MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome

David Serre[1,*], Byron H. Lee[2] and Angela H. Ting[1,*]

[1]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic Foundation, 9500 Euclid Ave, mail code NE50 and [2]Glickman Urological and Kidney Institute, Cleveland Clinic Foundation, 9500 Euclid Ave, mail code Q10, Cleveland, OH, 44195, USA

Received September 9, 2009; Revised October 14, 2009; Accepted October 15, 2009

ABSTRACT

DNA methylation is an epigenetic modification involved in both normal developmental processes and disease states through the modulation of gene expression and the maintenance of genomic organization. Conventional methods of DNA methylation analysis, such as bisulfite sequencing, methylation sensitive restriction enzyme digestion and array-based detection techniques, have major limitations that impede high-throughput genome-wide analysis. We describe a novel technique, MBD-isolated Genome Sequencing (MiGS), which combines precipitation of methylated DNA by recombinant methyl-CpG binding domain of MBD2 protein and sequencing of the isolated DNA by a massively parallel sequencer. We utilized MiGS to study three isogenic cancer cell lines with varying degrees of DNA methylation. We successfully detected previously known methylated regions in these cells and identified hundreds of novel methylated regions. This technique is highly specific and sensitive and can be applied to any biological settings to identify differentially methylated regions at the genomic scale.

INTRODUCTION

DNA cytosine methylation is the covalent addition of a methyl group to the 5 position of cytosine. In humans, DNA methylation occurs predominantly in a CpG dinucleotide context and is catalyzed by DNA methyltransferases (1–3). Dense clusters of CpG dinucleotides, termed CpG islands, are present in roughly 40% of gene promoters, and methylation of these regions is associated with transcriptional silencing (4,5). DNA methylation is essential for normal developmental processes, such as imprinting (6) and X chromosome inactivation (7). Dysregulation of DNA methylation occurs in disease states such as cancer, where promoter CpG island hypermethylation leads to inactivation of tumor suppressor genes (8,9). Thus, many tumor suppressors classically identified through mutation analyses, such as *APC* (10,11), *BRCA1* (12,13), and *CDKN2A* (14,15), have also been found to be transcriptionally silenced by promoter hypermethylation. Since epigenetic abnormalities are recognized to be integral to the pathogenesis of cancer, the Cancer Genome Atlas Project has aims to map DNA methylation in several common cancers. However, current methods all have major shortcomings that prevent a truly high-throughput, unbiased, and exhaustive profiling of genomic cytosine methylation.

Bisulfite sequencing, the gold standard for cytosine methylation analysis, provides single base pair resolution of methylation patterns but requires sequencing of the entire genome (16). The characterization of DNA methylation of a single genome by bisulfite sequencing currently requires around 150 lanes of Illumina Genome Analyzer II (GAII) in order to obtain sufficient coverage to accurately and quantitatively determine the methylation state of most cytosines. For this reason, it is impractical to apply this method to the study of multiple biological samples. Alternative approaches are based on specific enrichment of methylated portions of the genome. Methylation sensitive restriction enzyme digestion allows the enrichment of highly methylated regions of the genome (17). However, it introduces recognition site biases, has a relatively poor resolution, and is prone to false positives due to incomplete digestion.

*To whom correspondence should be addressed. Tel: +1 216 444 0682; Fax; +1 216 636 0009; Email: tinga@ccf.org
Correspondence may also be addressed to David Serre. Email: serred@ccf.org.

The authors wish it to be known that, in their opinion, first and third author should be regarded as joint First Authors.

Anti-5-methyl cytosine antibody immunoprecipitation captures any DNA fragment containing one or more methylated cytosines (18). As a result, sporadically methylated sequences can comprise a significant portion of the data generated by this method. Finally, detection of methylated regions following either one of these methods is often conducted by hybridizing the DNA to a tilling array (17–20). Array design introduces an ascertainment bias, constraining novelty of the results. Additionally, probe sequences heavily influence detection specificity and sensitivity. This is particularly problematic in the context of studying DNA methylation because many methylated regions have high GC content.

