Dynamic DNA methylation across diverse human cell lines and tissues

Katherine E. Varley,[1] Jason Gertz,[1] Kevin M. Bowling,[1] Stephanie L. Parker,[1,6] Timothy E. Reddy,[1,7] Florencia Pauli-Behn,[1] Marie K. Cross,[1] Brian A. Williams,[2] John A. Stamatoyannopoulos,[3,4] Gregory E. Crawford,[5] Devin M. Absher,[1] Barbara J. Wold,[2] and Richard M. Myers[1,8]

[1]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA; [2]Biology Division, and Beckman Institute, California Institute of Technology, Pasadena, California 91125, USA; [3]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA; [4]Department of Medicine, University of Washington, Seattle, Washington 98195, USA; [5]Department of Pediatrics, and the Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina 27708, USA

As studies of DNA methylation increase in scope, it has become evident that methylation has a complex relationship with gene expression, plays an important role in defining cell types, and is disrupted in many diseases. We describe large-scale single-base resolution DNA methylation profiling on a diverse collection of 82 human cell lines and tissues using reduced representation bisulfite sequencing (RRBS). Analysis integrating RNA-seq and ChIP-seq data illuminates the functional role of this dynamic mark. Loci that are hypermethylated across cancer types are enriched for sites bound by NANOG in embryonic stem cells, which supports and expands the model of a stem/progenitor cell signature in cancer. CpGs that are hypomethylated across cancer types are concentrated in megabase-scale domains that occur near the telomeres and centromeres of chromosomes, are depleted of genes, and are enriched for cancer-specific EZH2 binding and H3K27me3 (repressive chromatin). In noncancer samples, there are cell-type specific methylation signatures preserved in primary cell lines and tissues as well as methylation differences induced by cell culture. The relationship between methylation and expression is context-dependent, and we find that CpG-rich enhancers bound by EP300 in the bodies of expressed genes are unmethylated despite the dense gene-body methylation surrounding them. Non-CpG cytosine methylation occurs in human somatic tissue, is particularly prevalent in brain tissue, and is reproducible across many individuals. This study provides an atlas of DNA methylation across diverse and well-characterized samples and enables new discoveries about DNA methylation and its role in gene regulation and disease.

[Supplemental material is available for this article.]

In the early 1980s, several groups observed that the covalent addition of a methyl group to the cytosine base in mammalian genomic DNA at certain loci is associated with differential gene expression of nearby genes (Razin and Riggs 1980; Sutter and Doerfler 1980; van der Ploeg and Flavell 1980). This led to decades of research deciphering the patterns and purpose of this fifth DNA base in the human genome. It is an ongoing challenge to map the locations of methylated cytosines across the genome and to understand their roles in cell-type specific gene regulation (Ghosh et al. 2010), the establishment of gene expression patterns for stable differentiation (Lei et al. 1996; Okano et al. 1999; Jackson et al. 2004; Blelloch et al. 2006; Ji et al. 2010; Kim et al. 2010), and disease, including cancer where vast changes in DNA methylation patterns occur (Laird and Jaenisch 1994; Ushijima 2005; Sharma et al. 2010; Tsai and Baylin 2011). Bisulfite sequencing provides the most direct and highest resolution method to quantify DNA methylation in the genome, enabling the ability to calculate the fraction of molecules that are methylated at each individual cytosine sequenced (Frommer et al. 1992). The advent of next-generation DNA sequencing technologies has prompted the development of methods that take advantage of this vastly increased throughput, using bisulfite sequencing to query large subsets of the human genome (Meissner et al. 2008) and even whole human genomes (Cokus et al. 2008; Lister et al. 2009; Laurent et al. 2010; Li et al. 2010; Hansen et al. 2011; Berman et al. 2012; Hon et al. 2012).

The goal of this study was to generate a high-quality compendium of DNA methylation data across a large number of human cell lines and tissues. Reduced representation bisulfite sequencing (RRBS) was chosen because it provides quantitative, single-base resolution methylation profiles for a large subset of the human genome that is enriched for genic regions and CpG islands (CGIs) (Meissner et al. 2008). Other genomic assays have been performed in these samples as part of The ENCODE Project (The ENCODE Project Consortium 2007, 2011, 2012), providing a rich resource for integrated analysis of DNA methylation, gene expression, transcription factor binding, and chromatin modifications. We demonstrate that comparisons of methylation profiles across the diverse collection of samples in this study can be used to investigate cancer-associated methylation defects, cell-type specific

Present addresses: [6]Biomedical Sciences Graduate Program, University of California, San Francisco, CA 94143, USA; [7]Department of Biostatistics and Bioinformatics, and The Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708, USA.
[8]Corresponding author
E-mail rmyers@hudsonalpha.org
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.147942.112.

methylation, and epigenetic changes induced by cell culture. We quantify the reciprocal relationships of promoter and gene body methylation to expression levels and present the identification of a DNA methylation signature of intragenic (gene body) transcriptional enhancers marked by EP300 that clarifies the interpretation of gene body methylation. We report the discovery of reproducible non-CpG cytosine methylation in human somatic tissue, particularly in adult human brain tissues. The data described here provide an atlas of DNA methylation for investigating how this epigenetic mark relates to other molecular and phenotypic characteristics within these diverse cell types, including many commonly used cell line models. The data are readily available for visualization and analysis using the UCSC Genome Browser (Fujita et al. 2011; Raney et al. 2011) and will provide a valuable resource for future comparisons to other cell types, disease states, and functional genomic assays.

Results

Quantifying DNA methylation

We modified a previously published protocol for RRBS (Meissner et al. 2008) to create a streamlined workflow for this larger-scale implementation (Supplemental Fig. S1). For each reference cytosine sequenced, we computed the percentage of reads in which the cytosine was methylated (remained a C after bisulfite treatment) out of the total reads covering that position (Supplemental Fig. S2). This percent methylated (PM) value represents the percentage of molecules that were methylated at each cytosine.

In replicate growths of the human myeloid cell line K562, we found that lower read depth provided less reproducible measurements of PM between replicates. When restricted to CpGs with at least 10× coverage, the reproducibility of PM measurements across the full range of values improved between replicates (r = 0.987) and resulted in an average of 3.96 PM difference per CpG between replicates (Supplemental Fig. S3A). The resulting PM measurements were highly correlated with values obtained from an array-based methylation assay (Illumina Methyl450K, r = 0.954) (Supplemental Fig. S3B). We performed RRBS on 82 human cell lines and tissues in duplicate (sample information in Supplemental Table S1) and obtained at least 10× coverage for an average of 700,000 CpGs in each sample. The MspI restriction digest utilized for RRBS enriches for CpG-dense regions of the genome, including genes (twofold enrichment) and CpG islands (111-fold enrichment).

Global observations

We found that all samples, regardless of the disease state or tissue type, had similar distributions of methylation among the assayed CpGs (mean of pairwise $R^2 = 0.96$) (Supplemental Fig. S3C). In each sample, 5%–15% of assayed CpGs were completely methylated (PM ≥ 90), and 65%–80% of assayed CpGs were unmethylated (PM ≤ 10). This consistency could appear because the same CpGs are always methylated in all samples, or it could result from the same net amount of methylation placed on different loci between samples. Our data set supports the latter; the PM values of *individual* CpGs varied substantially across cell lines and tissues. Only 4% (27,053) of CpGs are unmethylated (≤10 PM) across all cell lines and tissues that we assayed, and these are located near the transcription start sites (χ^2, $P < 0.0004$) of genes with housekeeping functions ($P < 1.35 \times 10^{-10}$). The remaining 670,000 CpGs that we assayed exhibited differential methylation in this study, providing a rich data set for investigating epigenetic patterns.

To characterize cell-type specific methylation patterns, we analyzed 440,974 autosomal CpGs with at least 10× coverage in at least 90% of the samples. We performed unsupervised hierarchical clustering on the PM values for the top 5% of CpGs with the most variable methylation ($N = 22{,}696$, $\sigma \geq 32.6$). The samples all paired with their replicates and clustered into clades with distinct methylation patterns, and these clades corresponded to distinct types of samples, namely cancer cell lines, primary cell lines, tissues, and blood leukocytes (Fig. 1A; detailed tree in Supplemental Fig. S4). We divided these most variable CpGs based on whether they are located in gene regulatory regions (<2000 bp from the transcription start site) or in the body of genes (>2000 bp from the transcription start site) and found that both subsets recapitulate the classification of samples in the four clades (Supplemental Fig. S5). To determine if the detection of cell-type specific differences would be confounded by epigenetic variability introduced when cell lines were grown in different ENCODE Consortium laboratories, we obtained growth replicates of the same cell lines from different laboratories. We found that they clustered together based on cell type, not laboratory. Furthermore, for a particular cell type, replicates from within a lab and replicates grown in distant labs were equally similar (Supplemental Fig. S6).

