package[34] to find enriched categories ($P < 0.01$).

URLs. Complete set of T-DMR plots, http://rafalab.jhsph.edu/t-dmr3000.pdf; complete set of C-DMR plots, http://rafalab.jhsph.edu/c-dmr-all.pdf. The Stanley Medical Research Institute (SMRI) online genomics database, www.stanleygenomics.org.

Accession codes. NCBI GEO: Gene expression microarray data has been submitted under accession number GSE13471.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank B. Volgelstein (Johns Hopkins University School of Medicine) for providing colon tumors and matched normal mucosa samples. Postmortem brain, liver and spleen tissue was donated by The Stanley Medical Research Institute collection courtesy of M.B. Knable, E.F. Torrey and R.H. Yolken, whom we also thank for making available gene expression data for the brain and the liver tissue. We thank B. Carvalho for help with statistical software and C. Crainiceanu for advice with statistical methods. This work was supported by US National Institutes of Health grants P50HG003233 (A.P.F.), R37CA54358 (A.P.F.) and 5R01RR021967 (R.A.I.).

AUTHOR CONTRIBUTIONS

R.A.I. and A.P.F. designed the study and interpreted the results; R.A.I. designed new CHARM arrays and statistical methods with Z.W.; C.L.-A. performed bisulfite pyrosequencing, real-time quantitative PCR and sample preparation with C.M., K.G., M.R. and H.J.; B.W. and S.S. performed CHARM assays with sample preparation from M.W. and advice from J.B.P.; P.O. and H.C. performed functional assays; A.P.F. supervised the laboratory experiments and wrote the paper with R.A.I. and C.L.-A.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Feinberg, A.P. & Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301**, 89–92 (1983).
2. Feinberg, A.P. & Tycko, B. The history of cancer epigenetics. *Nat. Rev. Cancer* **4**, 143–153 (2004).
3. Baylin, S.B. & Ohm, J.E. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? *Nat. Rev. Cancer* **6**, 107–116 (2006).
4. Irizarry, R.A. *et al.* Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.* **18**, 771–779 (2008).
5. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
6. Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* **38**, 1378–1385 (2006).
7. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* **39**, 457–466 (2007).
8. Illingworth, R. *et al.* A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.* **6**, e22 (2008).
9. Shen, L. *et al.* Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genet.* **3**, 2023–2036 (2007).
10. Antequera, F. & Bird, A. CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr. Biol.* **9**, R661–R667 (1999).
11. Ehrlich, M. DNA methylation in cancer: too much, but also too little. *Oncogene* **21**, 5400–5413 (2002).
12. Hong, C. *et al.* Epigenome scans and cancer genome sequencing converge on WNK2, a kinase-independent suppressor of cell growth. *Proc. Natl. Acad. Sci. USA* **104**, 10974–10979 (2007).
13. Strathdee, G. *et al.* Inactivation of HOXA genes by hypermethylation in myeloid and lymphoid malignancy is frequent and associated with poor prognosis. *Clin. Cancer Res.* **13**, 5048–5055 (2007).
14. Cantor, A.B. *et al.* Antagonism of FOG-1 and GATA factors in fate choice for the mast cell lineage. *J. Exp. Med.* **205**, 611–624 (2008).
15. Segditsas, S. *et al.* Putative direct and indirect Wnt targets identified through consistent gene expression changes in APC-mutant intestinal adenomas from humans and mice. *Hum. Mol. Genet.* **17**, 3864–3875 (2008).
16. Jaeger, J. *et al.* Gene expression signatures for tumor progression, tumor subtype, and tumor thickness in laser-microdissected melanoma tissues. *Clin. Cancer Res.* **13**, 806–815 (2007).
17. Seal, S. *et al.* Truncating mutations in the Fanconi anemia J gene *BRIP1* are low-penetrance breast cancer susceptibility alleles. *Nat. Genet.* **38**, 1239–1241 (2006).
18. Pedersen, I.S. *et al.* Frequent loss of imprinting of PEG1/MEST in invasive breast cancer. *Cancer Res.* **59**, 5449–5451 (1999).
19. Gius, D. *et al.* Distinct effects on gene expression of chemical and genetic manipulation of the cancer epigenome revealed by a multimodality approach. *Cancer Cell* **6**, 361–371 (2004).
20. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA* **100**, 15776–15781 (2003).
21. Wakaguri, H., Yamashita, R., Suzuki, Y., Sugano, S. & Nakai, K. DBTSS: database of transcription start sites, progress report 2008. *Nucleic Acids Res.* **36**, D97–D101 (2008).
22. Glass, J.L. *et al.* CG dinucleotide clustering is a species-specific property of the genome. *Nucleic Acids Res.* **35**, 6798–6807 (2007).
23. Feinberg, A.P., Ohlsson, R. & Henikoff, S. The epigenetic progenitor origin of human cancer. *Nat. Rev. Genet.* **7**, 21–33 (2006).
24. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Statist Soc B (Methodol.)* **57**, 289–300 (1995).
25. Storey, J.D. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.* **31**, 2013–2035 (2003).
26. Irizarry, R.A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
27. Bullmore, E.T. *et al.* Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans. Med. Imaging* **18**, 32–42 (1999).
28. Efron, B., Tibshirani, R., Storey, J.D. & Tusher, V. Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **96**, 1151–1160 (2001).
29. Tost, J. & Gut, I.G. DNA methylation analysis by pyrosequencing. *Nat. Protocols* **2**, 2265–2275 (2007).
30. Higgs, B.W. *et al.* An online database for brain disease research. *BMC Genomics* **7**, 70 (2006).
31. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
32. Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)). *Methods* **25**, 402–408 (2001).
33. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
34. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2007).