To overcome these limitations, we have combined methyl CpG binding domain (MBD) precipitation of genomic DNA with massively parallel sequencing. Neither procedure introduces sequence bias, and the combination allows for high-throughput analysis of multiple samples. In this MBD-isolated Genome Sequencing (MiGS) method, we used recombinant MBD of MBD2 protein to precipitate densely methylated sequences obtained after random shearing of genomic DNA. *In vivo*, MBD2 binds specifically to methylated CpGs via its MBD and facilitates gene silencing through its innate transcriptional repression domain and recruitment of additional transcription inhibitors (21). Importantly, MBD binds with increasing affinity to multiple methylated cytosines in close proximity and will, therefore, predominantly precipitate biologically relevant, multiply methylated fragments as opposed to sporadically methylated CpGs of uncertain biological relevance (22). Random shearing of the genome by sonication minimizes sequence-specific fragmentation, as compared to restriction enzyme digestion. Finally, the massively parallel sequencing provides a high throughput detection method without any ascertainment bias.

We used MiGS to characterize genome-wide DNA methylation profiles of three isogenic human cancer cell lines harboring different levels of DNA methylation. The parental HCT116 colon cancer cells have average levels of DNA methylation found in many colon cancer cells. The DICER^{ex5} cells are derived from HCT116 through truncation of DICER1 alleles and show localized changes of DNA methylation at a small number of gene promoters (23). Finally, the *DNMT1, DNMT3b* double knockout (DKO) cells derived from HCT116 retain <5% of overall DNA methylation, as compared to HCT116 parental cells, and exhibited loss of promoter methylation at most loci analyzed so far (24). Our MiGS data describe the DNA methylation patterns in the entire genomes of these cell lines. We show that MiGS efficiently detected previously known DNA methylation and identified numerous novel DNA methylation sites. Our results strongly support that MiGS is a specific, sensitive, and high-throughput technique for the study of genome-wide DNA methylation patterns.

MATERIALS AND METHODS

Cell culture

HCT116 colon cancer cells were cultured in McCoy's 5A media supplemented with 10% fetal bovine serum. DKO (24) and DICER^{ex5} cells (25) are isogenic derivatives of HCT116 and were cultured in the same manner. Cells were harvested by scraping, and cell pellets were rinsed twice with 1X PBS.

DNA sample preparation

Genomic DNA was extracted from cell pellets by incubation in cell lysis buffer [20 mM Tris (pH 8), 20 mM EDTA, 2% SDS and 0.5 mg/ml Proteinase K] overnight at 55°C followed by phenol:chloroform extraction and ethanol precipitation. Genomic DNA was fragmented to ~150–600 bp by sonication in 1X BW buffer [4% glycerol, 1 mM $MgCl_2$, 0.5 mM EDTA, 0.5 mM DTT, 50 mM NaCl, 10 mM Tris–HCl (pH 7.4), 0.2% Tween-20 and 1X Complete EDTA-free Protease Inhibitor cocktail], and fragment sizes were verified by gel electrophoresis in 1% agarose gels. The fragmented samples were purified through QIAquick PCR cleanup columns (Qiagen) following manufacturer's protocol to exclude fragments smaller than 100 bp. Five micrograms of purified DNA fragments per sample was used in each immunoprecipitation reaction.

Isolation of methylated DNA by recombinant MBD

His-tagged recombinant MBD of MBD2 (MBD2_MBD) protein was expressed and purified as previously described (22,26). Recombinant MBD2_MBD was conjugated to magnetic beads (Dynal) by incubation overnight with rotation at 4°C. The conjugation mixture contained 41.7 μl/ml protein G magnetic bead slurry (Invitrogen), 10 μg/ml anti-his RGS antibody (Qiagen), 1 μg/ml self-ligated pCR4-TOPO vector (Invitrogen), 160 nM purified recombinant MBD2_MBD proteins and 1X BW buffer. The MBD-magnetic bead conjugates were washed with 1 volume of cold 1X BW buffer and then re-suspended in 1 volume fresh 1X BW buffer. Five milliliters of the re-suspended mixture was used to precipitate 5 μg of sheared DNA fragments by incubation overnight with rotation at 4°C. The precipitation complex was washed 5 × with 1 volume of 1X BW buffer. DNA was eluted by incubation of the precipitation complex in freshly prepared Elution Buffer [20 ml Tris (pH 8), 10 mM EDTA, 0.5% SDS and 500 μg/ml Proteinase K] at 60°C for 2 h with shaking. The eluted DNA was extracted with phenol:chloroform, precipitated by isopropanol, and re-suspended in 45 μl dH_2O.