Aberrant methylation across cancer cell lines

The dominant signal in this data set is cancer-specific hypermethylation found at 66,570 CpGs in the cancer cell lines (Fig. 1A), including lines derived from breast, prostate, lung, ovarian, endometrial, liver, and pancreatic cancer, as well as neuroblastoma and several leukemias (Kolmogorov-Smirnov, $P < 1 \times 10^{-7}$). Of the loci that are hypermethylated across cancers, 48,787 (73%) reside in CGIs (Table 1) and represent a significant portion of the 148,465 island CpGs assayed (33% vs. expected 15%, Fisher's exact test, $P < 0.05$). The observation of hypermethylation at CGIs across the genome, regardless of their proximity to genes, in 18 diverse cancer cell lines is consistent with reports of a CpG island methylator phenotype (CIMP) that was first described in colorectal cancer (Toyota et al. 1999) and has since been documented in many cancer types (Teodoridis et al. 2008; Liu et al. 2011; Chen et al. 2012; Turcan et al. 2012). An additional 7377 (11%) *nonisland* CpGs are significantly hypermethylated in cancer, and these reside in promoters and bodies of genes encoding proteins with sequence-specific DNA binding transcription factor activity (hypergeometric enrichment for GO:0003700, $P = 5.4 \times 10^{-15}$) (Table 1). The presence of increased methylation in both the promoter and gene body of these transcription factor genes indicates dysregulation of methylation at these genes in cancer cell lines.

It has been observed that hypermethylation in cancer is enriched at loci that in embryonic stem cells (ESCs) are unmethylated, have bivalent chromatin marks, and are reversibly repressed by the Polycomb Repressive Complex. This observation has led to the model of a stem/progenitor cell signature in cancer (Ohm et al. 2007; Schlesinger et al. 2007; Widschwendter et al. 2007; Easwaran et al. 2012). We determined the overlap between the loci that are hypermethylated across the cancer cell lines in our data set and the location of binding sites for 149 transcription factors that were assayed by chromatin immunoprecipitation sequencing (ChIP-seq) experiments performed by our lab and others in The ENCODE Project Consortium. For each transcription factor that overlapped loci that are hypermethylated in cancer cell lines,

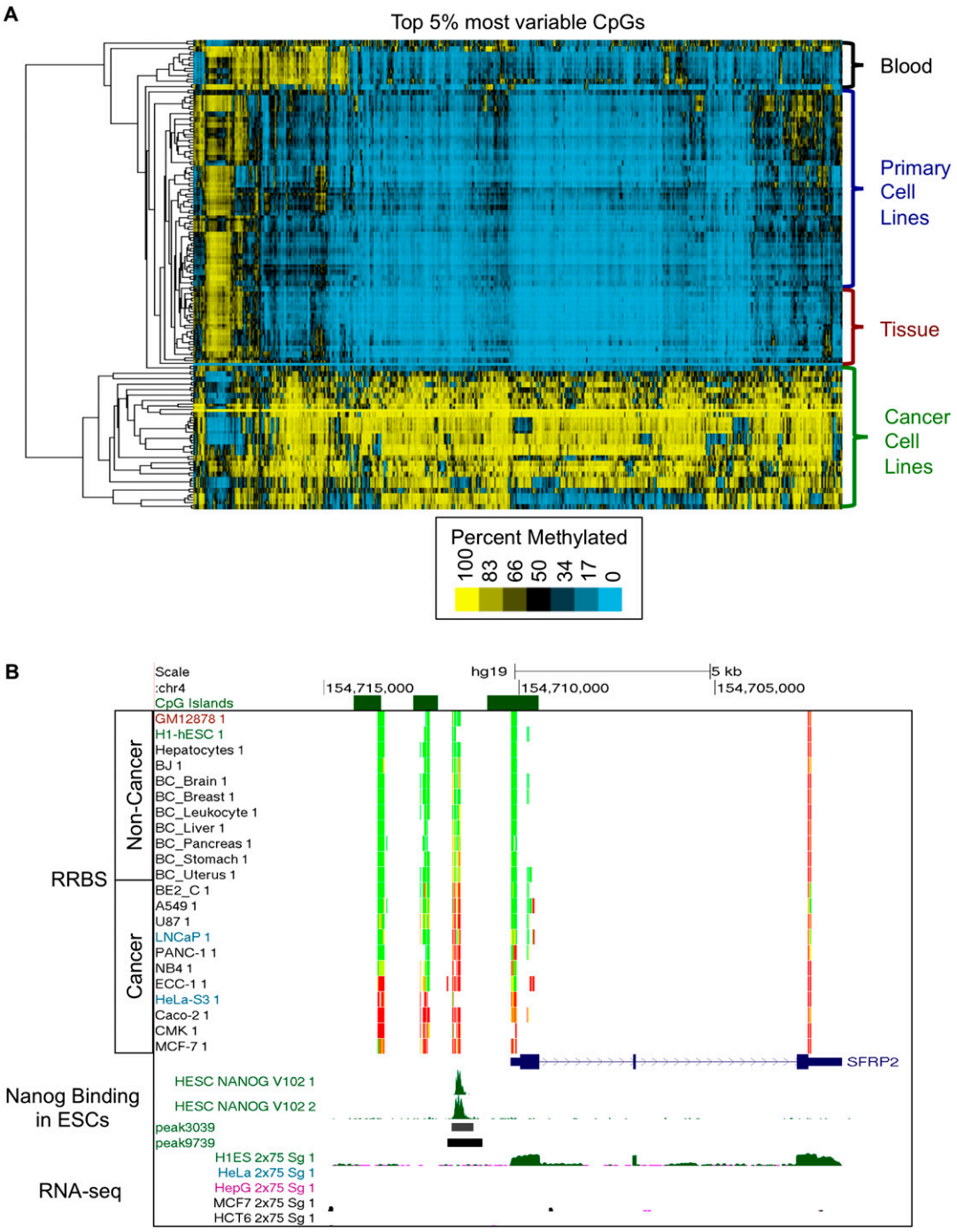

Figure 1. Methylation patterns distinguish cell types and reveal aberrant hypermethylation across cancers. (*A*) Unsupervised hierarchical clustering of the *top* 5% of CpGs with the most varying methylation across 82 samples distinguishes four major clades, identified as cancer cell lines, tissues, primary cell lines, and blood leukocytes. (*B*) Loci that are hypermethylated across cancers are significantly enriched for sites that are bound by NANOG in embryonic stem cells. UCSC Genome Browser visualization of the *SFRP2* gene showing DNA methylation data, NANOG binding sites in the embryonic stem cell line H1-hESC (H1-hESC), and RNA-seq data. The color in the RRBS track indicates the percent of molecules that are methylated at each CpG position. (Red) 100%, (yellow) 50%, (green) 0%. Hypermethylation across the cancers occurs in the *SFRP2* gene promoter where NANOG, a transcription factor, binds in H1-hESC. NANOG binding in H1-hESC is visualized as green peaks in both ChIP-seq replicates, and peak boundaries are depicted as black and gray boxes *below* the raw signal (darker boxes indicate a more significant peak). The RNA-seq data demonstrate that *SFRP2* is expressed in H1-hESC and not expressed in the cancer cell lines (HeLa, HepG2, MCF7, and HCT116).

we report the enrichment and statistics in Supplemental Table S2. Consistent with previous reports, we found that hypermethylated CpGs have a significant overlap with loci that are bound by SUZ12 in ESCs, a component of Polycomb Repressive Complex 2 (Fisher's exact test, Benjamini–Hochberg (BH)-adjusted, $P = 9.27 \times 10^{-142}$). The corepressor CTBP2 was also enriched at these sites and sig-

Table 1. Genomic context of cell-type specific methylation

		Near TSS (<2 Kbp)		Far from TSS in gene body		Intergenic	
	CGI	Hypo	Hyper	Hypo	Hyper	Hypo	Hyper
Cancer-specific	+	78 (0.3%)	26,434 (99.7%)	130 (1.5%)	8701 (98.5%)	291 (2.1%)	13,652 (97.9%)
	–	673 (15.5%)	3657 (84.5%)	1628 (30.4%)	3720 (69.6%)	3890 (51.1%)	3716 (48.9%)
Blood-specific	+	59 (4.2%)	1359 (95.8%)	103 (11.9%)	763 (88.1%)	50 (8.9%)	511 (91.1%)
	–	265 (28.3%)	673 (71.7%)	456 (29.9%)	1069 (70.1%)	360 (29.9%)	843 (70.1%)
Tissues vs. primary cell lines	+	71 (44.9%)	87 (55.1%)	257 (100%)	0 (0%)	43 (43.4%)	56 (56.6%)
	–	152 (33.7%)	299 (66.3%)	718 (100%)	0 (0%)	239 (43.1%)	316 (56.9%)

nificantly overlapped SUZ12-bound loci in ESCs (Fisher's exact test, BH-adjusted, $P = 9.75 \times 10^{-20}$). Additionally, we discovered that hypermethylated loci were also enriched for sites where NANOG is bound in ESCs, which is a transcription regulator essential for maintaining pluripotency in ESCs (Fisher's exact test, BH-adjusted, $P = 4.06 \times 10^{-4}$). The hypermethylated NANOG binding sites do not overlap the SUZ12-bound loci, and the genes nearest the hypermethylated NANOG binding sites are enriched for genes that are overexpressed in human ESCs ($P = 6.64 \times 10^{-5}$) (Ben-Porath et al. 2008). These observations support the role of embryonic transcriptional regulators in directing aberrant methylation that leads to cancer but add complexity to the model. In contrast to the polycomb repressive complex recruiting hypermethylation in cancer to persistently silence differentiation-inducing genes, our results suggest that hypermethylation also occurs at NANOG-bound loci that are active in ESCs. While these are seemingly opposing effects, it is possible that both mechanisms lead to tumorigenesis by silencing genes that lead to differentiation, as well as silencing genes that control the stable, nonneoplastic division of a pluripotent cell. This is supported by the observation that NANOG binding site hypermethylation in cancer occurs at genes enriched for transcription factor activity ($P = 9.78 \times 10^{-5}$), including developmental regulators such as *FOXD3*, a negative regulator of cell cycle (Abel and Aplin 2010), and *CDX2*, whose silencing is associated with cancer progression (Huang et al. 2012; Knosel et al. 2012). Figure 1B depicts another example, the *SFRP2* gene, a modulator of Wnt signaling, whose promoter is unmethylated and bound by NANOG in the H1 ESCs where the gene is expressed. In cancer, the NANOG binding site in the promoter is methylated, and the gene is silenced.