Library preparation and sequencing

Ten nanograms of MBD-isolated DNA per sample was processed with the ChIP-Seq Sample Prep Kit (Illumina) following manufacturer's protocol to construct the sequencing libraries. Sequencing libraries were analyzed on the BioAnalyzer (Agilent), and a tight distribution of insert size around 120 bp was observed for all libraries.

Each library was sequenced on the GAII to generate 36-bp long reads.

Mapping of sequence reads

All sequence reads were mapped to the reference human genome (NCBI Build 36.1, UCSC Hg18) using the Bowtie algorithm (27). Only reads that mapped unambiguously to single genomic locations were considered for further analysis. In addition, when several reads mapped to the exact same location and in the same orientation, these reads were deemed to represent amplified products generated from a single library insert. Only one of such reads was considered for further analysis. Since the average DNA fragment length used to generate reads was 120 bp, the reference human genome was split into non-overlapping 100-bp windows. Each 36-bp sequence was extended to 120 bp and assigned to the 100-bp window that was most covered by this extended read. Sequence reads not mapped to the reference genome were mapped using Bowtie (27) against the raw Sanger sequence data for Craig Venter's genome (28).

Repeat element analysis

Sequence reads were analyzed using RepeatMasker (29) to ascertain the composition of repeat elements in each dataset.

Determination of significantly methylated loci

To determine the significance cut-off for considering one locus methylated, we assumed that windows with one or two reads were mainly background to fit a null model, which determines the number of windows with a given number of reads expected solely by chance. By comparing the observed number of windows with a given number of reads to the expected number of windows with the same given number of reads under the null model, a false discovery rate (FDR) was calculated for each window. Using the average expected number of reads per window (the total number of reads divided by the total number of 100-bp windows in the genome) as λ yielded similar FDRs. Differential methylation between HCT116 and DICER^{ex5} was assessed by testing whether the number of reads observed in each window differ between samples corrected for the total number of uniquely mapped reads in each sample using Fisher's exact test.

Sampling curves

To estimate whether the number of reads generated was sufficient to assess genome-wide DNA methylation patterns, sampling curves were generated for HCT116 and DICER^{ex5}. We combined the dataset analyzed in this article with additional 28 256 599 and 19 669 907 reads generated for HCT116 and DICER^{ex5} from the same libraries. 1, 2.5, 5, 7.5, 10, 15, 20, 30 and 35 million reads were randomly sampled without replacement to determine how many methylated windows could be identified in each subset. This analysis was performed separately for all methylated windows ($\geq$4 reads) and for highly methylated windows ($\geq$10 reads). Ten random iterations were performed at each chosen number of reads, and the average numbers are displayed in Figure 2.

Genomic annotations

Each methylated window was categorized according to its relative position to RefSeq gene annotations (UCSC Hg18). Windows located within 500 bp of transcription start sites were annotated as 5′-end, and windows within 500 bp of transcription stop sites were annotated as 3′-end. Windows located between the most upstream transcription start site and the most downstream stop site, but not previously annotated as 5′-end or 3′-end, were annotated as genic. All remaining windows were considered as intergenic. CpG island and miRNA annotations were retrieved from the UCSC Genome Browser. For miRNA, 500 bp on either side of the annotated coding sequence were used to include possible regulatory elements. A summary of the Broad Institute datasets for GM12878, HUVEC, K562 and NHEK cells was used for annotating CTCF binding sites.

Gene expression analysis

Differential gene expression data from HCT116, DAC-treated HCT116, and DKO cells were obtained from Gene Expression Omnibus (GSE4763). For each probe on the Agilent microarray, the mean log2 change in gene expression was calculated for the DAC-treated HCT116/HCT116 (chemical demethylation) and DKO/HCT116 (genetic demethylation) pairs. The significance of differential gene expression for genes with methylation differences was assessed using Student's *t*-test.

Bisulfite sequencing

Genomic DNA was bisulfite converted using the EpiTect kit (Qiagen) following manufacturer's protocol. The PCR amplicons were gel-purified and cloned into pCR4-TOPO vector (Invitrogen). At least 10 clones were sequenced individually on an ABI3730xl DNA Analyzer to ascertain the methylation patterns of each locus. The percentage of methylation is calculated as number of methylated cytosines divided by the total number of cytosines in all the amplicons analyzed. The percentages are rounded to the nearest integer in Table 2.

RESULTS

MiGS accurately identifies methylated regions

We performed MBD precipitation of methylated genomic DNA in HCT116, DICER^{ex5} and DKO cell lines, and prepared sequencing libraries for each sample. The libraries for HCT116, DICER^{ex5} and DKO were sequenced using 2, 2 and 1 lanes, respectively of a GAII. We generated 19 041 613 reads from HCT116, 18 315 610 reads from DICER^{ex5} and 4 393 056 reads from DKO. Approximately 30% of all reads could be unambiguously mapped to a unique location in the human genome (NCBI Build 36.1) using the Bowtie algorithm (27) (Table 1). Another 10–12% of the reads mapped to multiple locations in the reference genome. Finally, we postulated

Table 1. Mapping distribution of DNA sequence reads

	HCT116	DICER^{ex5}	DKO
No. of reads	19 041 613	18 315 610	4 393 056
Reads mapped uniquely to Human Genome	7 102 330 (37%)	6 920 203 (38%)	1 207 400 (27%)
Non-repeat	5 767 100	5 620 190	977 692
Repeat elements	1 335 230	1 300 013	229 708
Reads mapped to multiple locations on the Human Genome	1 978 515 (10%)	2 138 088 (12%)	439 995 (10%)
Non-repeat	811 412	793 801	103 734
Repeat elements	1 167 103	1 344 287	336 261
Reads mapped to the unassembled Human Genome	4 210 471 (22%)	3 811 890 (21%)	167 253 (4%)
Non-repeat	3 368 209	3 083 404	159 548
Repeat elements	842 262	728 486	7705
Others not mapped	5 694 794 (30%)	5 445 429 (30%)	2 578 408 (59%)

Table 2. Experimental validation by bisulfite sequencing

Locus analyzed		MiGS (No. of methylated windows)				Bisulfite Seq. (% MetC)		
Coordinates	Name[a]	Total	HCT116	DICER^{ex5}	DKO	HCT116	DICER^{ex5}	DKO
Significantly methylated in HCT116 and DICER^{ex5} in MiGS assay								
chr19:1507384-1507703	*MEX3D*	5	5	5	5	99	99	41
chr5:92949531-92949838	*NR2F1*	4	4	4	4	100	99	37
chr1:154452833-154453297	*PMF1*	5	5	5	3	98	98	31
chr1:211190670-211190970	*VASH2*	4	3	2	0	76	28	16
chrX:21302569-21302868	*CNKSR2*	4	4	4	1	95	96	13
chr1:247108292-247108567	*ZNF672*	4	4	4	4	98	98	11
chrX:23262922-23263235	*PTCHD1*	4	4	4	2	94	96	6
chr1:164401866-164402277	*FAM78B*	5	5	5	1	93	94	1
chr16:33869047-33869405	Peri-16	5	5	5	0	96	96	1
chr10:83624002-83624444	*NRG3*	5	5	5	0	94	92	1
chr20:26136535-26136825	*mir663*	4	3	3	0	99	98	1
chr1:47654991-47655443	*FOXE3*	6	6	6	0	97	98	1
chr10:79066821-79067106	*KCNMA1*	4	4	4	0	98	99	1
chr9:23810766-23810968	*ELAVL2*	3	3	3	0	98	98	1
chr9:25667441-25667703	*TUSC1*	4	4	4	0	95	92	1
chr10:31649125-31649519	*ZEB1*	5	5	1	0	72	25	0
chr3:128830800-128831068	*PODXL2*	3	3	3	0	87	34	NA[b]
No evidence of methylation in HCT116 and DICER^{ex5} in MiGS assay								
chr19:44113033-44113323	S/M[c]	4	0	0	0	1	NA	NA
chr8:145209542-145209851	*GPAA1*	4	0	0	0	1	NA	NA
chr7:105539547-105539910	*SYPL1*	5	0	0	0	0	NA	NA
chr18:10515854-10516103	*NAPG*	4	0	0	0	0	NA	NA
chr1:143920369-143920682	*NOTCH2NL*	4	0	0	0	0	NA	NA