Hypomethylation across the genome has been reported in cancer (Irizarry et al. 2009; Hansen et al. 2011; Berman et al. 2012; Hon et al. 2012). Although RRBS enriches for CpG-rich regions of the genome that tend to be hypermethylated in cancer, we also query thousands of positions in low CpG-density regions. We detected 6691 positions that were significantly hypomethylated across the cancer types when compared to the primary cell lines and tissues (Kolmogorov-Smirnov, $P < 1 \times 10^{-7}$). We discovered that these hypomethylated loci were colocalized in the genome and identified 114 independent megabase windows that had significantly more hypomethylated loci than expected, taking into account the nonrandom genomic coverage of the assay (Fisher's exact test, $P = 2 \times 10^{-5}$) (genomic coordinates listed in Supplemental Table S3). These megabase-size hypomethylated domains were significantly depleted of genes (binomial, $P < 3.48 \times 10^{-40}$) and were significantly enriched in the 10-Mbp ends of chromosomes and near the centromere, although they are not usually directly adjacent to the telomeres or centromeres (binomial, $p = 4.41 \times 10^{-19}$) (Fig. 2A). Recent reports have found that some tumors exhibit hypomethylation corresponding to lamina-associated domains (Berman et al. 2012), but we did not observe an enrichment for lamin B1 binding in the hypomethylated domains we identified. It has also been reported that hypomethylated regions in cancer correspond to H3K27me3 (Hon et al. 2012; Statham et al. 2012), and we found that the domains we identified overlap long tracks of H3K27me3, as well as corresponding stretches of EZH2 binding. We compared the presence of H3K27me3 in these regions between cancer cell lines and noncancer cell lines and discovered that the long tracks of H3K27me3 were specific to the cancer cell lines (χ^2 test, $P = 2.4 \times 10^{-27}$). Brinkman et al. demonstrated that when DNA methyltransferases are knocked-out in ESCs, broad local enrichments (BLOCs) of H3K27me3 appear in place of high levels of methylation (Brinkman et al. 2012). It is plausible that the same process is directing H3K27me3 BLOC formation at these hypomethylated domains in cancer. Among the ChIP-seq experiments comprising 149 transcription factors, we did not identify any factors that were particularly enriched in these domains. Notably, these hypomethylated domains occasionally contain a gene that is expressed in specific cancer samples, and those samples exhibit gene-body methylation within the hypomethylated domain corresponding to the gene's expression. This indicates that gene expression and gene-body methylation are not occluded from these regions by the unmethylated tracks of H3K27me3 surrounding the gene. The prevalence of these domains across cancer types warrants further investigation of these regions as predictive biomarkers and to uncover the mechanisms driving these massive defects. Figure 2B depicts an example of a domain at the end of the q-arm of chromosome 22, where a 2-Mb gene-depleted region is specifically hypomethylated across cancer cell lines and exhibits long tracks of cancer-specific H3K27me3 and EZH2 binding. This hypomethylated domain is flanked by methylation corresponding to the gene-body methylation of expressed genes.

The single nucleotide resolution of bisulfite sequencing allows us to detect both methylation and DNA sequence variants in the same molecules, and we used this information to identify loci with allele-specific or allele-biased methylation (Gertz et al. 2011). We identified 1144 CpGs that were adjacent to an allelic variant and exhibited allele-biased methylation in the noncancer samples from different tissues and individuals. We found that 1027 (90%) of these CpGs, which are allelically methylated in noncancer samples, exhibit aberrant methylation in cancer cell lines (Kolmogorov-Smirnov, FDR-adjusted, $P < 0.05$) (Supplemental Fig. S7A). This aberrant methylation occurs as either hypermethylation (gain of methylation on the unmethylated allele) (example in Supplemental Fig. S7B) or hypomethylation (loss of methylation on the methylated allele) (example in

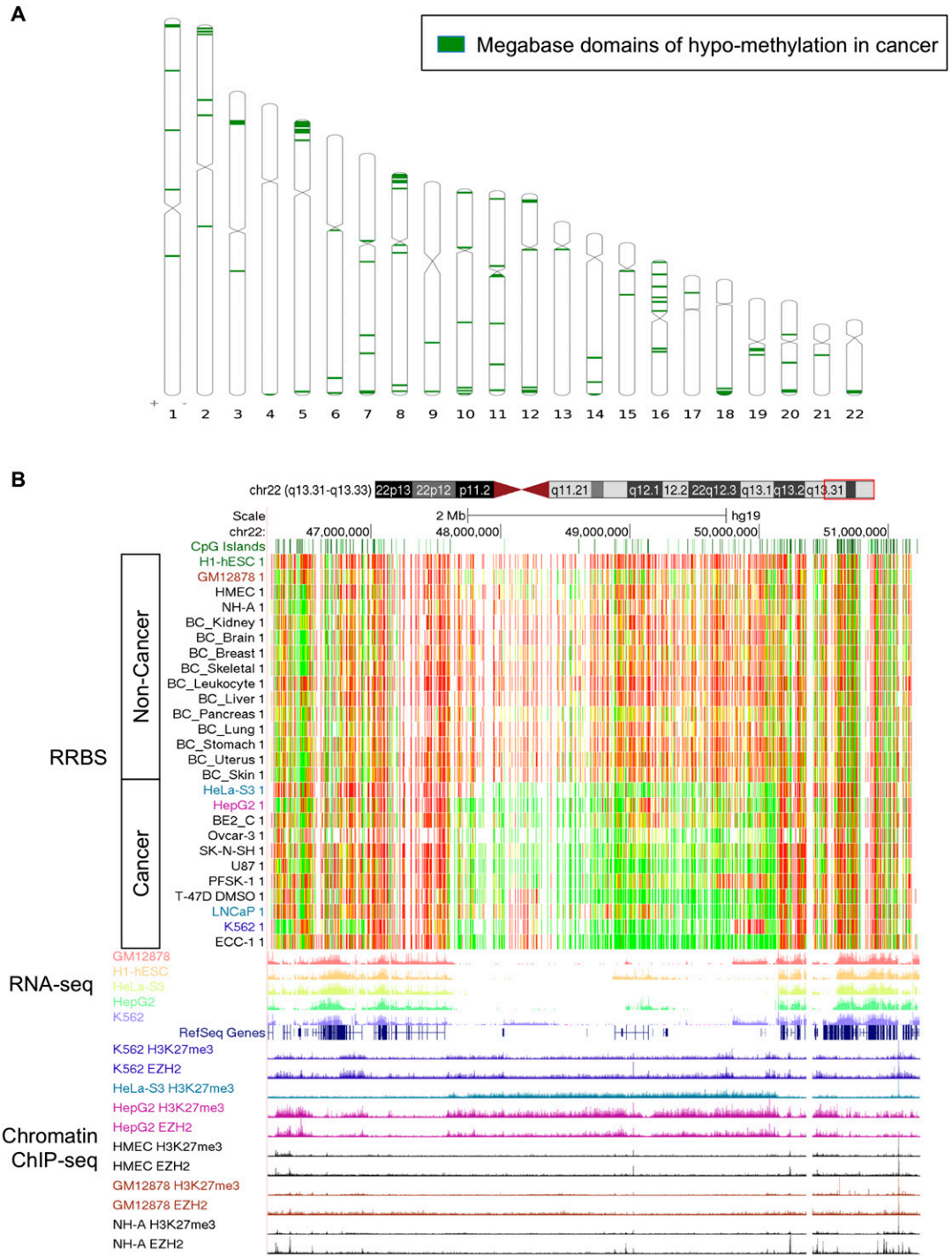

Figure 2. Megabase-size domains are hypomethylated across cancers. (*A*) We identified 114 megabase windows in the genome that are significantly hypomethylated across cancer cell lines, compared to primary cell lines and tissues. These domains are enriched near the ends and centromeres of chromosomes. (*B*) UCSC Genome Browser visualization of a 2-Mb hypomethylated domain on the q-arm of chromosome 22. The color in the RRBS track indicates the percent of molecules that are methylated at each CpG position. (Red) 100%, (yellow) 50%, (green) 0%. Hypomethylation across cancers occurs in the 2-Mb gene-depleted region. RNA-seq demonstrates that the methylated regions flanking the cancer-specific hypomethylated domain contain genes that are expressed in both the cancer (HeLa, HepG2, and K562) and noncancer samples (GM12878 and H1-HESC). The chromatin ChIP-seq tracks demonstrate that the hypomethylated region is marked by cancer-specific repressive H3K27me3 and EZH2 binding (cancer = K562, HeLa, HepG2; noncancer = HMEC, GM12878, NH-A).

Supplemental Fig. S7C). Loss-of-imprinting has been previously reported at particular imprinted loci in cancer (Feinberg et al. 2002; Bjornsson et al. 2007; Feinberg 2007; Monk 2010), and these observations demonstrate that the majority of allelically methylated loci are dysregulated in cancer, regardless of whether they are imprinted.