[a] Name of the closest RefSeq gene.
[b] Bisulfite sequencing not performed.
[c] SARS2/MRPS12 bidirectional promoter.

that many sequence reads that did not match sequences on the assembled human genome originate from centromeric and sub-telomeric regions of the genome that have not been successfully assembled. These sequences, however, should be present in the shotgun sequencing data of Venter's genome (28). Indeed, 4–22% of sequencing reads matched to the unassembled Venter's genome dataset but not in the assembled reference genome (Table 1). Finally, we annotated 13–20% of the total sequences as repeat elements (Supplementary Figure S1). Sequencing reads that cannot be mapped to the human genome likely (i) contain too many nucleotide differences to the reference sequences due to sequencing errors, polymorphisms, or insertions/deletions or (ii) originate from regions overlapping genomic rearrangements (D.S. and A.H.T., unpublished data).

Reads that were unambiguously mapped to a single location in the reference human genome were assigned to non-overlapping 100-bp windows. To determine which windows contained more reads than would be expected by chance, we fitted a null distribution to our data and estimated a FDR for each window. We determined that 100-bp windows containing four or more reads were unlikely to be generated by chance ($\text{FDR} < 0.01$) and therefore significantly methylated. Overall we identified 171 338, 166 568 and 59 774 non-overlapping windows with significant evidence of DNA methylation in HCT116, DICER^{ex5}, and DKO, respectively and focused on these regions in subsequent analyses.

For validation, bisulfite sequencing was performed on 17 randomly selected regions that were identified as methylated by MiGS. These regions represent a wide