Cell-type specific methylation

When we isolated and subjected noncancer samples to unsupervised hierarchical clustering of the top 5% of CpGs with the most variable methylation (N = 22,152, $\sigma \geq 24.5$), the samples again clustered into clades corresponding to distinct types of biological samples: tissues, primary cell lines, embryonic cell lines, and blood leukocytes (Fig. 3A; detailed tree in Supplemental Fig. S8). The blood leukocyte-derived samples, including both peripheral blood leukocytes and Epstein-Barr virus (EBV)-immortalized lymphoblastoid cell lines, clustered together and displayed a distinct pattern of methylation at 6511 CpGs (Kolmogorov-Smirnov, $P < 1 \times 10^{-7}$). Blood-specific hypermethylation occurs at CGIs (Table 1), but unlike the ubiquitous CGI methylation in cancer cell lines, only a select subset of CGIs (66%) exhibit hypermethylation when compared to tissue and primary cell lines. Epigenetic regulation of genes involved in blood leukocyte development was evident and included specific methylation at genes involved in regulation of body fluid levels (GO:0050878, $P = 3.25 \times 10^{-4}$), blood coagulation (GO:0007596, $P = 5.33 \times 10^{-4}$), hematopoietic or lymphoid organ development (GO:0048534, $P = 8.52 \times 10^{-4}$), and B cell activation (GO:0042113, $P = 9.24 \times 10^{-4}$).

We identified seven tissue types that were represented by both primary cell lines and primary tissues and performed ANOVA to identify CpGs whose methylation is significantly associated with the tissue of origin (regardless of whether it has been grown in culture). We identified 117,795 CpGs whose methylation was significantly associated with the tissue of origin and was consistent in both the primary cell lines and the primary tissues (FDR < 0.05) (subset visualized in Fig. 3B). These data support previous observations that there is a large number of tissue-specific differentially methylated regions (tDMRs) (Rakyan et al. 2008) and that primary cell lines can serve as models for understanding epigenetic tissue-specific gene regulation at a large number of loci. However, as described above, we found that unsupervised hierarchical clustering of the most variable CpGs across samples divided the primary cell lines from the tissues (Figs. 1A, 3A). We identified 2238 CpGs that significantly discriminate primary cell lines from tissue samples (Kolmogorov-Smirnov, $P < 1 \times 10^{-7}$) (Table 1; loci listed in Supplemental Table S4). These methylation differences associated with cell culture occur predominately as unmethylated CpGs in the bodies of genes involved in regulating cellular proliferation (GO:0042127, $P = 5.17 \times 10^{-4}$). Studies of DNA methylation that use cell lines as model systems could use this list to reduce false positives due to the epigenetic effects of cell culture.

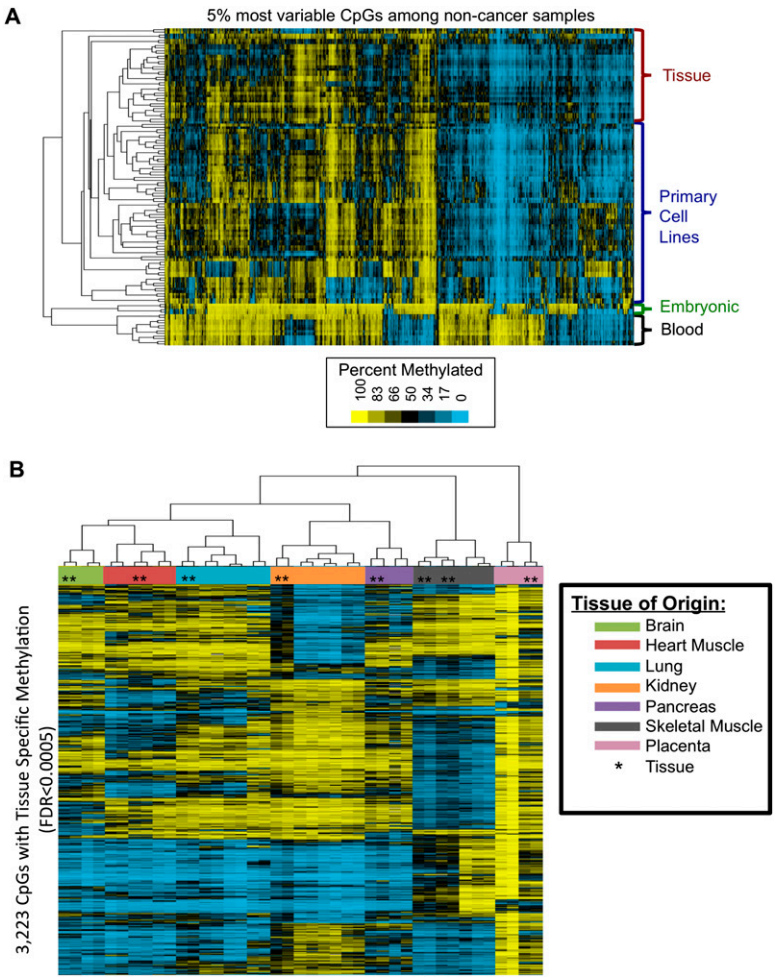

Figure 3. Noncancer samples exhibit methylation differences associated with cell culture, as well as tissue-specific methylation that is preserved between primary cell lines and tissues. (*A*) Unsupervised hierarchical clustering of the *top* 5% of CpGs with the most varying methylation across noncancer samples separates clades of samples characterized as tissues, primary cell lines, embryonic cell types, and blood leukocytes. The tissues and primary cell lines are divided into separate clades by a cell culture-associated methylation signature. (*B*) Seven tissue types were represented by both primary cell lines and tissues in this data set (tissue types listed in legend). ANOVA identified 117,795 CpGs significantly associated with tissue of origin (FDR < 0.05). For this visualization, we performed unsupervised hierarchical clustering on the 3223 significant CpGs with the largest standard deviation of PM values across the samples (SD ≥ 26). Both primary cell lines and primary tissues share a common tissue-specific methylation pattern, and the heat map displays the methylation patterns associated with each tissue of origin. Many CpGs are partially methylated in the tissues (black = 50%) at loci where the cell lines are completely methylated (yellow = 100%), indicating that heterogeneity among the cell types comprising the tissues results in a dampened signal compared to the pure cell population isolated in a cultured cell line (tissues marked by *, cell lines unmarked).

Context-dependent DNA methylation signatures of gene expression

The current models describing the relationship between DNA methylation and gene expression indicate that promoter methylation is associated with gene silencing, and gene body methylation is associated with expression (Doerfler et al. 1989; Jones and Baylin 2002; Lorincz et al. 2004; Ball et al. 2009; Illingworth et al. 2010; Maunakea et al. 2010; Aran et al. 2011; Deaton et al. 2011). RNA-seq has been performed as part of The ENCODE Project on several of the samples included

in our DNA methylation study, providing an opportunity to explore and quantify this relationship. We divided the CpGs near genes into four categories that were distinguished based on whether or not they reside in CGIs and whether they are near the TSS (<2 kb upstream or downstream) or far from the transcription start site in the gene body (≥2 kb). We computed the Pearson correlation coefficient between RPKM values (reads per kilobase of transcript per million reads) (Mortazavi et al. 2008) measured by RNA-seq and PM values measured by RRBS. In this data set, the vast majority of CpGs close to the TSS, regardless of whether they reside in CGIs, are negatively correlated with gene expression (median r = −0.3756), i.e., increased methylation is associated with lower levels of gene expression (Fig. 4A,B). In contrast, the *nonisland* CpGs far from the TSS in the body of the gene are positively correlated with expression (median r = 0.44450), i.e., increased methylation is associated with higher levels of gene expression (Fig. 4D). For these promoter and nonisland gene body CpGs, the current model of the relationship between methylation and expression holds across these cell lines. However, we also identified an exception to the expected pattern: *island* CpGs residing in gene bodies displayed a bimodal distribution of correlation with expression levels (Fig. 4C). This indicates that a substantial subset of CGIs in the gene body are negatively correlated with expression, similar to promoters, rather than positively correlated with expression like other gene body CpGs.

To investigate gene regulatory processes that could account for this finding, we integrated data sets from The ENCODE Project, including CAGE tag sequencing, histone modification, and transcription factor ChIP-seq. Recent studies have proposed that cryptic or alternative promoters, marked by H3K4me3 and CAGE tags, may appear as promoter-like methylation in gene bodies (Illingworth et al. 2010; Maunakea et al. 2010; Deaton et al. 2011). We found that H3K4me3 and CAGE tags account for 8.5% (479/6155) of the gene body CGI CpGs that were negatively correlated with expression, providing evidence that this subset of CGIs indeed reside in alternative promoters (Fisher's exact test, *P*-value = 1.06×10^{-7} [H3K4me3] and 0.009124 [CAGE tags]).