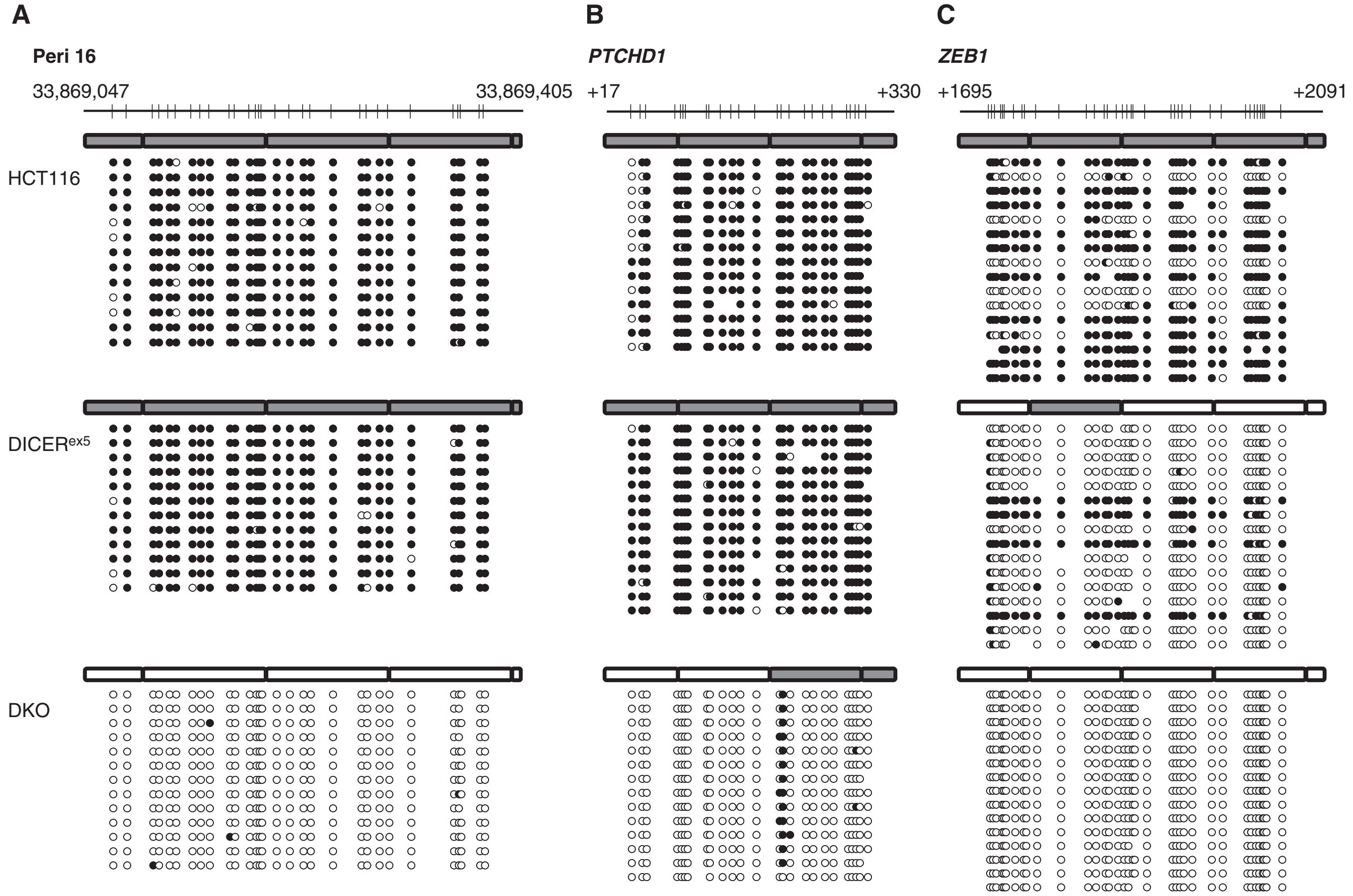

Figure 1. Bisulfite sequencing of newly identified methylated regions. Bisulfite sequencing was performed on (**A**) peri-centromeric CpG island on chromosome 16, (**B**) promoter CpG island of *PTCHD1* and (**C**) *ZEB1* intronic CpG island in HCT116, DICER^{ex5} and DKO cells. Genomic coordinates (NCBI Build 36.1) for the bisulfite sequencing amplicon of chromosome 16 peri-centromeric CpG island are shown. Positions relative to the transcription start site are indicated for the amplicons of *PTCHD1* and *ZEB1*. Each circle represents a CpG dinucleotide with black circles representing methylated cytosines and white ones representing unmethylated cytosines. Each row represents one individual allele sequenced. Rectangles above each cell line represent the non-overlapping 100-bp windows covered by each amplicon. Methylated windows identified by MiGS are shaded in gray while unmethylated windows are in white.

variety of genomic contexts, including peri-centromeric CpG islands, gene promoters, intragenic and intergenic sequences (Table 2 and Supplementary Table S1). All 17 tested regions were validated by bisulfite sequencing (Table 2). For example, MiGS determined that the peri-centromeric CpG island of chromosome 16 was methylated in HCT116 and DICER^{ex5} cells but not in DKO cells (Supplementary Figure S2). Bisulfite sequencing confirmed these patterns (Figure 1A). Conversely, the promoter CpG island of *PTCHD1* was determined to have high levels of methylation in both HCT116 and DICER^{ex5} cells and low, but significant, levels of methylation in DKO cells (Supplementary Figure S3). Bisulfite sequencing of this region substantiated dense methylation in HCT116 and DICER^{ex5} cells and robust residual methylation at two CpG sites in the DKO cells (Figure 1B). It is important to note that DKO cells have lost >95% of its global DNA methylation (24). Overall, we identified 59 774 methylated 100-bp windows in DKO cells. Bisulfite sequencing of seven such regions showed low but consistent levels of residual methylation (Table 2). In this context, MBD bound tightly to DNA fragments with residual methylation rather than binding without specificity. Consequently, we often observed in DKO cells a disproportionally large number of reads for regions with very low but robust DNA methylation. Nonetheless, loci that are entirely lacking DNA methylation in DKO cells still remain free of reads (Table 2). This indicates that MiGS is highly specific and sensitive. Finally, we identified differential methylation between HCT116 and DICER^{ex5} cells at the intronic CpG island of *ZEB1* (Fisher's exact test, two-tailed, $P = 3.1 \times 10^{-9}$) (Supplementary Figure S4). Bisulfite sequencing corroborated this finding (Figure 1C). *VASH2* ($P = 5.1 \times 10^{-4}$) and *PODXL2* ($P = 1.9 \times 10^{-6}$), also deemed to be differentially methylated between HCT116 and DICER^{ex5} cells, were both confirmed by bisulfite sequencing (Table 2).