We sought to identify other function elements that could explain the remaining gene body CGIs that have epigenetic regulation similar to promoters. A recent study showed that DNA binding factors influence DNA methylation and that the methylation signature could be used to identify CpG-poor distal enhancers in the mouse genome (Stadler et al. 2011). The causes and consequences of gene-body methylation are not well understood, making it unclear whether active enhancers could be unmethylated when embedded in the densely methylated gene body of an expressed gene. We investigated whether the unmethylated CpG-rich regions that we observe within methylated gene bodies might be intragenic enhancers. We performed ChIP-seq for the transcriptional coactivator EP300, a factor known to bind to transcriptional enhancers (Visel et al. 2009). The vast majority of CpGs in EP300 binding sites were unmethylated (PM ≤ 10; 99.3% in HepG2, 98.2% in GM12878, 99.6% in hESC H1). When gene body CGIs contain EP300 binding sites, they are more strongly inversely correlated with gene expression than are CGIs not bound by EP300 (Kolmogorov-Smirnov test, *P*-value < 2.2×10^{-16}) (Supplemental Fig. S9). These EP300-bound CGI intragenic enhancers account for an additional 8% (452/5676) of gene body CGIs that are negatively correlated with expression. The signature of CpG-rich active enhancers in gene bodies was not observed at CpG-poor regions, indicating that this escape from gene-body methylation is associated with CGIs. To identify other transcription factors associated with these unmethylated gene body CpG

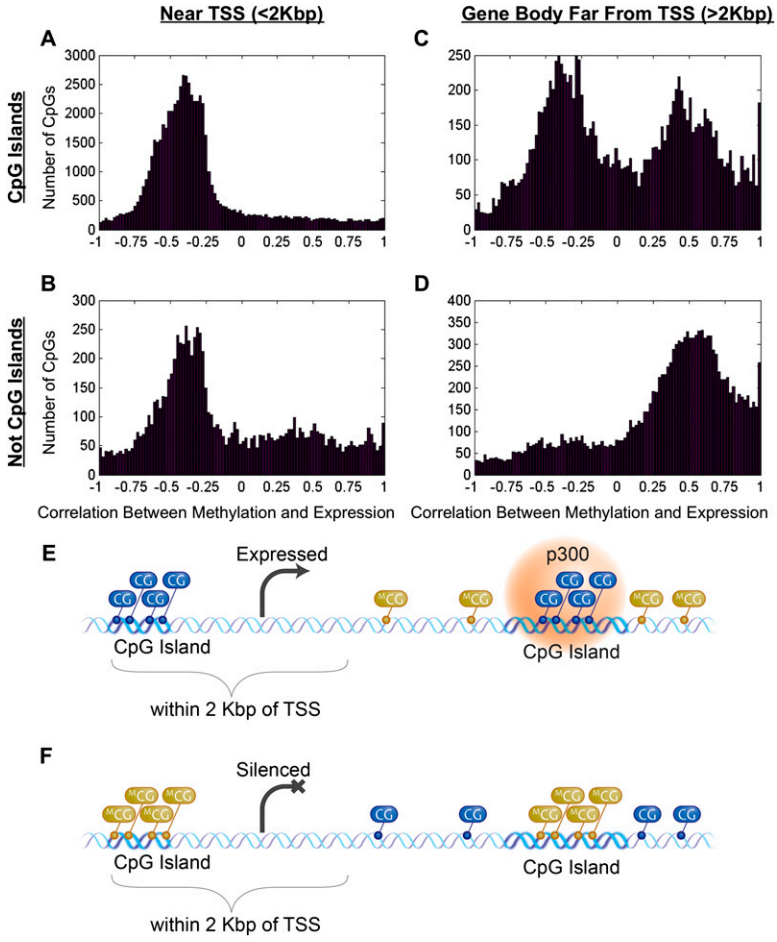

Figure 4. Correlation between CpG methylation and gene expression depends on genomic context. (*A,B*) CpGs <2000 bp away from the transcription start site (TSS) are negatively correlated with expression, regardless of whether they reside in a CpG island. (*C*) CpGs that are in gene bodies far from the TSS (>2000 bp away) and reside in CpG islands can be either positively or negatively correlated with gene expression. (*D*) CpGs that are in the gene body far from the TSS (>2000 bp away) and do not reside in CpG islands and are positively correlated with gene expression. (*E*) Model of relationship between methylation and gene expression. Expressed genes are associated with unmethylated promoters, methylated gene bodies, and unmethylated intragenic CpG island EP300-bound enhancers. (*F*) Silenced genes are associated with methylated promoters, unmethylated gene bodies, and methylated intragenic CpG island enhancer elements.

islands that may reveal the role of the remaining loci that are not associated with EP300 binding, H3K4me3, or CAGE tags, we overlapped them with the binding sites for the remaining 148 transcription factors in ChIP-seq data sets generated by The ENCODE Project Consortium. We did not identify any significant enrichment for specific transcription factors. The catalog of EP300-bound enhancers and other nonpromoter regulatory elements is not comprehensive. It is possible that these are enhancers bound by EP300 below the sensitivity of ChIP-seq, or they could be regulatory regions bound by transcription factors or noncoding RNAs not yet studied by The ENCODE Project Consortium.

Together, these results support a revised model of the expected DNA methylation state found at expressed and silenced genes, depicted in Figure 4, E and F. In expressed genes, active intragenic enhancers bound by EP300 appear as patches of unmethylated CGIs amid the dense methylation found in the body of the expressed genes, and these enhancers have methylation that is concordant with the unmethylated CpGs several thousand base pairs away in the promoter and near the TSS. The reverse patterns are associated with silenced genes. This revised model changes our expectations of the type of methylation we find in gene regions and helps to more accurately interpret gene body methylation.

Non-CpG cytosine methylation

DNA methylation predominately occurs at CpG dinucleotides in the human genome, but there have been recent reports that non-CpG cytosine methylation occurs at a lower, but appreciable, level in embryonic and pluripotent cells (Lister et al. 2009; Ziller et al. 2011). We identified 56,287 cytosines that were not located at CpG positions in the reference genome that exhibited methylation (PM > 10) in at least one sample. We eliminated 30,773 (55%) of these methylated cytosines from further analysis because they were adjacent to genetic variants in these samples that created a CpG dinucleotide that became methylated, a finding that suggests that there is a large number of polymorphic CpGs that show epigenetic diversity between individuals. To reduce false-positives due to stochastic errors in bisulfite conversion, we identified 2466 non-CpG cytosines that were methylated (PM > 10) in both replicates of any sample and determined how many of these loci were methylated in each sample (Supplemental Table S5; data in Supplemental Table S6). We found the largest number of methylated non-CpG cytosines (N = 1418) in the human embryonic stem cell line H1 (H1-hESC), consistent with previous reports in this cell type (Fig. 5A). Adult human brain tissue had the second highest number of methylated non-CpG cytosines, with 666 non-CpG cytosines methylated in both replicates, and was more than twice as high as the following sample (Supplemental Table S5). The non-CpG methylation we observed in the brain samples occurred at a different set of loci than the non-CpG methylation observed in ESCs (Fig. 5A). This was unexpected in light of the recent report of the near complete absence of non-CpG cytosine methylation in human somatic cell types, although adult

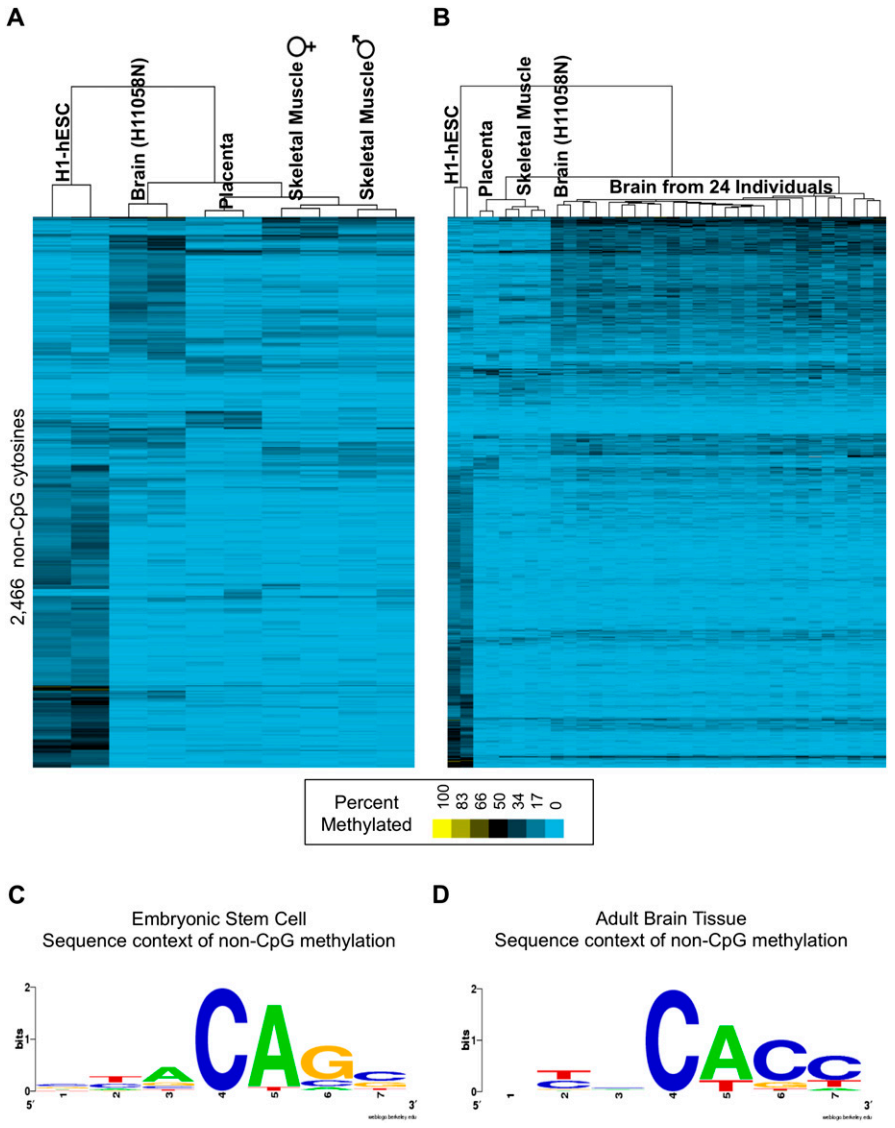

Figure 5. Non-CpG cytosine methylation. (*A*) We examined 82 cell lines and tissues and identified 2466 non-CpG cytosines that were methylated in both replicates. The samples with more than 200 methylated non-CpG cytosines are depicted. The human embryonic cell line (H1-hESC) contained 1418 methylated non-CpG cytosines, followed by adult human brain tissue (N = 666), placental tissue (N = 249), and skeletal muscle from two individuals (female N = 261, male N = 235). (*B*) The non-CpG cytosine methylation identified in the brain tissue was confirmed across post-mortem brain samples from 24 different individuals and occurs at a set of loci distinct from those methylated in the other samples. (*C*) The non-CpG cytosine methylation found in the embryonic stem cell line occurred primarily at the CAG sequence context, consistent with previous reports (Lister et al. 2009). (*D*) The non-CpG cytosine methylation discovered in the adult human brain tissue occurred primarily in the CACC sequence context.

human brain tissue was not studied (Ziller et al. 2011). The observation of non-CpG methylation in somatic tissues challenges current hypotheses that a pluripotent-specific regulation or noise is involved in establishing this mark. The other samples with more than 200 non-CpG cytosines methylated in both replicates included skeletal muscle tissue from two different individuals and placental tissue (Fig. 5A; Supplemental Table S5). To further investigate non-CpG cytosine methylation in the brain, we performed RRBS on brain samples from 24 additional individuals. These fresh-frozen samples were collected post-mortem from the dorsolateral prefrontal cortex (DLPFC) of healthy control donors as part of the Pritzker Neuropsychiatric Disorders Research Consortium. The non-CpG cytosines that were methylated in the first brain sample were also methylated (PM > 10) in these additional brain samples (Fig. 5B; data in Supplemental Table S7). The non-CpG loci that are methylated in adult brain tissue are distinct from the set of non-CpG cytosines that were methylated in ESCs (Fig. 5B). The non-CpG cytosine methylation that we observed in the brain predominately occurs at CAC trinucleotides (Fig. 5D), which is different from the CAG trinucleotide context that we and others have observed in ESCs (Fig. 5C; Lister et al. 2009; Ziller et al. 2011). Non-CpG methylation was recently reported in mouse frontal cortex (Xie et al. 2012) in a similar sequence context that we find in human brain, suggesting that mice may serve as a valuable experimental model for understanding this new methylation pattern. We used GREAT (Genomic Regions Enrichment of Annotations Tool) (McLean et al. 2010) to determine if these methylated non-CpG cytosines are associated with any functional enrichment and found that loci in brain tissue are found near genes enriched for "blood vessel development" (GO:0001568, GREAT hyper FDR q-val = 5.34×10^{-6}), while the loci methylated in hESC are found near genes related to small GTPase regulator activity (GO:0005083 hyper FDR q-val =4.91×10^{-4}), suggesting that these events are associated with different gene regulatory processes in each cell type.

Discussion

We have described an epigenomics resource generated by the ENCODE Consortium: large-scale single-base resolution DNA methylation profiling on a diverse collection of 82 human cell lines and tissues using reduced representation bisulfite sequencing. Many of these samples have been characterized with other genomic assays by The ENCODE Project Consortium, providing a rich resource for exploring functional changes associated with DNA methylation. We demonstrated that cell lines grown in replicate in multiple laboratories display stable DNA methylation signatures, that comparing methylation profiles between samples identifies methylation profiles relevant to the functional differences between cell types, and that these data provide a catalog of aberrant methylation found in cancer cell lines.

We discovered that cancer-specific hypermethylation is enriched at sites where NANOG binds in ESCs. This observation complements the previous report of a stem/progenitor signature in cancer but expands it beyond loci that are bound by Polycomb Repressive Complex 2 to include loci that are bound and activated by NANOG, a seemingly contradictory process. Further investigation is needed to understand when, during the progression of cancer, NANOG is present and how would it attract the methylation machinery to result in this aberrant hypermethylation. Additionally, we discovered that hypomethylation that is consistent across cancer types occurs in megabase-scale domains near the ends of chromosomes that contain long tracks of cancer-specific repressive H3K27me3. Further investigation is needed to understand how and why H3K27me3 repression is utilized in these regions rather than DNA methylation and whether their location on the ends of chromosomes is indicative of a structural mechanism or scaffold interaction that leads to their hypomethylation. As more genome-scale assays are performed on these samples, we are hopeful that further integrated analysis will shed light on the causes and consequences of these prolific methylation defects in cancer.

We demonstrate how integrated analysis enabled the quantification of known relationships between DNA methylation and gene expression and describe the characterization of DNA methylation at intragenic enhancers. We present a revised model of the types of methylation found in the body of expressed and silenced genes that includes our finding that unmethylated EP300-bound CGIs can be embedded in the densely methylated bodies of expressed genes. This model can be used to more accurately predict the effects of aberrant DNA methylation found in disease association studies where EP300 binding and gene expression data are not available.

The single-base resolution of RRBS enabled the detection of non-CpGs cytosine methylation across the diverse samples in this study, which led to the discovery that adult human brain tissue from many different individuals contains methylated non-CpG cytosines. We find that these loci are different from those previously identified in ESCs and occur in a CACC sequence context, rather than at CAG trinucleotides. This data set provides a launching point for the investigation of the mechanisms and function of this newly characterized phenomenon. It is intriguing that this mark is particularly abundant in brain tissue but not in the brain-derived cell lines in our study. It is plausible that the non-CpG methylation occurs in a particular type of brain cell that was not among the cell lines in this study or that the non-CpG methylation is eliminated during cell culture growth. As more genomic assay protocols are adapted to work with small amounts of tissue, we are hopeful that cell types and factors associated with the non-CpG methylation at these loci will be revealed.

Overall, we hope that this atlas of methylation across diverse samples, including many commonly used cell line models, proves to be a valuable resource for exploring how DNA methylation relates to other molecular and phenotypic characteristics.

Methods

Cell lines and tissues

Samples included in this study are listed in Supplemental Table S1. Detailed information about the samples can be obtained from the ENCODE Common Cell Types websites at (http://genome.ucsc.edu/ENCODE/cellTypes.html).

Reduced representation bisulfite sequencing experimental procedure

We modified the previously published protocol for RRBS (Meissner et al. 2008) to create a streamlined workflow for this larger-scale implementation. We designed the reactions to eliminate the phenol extraction and ethanol precipitation steps as well as one of the gel extraction steps. We also changed the PCR conditions to amplify fragments with diverse GC content and a broad range of sizes uniformly. A protocol overview is depicted in Supplemental

Figure S1, and specifics are provided as follows. We used the Qiagen DNeasy Blood and Tissue Kit to extract DNA. We then digested 1 μg genomic DNA with 1 μL 20U/μL MspI restriction enzyme (New England Biolabs [NEB]) in 1× NEBuffer 2 in a total reaction volume of 50 μL. This reaction was incubated at 37°C for 30 min, followed by heat inactivation at 80°C for 20 min. We then filled in the overhangs and added a 3′ A tail by adding dNTP mix to 33 μM and 1 μL 5U/ μL Klenow Fragment (3′–5′ exo-) in a total reaction volume of 55 μL. This reaction was incubated at 37°C for 30 min, followed by heat inactivation at 75°C for 20 min. We then purified the DNA with a Qiagen MinElute column. We purchased two methylated DNA oligonucleotides from IDT (www.idtdna.com) as follows: ilAdap Methyl PE1 (ACACTCTTTCCCTACACGACGCT CTTCCGATC*T) and ilAdap Methyl PE2 (5′P-GATCGGAAGAG CGGTTCAGCAGGAATGCCGA*G), where all C's are 5-methyl cytosine DNA nucleotides, 5′P indicates a 5′ phosphate, and an asterisk indicates a phosphorothioate bond. We then annealed these oligos to form a stock of 40 μM duplex DNA adapters in a reaction containing 40 μM ilAdap Methyl PE1, 40 μM Methyl PE2, 1× T4 DNA Ligase Buffer (NEB) in a total volume of 50 μL. We incubated this reaction at 95°C for 5 min, then 70°C for 1 min, then 60°C for 1 min, then 50°C for 1 min, then 40°C for 1 min, then 30°C for 1 min. We stored these annealed adapters at −30°C for future use. We ligated the annealed methylated Illumina adapters in a reaction containing 10 μL purified DNA, 1× T4 DNA Ligase Buffer (NEB), 1 μL 400U/μL T4 DNA Ligase (NEB), and 1 μL 40 μM annealed methylated adapters in a total volume of 20 μL. This reaction was incubated at 20°C for 15 min, followed by heat inactivation at 65°C for 10 min. We electrophoresed the 20 μL ligation reaction in a 2.5% Seaplaque Agarose (Lonza) gel. The desired restriction fragments are between 40 and 120 bp, and the adapters add 33 bp onto each end of the restriction fragments, so we used a razor blade to isolate the agarose gel section containing DNA between 106 and 186 bp, while avoiding the adapter self-ligation products that appear <100 bp. We then purified the DNA using a Qiagen Qiaquick Gel Extraction kit as described in the manufacturer's instructions, except that we did not heat the gel fragment to dissolve it, and we eluted the purified DNA from the column using 22 μL buffer EB. We then used 20 μL of this purified DNA in the sodium bisulfite conversion, which was performed using the EZ DNA Methylation Gold Kit (Zymo Research). We purchased PCR primers that would amplify the adapter-ligated DNA and add the cluster generation sequences to the amplicons for Illumina sequencing. We purchased these PCR primers from IDT as follows: ilPCR PE1 (AATGATACGGCGACCACCGAGATC TACACTCTTTCCCTACACGACGCTCTTCCGATC*T) and ilPCR PE2 (CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCC TGCTGAACCGCTCTTCCGATC*T), where an asterisk indicates a phosphorothioate bond. The DNA that was purified from the bisulfite conversion kit was then PCR-amplified in a reaction containing 5 U Platinum Taq DNA Polymerase (Invitrogen), 10× PCR Buffer without $MgCl_2$ (Invitrogen), 2 mM $MgCl_2$, 0.5 μM ilPCR PE1 DNA oligonucleotide primer, 0.5 μM ilPCR PE2 DNA oligonucleotide primer, 0.5 mM each dNTP and 0.5 M Betaine (Sigma-Aldrich) in a total reaction volume of 50 μL. The reaction was incubated at 98°C for 1 min, followed by 20 cycles of (95°C for 30 sec and 62°C for 3 min). We confirmed the amplification and correct product size range by running one-fifth of the reaction on a 2% agarose gel. We then purified the remaining PCR product with a Qiagen Qiaquick column, eluting in 25 μL buffer EB, and quantified the purified product using the Quant-IT High Sensitivity dye kit (Invitrogen) on a Qubit fluorometer (Invitrogen). We typically obtained 250–750 ng of library material. We then diluted each library to 10 nM and proceeded to sequence each library in a single lane on the Illumina Genome Analyzer IIx sequencing machine according the manufacturers instructions. We typically achieved better quality scores and alignment with slightly lower cluster density compared to other libraries with more even base-representation, and we empirically determined that clustering the sample at 5 pM was optimal.

Sequence alignment and calculating percent methylated value for each cytosine

The sequence data for this project were acquired between June 2009 and December 2010. On our Illumina Genome Analyzer IIx sequencing machine, we sequenced one library per lane and obtained between 8 million and 31 million single-end 36-bp reads per library. We aligned these reads to a modified reference genome sequence that was created to reflect both the reduced representation of the genome due to the MspI restriction digest as well as the sodium bisulfite conversion which creates a T in the sequencing reads rather than a C at all unmethylated bases. To create this reference, we first parsed the hg19 reference genome to identify all of the MspI restriction fragments <500 bp. We then isolated the 36-bp ends of these fragments into a fasta file and converted every C in the reference to a T and recorded the position of these reference cytosines in the name of the reference sequence. To achieve optimal alignment that is not biased by the methylation state of a molecule, we also created a copy of our sequencing read files, converted every C in the read to a T, and recorded the position of these read cytosines in the name of the read (Supplemental Fig. S2). We then used bowtie (Langmead et al. 2009) to align these converted reads to the custom reference sequence and required that the alignment be optimal and unique in the reference and only align to the proper strand (bowtie options –best, -m 1,–norc). On average, we uniquely aligned 53.2% (200,000/375,603) of the genomic MspI digest restriction fragments in the selected size range (40–120 bp), which resulted in coverage of an average of 1.2 million CpGs in each sample. This is only 8.6% of the 14 million nonrepetitive CpGs in the human genome but represents a 1.9-fold enrichment for genic regions and a 111-fold enrichment for CGIs. We then parsed the alignment file and the encoded read and reference names to determine how many reads covered each reference cytosine position and what percentage of those reads contained a C at each reference cytosine position (Supplemental Fig. S2). This percent methylated value approximates the percentage of molecules in the sample that were methylated at each individual cytosine. We compute the bisulfite conversion rate of each sample by determining the percent of non-CpG cytosines that are methylated (PM ≥ 10), which is an underestimate of the conversion rate in samples with biological non-CpG methylation. Each sample must meet quality control criteria before data release, including a bisulfite conversion rate ≥98.5%, a complex library with more than 500,000 CpGs with at least 10× coverage, and a correlation coefficient of greater than 0.9 between replicates.

Methyl 450 array methods

Illumina Methylation450 arrays were run using standard Illumina protocols. Briefly, 500 ng of DNA from each cell line was bisulfite-converted with the Zymo Research EZ DNA Methylation kit, amplified, hybridized, and stained with standard Illumina reagents. The intensity data were imported into Illumina's GenomeStudio software, and standard beta scores were exported and used in the analysis.

Analysis of methylation across samples

Once we compiled the percent methylated values for all cytosines shared across samples, we then performed extensive analysis of the

trends in these data. Statistical associations including mean calculations, standard deviation calculations, Pearson correlations, binomial tests, Fisher's exact tests, χ^2 tests, hypergeometric tests, and Kolmogorov-Smirnov tests were calculated using Matlab (The Mathworks, Inc.), and the statistical package R (The R Foundation for Statistical Computing; http://www.R-project.org). Clustergrams were created using the average linkage of Euclidean distance in Cluster3.0 (de Hoon et al. 2004) and visualized using Java TreeView 1.1.4r3 (Saldanha 2004). For the annotation of cytosine positions relative to gene features, we used the genomic coordinates for gene features from the hg19 refGene table of the UCSC Genome Browser (Fujita et al. 2011; Raney et al. 2011). Similarly, we used the genomic positions of the CpG islands track on the UCSC Genome Browser to annotate CGI occupancy. To measure the Pearson correlation between methylation and expression, we excluded CpGs whose methylation did not vary by 10 PM across the cell lines to avoid spurious correlations to noise in the methylation measurements. For identifying transcription factor binding sites associated with loci hypermethylated across cancer cell lines, we used the compiled supertrack data set containing binding sites for 149 transcription factors from ENCODE ChIP-seq experiments, which is available from the UCSC Genome Browser (http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=305048059&c=chr2&g=wgEncodeRegTfbsClusteredV2). For analyzing the binding sites of the specific transcription factors, we used individual ChIP-seq data sets. The data for SUZ12 binding sites in H1-hESC are available from the UCSC Genome Browser (http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=305048059&g=wgEncodeSydhTfb). The data for NANOG binding sites in H1-hESC are also available from the UCSC Genome Browser (http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=305048059&g=wgEncodeHaibTfbs). For characterizing the hypomethylated domains found across cancer cell lines, we used the H3K27me3 and EZH2 binding ChiP-seq data collected by the Broad Institute as part of The ENCODE Project, which were obtained from the UCSC Genome Browser (http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=286312585&c=chr4&g=wgEncodeBroadHistone). The nuclear lamina-associated domains data were obtained from the UCSC Genome Browser (http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=305048059&c=chr2&g=laminB1Super). The RNA-seq data for the HeLa, hESC H1, K562, HepG2, and GM12878 cell lines were collected as part of The ENCODE Project and can be found under "RNA-seq from ENCODE/Caltech" on the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=193248635&c=chr10&g=wgEncodeCaltechRnaSeq). The H3K4me3 ChIP-seq data for the HeLa, K562, HepG2, and GM12878 cell lines were collected as part of The ENCODE Project and can be found under "Histone Modifications by ChIP-seq from ENCODE/University of Washington" on the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=203697013&c=chr5&g=wgEncodeUwHistone). The CAGE tag data from whole cell polyA+ fractions of the HeLa, hESC H1, K562, HepG2, and GM12878 cell lines were collected as part of The ENCODE Project and can be found under (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=210114571&c=chr21&g=wgEncodeRikenCage). The EP300 ChIP binding site information was generated by our group as part of The ENCODE Project and can be found under "ENCODE Transcription Factor Binding Sites by ChIP-seq from HudsonAlpha Institute" on the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=210114571&c=chr21&g=wgEncodeHaibTfbs). Binding sites for EP300 identified in GM12878, H1 hESC and HepG2 were combined to create a list of potential enhancers. Genetic polymorphism that created CpG dinucleotides were identified as those positions where the bowtie alignment identified the same mismatched base in at least 10% of the reads with a minimum read depth of five. Functional annotation and enrichment of genes was obtained using the gene ontology search program Gorilla (http://cbl-gorilla.cs.technion.ac.il/) (Eden et al. 2009), using the option that calculates enrichment in a target gene list over a background list of all genes covered in the RRBS libraries. The motif logos representing the sequence context of non-CpG cytosine methylation were created using WebLogo (http://weblogo.berkeley.edu) (Crooks et al. 2004).

Data access

The DNA methylation data generated as part of The ENCODE Project are available for visualization and download from the UCSC Genome Browser (GRCh37/hg19) (http://www.genome.ucsc.edu) under the Regulation heading in the ENCODE DNA Methylation tracks. All of the RNA-seq and ChIP-seq data used in the analysis are also available for visualization and download from the UCSC Genome Browser (links are listed in Supplemental Table S1). The DNA methylation data are also available from the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) through accession numbers GSE27584 and GSE42590.

Acknowledgments

This work was supported by NHGRI grant number U54 HG004576 to R.M.M. and B.J.W. as part of The ENCODE Project. We thank members of the Pritzker Neuropsychiatric Disorders Research Consortium, particularly Drs. William Bunney, Edward Jones (deceased), Huda Akil, Stan Watson, Alan Schatzberg, and Jack Barchas for providing the 24 post-mortem brain tissues for validating the non-CpG methylation that we discovered in this study, and J.D. Frey for his assistance with figure illustrations.

References

Abel EV, Aplin AE. 2010. FOXD3 is a mutant B-RAF-regulated inhibitor of G(1)-S progression in melanoma cells. *Cancer Res* **70:** 2891–2900.

Aran D, Toperoff G, Rosenberg M, Hellman A. 2011. Replication timing-related and gene body-specific methylation of active human genes. *Hum Mol Genet* **20:** 670–680.

Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ, Church GM. 2009. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* **27:** 361–368.

Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, Regev A, Weinberg RA. 2008. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat Genet* **40:** 499–507.

Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, Noushmehr H, Lange CP, van Dijk CM, Tollenaar RA, et al. 2012. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat Genet* **44:** 40–46.

Bjornsson HT, Brown LJ, Fallin MD, Rongione MA, Bibikova M, Wickham E, Fan JB, Feinberg AP. 2007. Epigenetic specificity of loss of imprinting of the IGF2 gene in Wilms tumors. *J Natl Cancer Inst* **99:** 1270–1273.

Blelloch R, Wang Z, Meissner A, Pollard S, Smith A, Jaenisch R. 2006. Reprogramming efficiency following somatic cell nuclear transfer is influenced by the differentiation and methylation state of the donor nucleus. *Stem Cells* **24:** 2007–2013.

Brinkman AB, Gu H, Bartels SJ, Zhang Y, Matarese F, Simmer F, Marks H, Bock C, Gnirke A, Meissner A, et al. 2012. Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res* **22:** 1128–1138.

Chen HY, Zhu BH, Zhang CH, Yang DJ, Peng JJ, Chen JH, Liu FK, He YL. 2012. High CpG island methylator phenotype is associated with lymph node metastasis and prognosis in gastric cancer. *Cancer Sci* **103:** 73–79.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452:** 215–219.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator. *Genome Res* **14:** 1188–1190.

de Hoon MJ, Imoto S, Nolan J, Miyano S. 2004. Open source clustering software. *Bioinformatics* **20:** 1453–1454.

Deaton AM, Webb S, Kerr AR, Illingworth RS, Guy J, Andrews R, Bird A. 2011. Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome Res* **21:** 1074–1086.
Doerfler W, Hoeveler A, Weisshaar B, Dobrzanski P, Knebel D, Langner KD, Achten S, Muller U. 1989. Promoter inactivation or inhibition by sequence-specific methylation and mechanisms of reactivation. *Cell Biophys* **15:** 21–27.
Easwaran H, Johnstone SE, Van Neste L, Ohm J, Mosbruger T, Wang Q, Aryee MJ, Joyce P, Ahuja N, Weisenberger D, et al. 2012. A DNA hypermethylation module for the stem/progenitor cell signature of cancer. *Genome Res* **22:** 837–849.
Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10:** 48.
The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.
The ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9:** e1001046.
The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.
Feinberg AP. 2007. An epigenetic approach to cancer etiology. *Cancer J* **13:** 70–74.
Feinberg AP, Cui H, Ohlsson R. 2002. DNA methylation and genomic imprinting: Insights from cancer into epigenetic mechanisms. *Semin Cancer Biol* **12:** 389–398.
Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci* **89:** 1827–1831.
Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, et al. 2011. The UCSC Genome Browser database: Update 2011. *Nucleic Acids Res* **39:** D876–D882.
Gertz J, Varley KE, Reddy TE, Bowling KM, Pauli F, Parker SL, Kucera KS, Willard HF, Myers RM. 2011. Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *PLoS Genet* **7:** e1002228.
Ghosh S, Yates AJ, Fruhwald MC, Miecznikowski JC, Plass C, Smiraglia D. 2010. Tissue specific DNA methylation of CpG islands in normal human adult somatic tissues distinguishes neural from non-neural tissues. *Epigenetics* **5:** 527–538.
Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D, et al. 2011. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43:** 768–775.
Hon GC, Hawkins RD, Caballero OL, Lo C, Lister R, Pelizzola M, Valsesia A, Ye Z, Kuan S, Edsall LE, et al. 2012. Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res* **22:** 246–258.
Huang LP, Yu YH, Sheng C, Wang SH. 2012. Up-regulation of cadherin 17 and down-regulation of homeodomain protein CDX2 correlate with tumor progression and unfavorable prognosis in epithelial ovarian cancer. *Int J Gynecol Cancer* **22:** 1170–1176.
Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr AR, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, Bird AP. 2010. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet* **6:** e1001134.
Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, et al. 2009. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* **41:** 178–186.
Jackson M, Krassowska A, Gilbert N, Chevassut T, Forrester L, Ansell J, Ramsahoye B. 2004. Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells. *Mol Cell Biol* **24:** 8862–8871.
Ji H, Ehrlich LI, Seita J, Murakami P, Doi A, Lindau P, Lee H, Aryee MJ, Irizarry RA, Kim K, et al. 2010. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **467:** 338–342.
Jones PA, Baylin SB. 2002. The fundamental role of epigenetic events in cancer. *Nat Rev Genet* **3:** 415–428.
Kim K, Doi A, Wen B, Ng K, Zhao R, Cahan P, Kim J, Aryee MJ, Ji H, Ehrlich LI, et al. 2010. Epigenetic memory in induced pluripotent stem cells. *Nature* **467:** 285–290.
Knosel T, Chen Y, Hotovy S, Settmacher U, Altendorf-Hofmann A, Petersen I. 2012. Loss of desmocollin 1-3 and homeobox genes PITX1 and CDX2 are associated with tumor progression and survival in colorectal carcinoma. *Int J Colorectal Dis* **27:** 1391–1399.
Laird PW, Jaenisch R. 1994. DNA methylation and cancer. *Hum Mol Genet* **3:** 1487–1495.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25.
Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J, et al. 2010. Dynamic changes in the human methylome during differentiation. *Genome Res* **20:** 320–331.
Lei H, Oh SP, Okano M, Juttermann R, Goss KA, Jaenisch R, Li E. 1996. De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development* **122:** 3195–3205.
Li Y, Zhu J, Tian G, Li N, Li Q, Ye M, Zheng H, Yu J, Wu H, Sun J, et al. 2010. The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol* **8:** e1000533.
Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462:** 315–322.
Liu JB, Zhang YX, Zhou SH, Shi MX, Cai J, Liu Y, Chen KP, Qiang FL. 2011. CpG island methylator phenotype in plasma is associated with hepatocellular carcinoma prognosis. *World J Gastroenterol* **17:** 4718–4724.
Lorincz MC, Dickerson DR, Schmitt M, Groudine M. 2004. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat Struct Mol Biol* **11:** 1068–1075.
Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, et al. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466:** 253–257.
McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of *cis*-regulatory regions. *Nat Biotechnol* **28:** 495–501.
Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454:** 766–770.
Monk D. 2010. Deciphering the cancer imprintome. *Brief Funct Genomics* **9:** 329–339.
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628.
Ohm JE, McGarvey KM, Yu X, Cheng L, Schuebel KE, Cope L, Mohammad HP, Chen W, Daniel VC, Yu W, 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.
Okano M, Bell DW, Haber DA, Li E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99:** 247–257.
Rakyan VK, Down TA, Thorne NP, Flicek P, Kulesha E, Graf S, Tomazou EM, Backdahl L, Johnson N, Herberth M, et al. 2008. An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res* **18:** 1518–1529.
Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, Barber GP, Meyer LR, Sloan CA, Malladi VS, Roskin KM, et al. 2011. ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res* **39:** D871–D875.
Razin A, Riggs AD. 1980. DNA methylation and gene function. *Science* **210:** 604–610.
Saldanha AJ. 2004. Java Treeview–extensible visualization of microarray data. *Bioinformatics* **20:** 3246–3248.
Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M, Zimmerman J, Eden E, Yakhini Z, Ben-Shushan E, Reubinoff BE, et al. 2007. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat Genet* **39:** 232–236.
Sharma S, Kelly TK, Jones PA. 2010. Epigenetics in cancer. *Carcinogenesis* **31:** 27–36.
Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, et al. 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480:** 490–495.
Statham AL, Robinson MD, Song JZ, Coolen MW, Stirzaker C, Clark SJ. 2012. Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. *Genome Res* **22:** 1120–1127.
Sutter D, Doerfler W. 1980. Methylation of integrated adenovirus type 12 DNA sequences in transformed cells is inversely correlated with viral gene expression. *Proc Natl Acad Sci* **77:** 253–256.
Teodoridis JM, Hardie C, Brown R. 2008. CpG island methylator phenotype (CIMP) in cancer: Causes and implications. *Cancer Lett* **268:** 177–186.
Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JP. 1999. CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci* **96:** 8681–8686.
Tsai HC, Baylin SB. 2011. Cancer epigenetics: Linking basic biology to clinical medicine. *Cell Res* **21:** 502–517.
Turcan S, Rohle D, Goenka A, Walsh LA, Fang F, Yilmaz E, Campos C, Fabius AW, Lu C, Ward PS, et al. 2012. IDH1 mutation is sufficient to

establish the glioma hypermethylator phenotype. *Nature* **483:** 479–483.
Ushijima T. 2005. Detection and interpretation of altered methylation patterns in cancer cells. *Natl Rev* **5:** 223–231.
van der Ploeg LH, Flavell RA. 1980. DNA methylation in the human γ δ β-globin locus in erythroid and nonerythroid tissues. *Cell* **19:** 947–958.
Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457:** 854–858.
Widschwendter M, Fiegl H, Egle D, Mueller-Holzner E, Spizzo G, Marth C, Weisenberger DJ, Campan M, Young J, Jacobs I, et al. 2007. Epigenetic stem cell signature in cancer. *Nat Genet* **39:** 157–158.
Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B. 2012. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148:** 816–831.
Ziller MJ, Muller F, Liao J, Zhang Y, Gu H, Bock C, Boyle P, Epstein CB, Bernstein BE, Lengauer T, et al. 2011. Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS Genet* **7:** e1002389.

Received August 16, 2012; accepted in revised form November 21, 2012.