We compared our results with previously reported methylated gene promoters in HCT116 to estimate our false negative rate (30) (Supplementary Table S2). We defined the proximal promoter as 500 bp on either side of the transcription start site. Our assay identified at least one methylated 100-bp window within the proximal

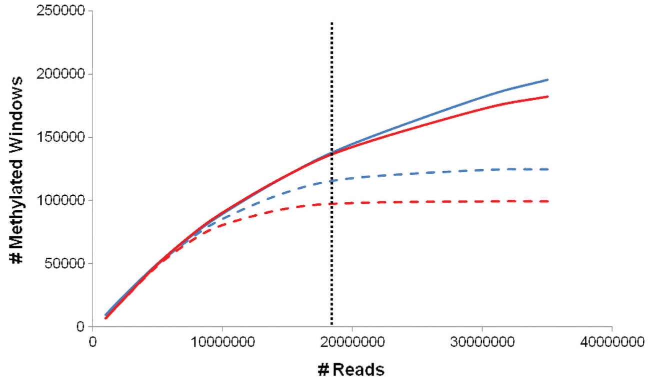

Figure 2. Sampling curves for HCT116 and DICER^{ex5}. The curves show the number of 100-bp windows identified as being methylated (y-axis) for a given number of sequences randomly drawn from the entire dataset (x-axis). Blue lines represent sampling curves for HCT116, and red lines correspond to data for DICER^{ex5}. The solid lines show the results obtained when considering all significantly methylated windows (≥4 reads), while the dash lines represent the results for highly methylated loci (≥10 reads) only. The vertical black line shows the number of reads used in this study.

promoter of 63 out of the 72 known methylated genes (87.5%). For two of the nine false negatives, *EPHA4* and *ZNF550*, methylation signals were detected close to our definition of the promoter. For *EPHA4*, methylation was detected in the CpG island in the first intron. *ZNF550* showed methylation in a CpG island 4 Kb upstream of the transcription start site. We also analyzed by bisulfite sequencing five randomly selected CpG islands without evidence of DNA methylation in our datasets. All five regions (*GPAA1*, *SYPL1*, *NAPG*, *S/M* and *NOTCH2NL*) were found to be free of DNA methylation (Table 2 and Supplementary Figure S5). Together, these data indicate that MiGS can detect DNA methylated regions on a genome-wide scale with very low false positive and false negative rates.

Finally, we estimated whether the depth of sequencing was sufficient to comprehensively characterize the entire methylome. HCT116 and DICER^{ex5} cells have similar levels of global methylation, and we would expect their methylation profiles to be similar. Indeed, 70.8% of all significantly methylated windows were detected in both datasets. Additionally, windows with high number of reads in HCT116 also displayed a high number of reads in DICER^{ex5} (Pearson's $r^2 = 0.84$, $P < 2.2 \times 10^{-16}$). Furthermore, we combined the present dataset (two lanes of sequencing each for HCT116 and DICER^{ex5}) with additional sequences generated from 3 and 2 lanes of GAII from the same libraries. Using random sampling, we assessed the number of methylated loci that can be detected with different number of sequencing reads. The sampling curves showed that we did not reach saturation in either sample and more methylated loci could be identified with additional sequencing (Figure 2). Nonetheless, >95% of highly methylated regions (≥10 reads) were captured with fewer than 8 million reads, roughly the amount of data generated from a single lane of GAII (Figure 2). Importantly, 56 out of 63 previously known methylated loci confirmed by MiGS (Supplementary Table S2) fell within this highly methylated category. Thus, while additional sequences might be necessary to identify all methylated loci, data from two sequencing lanes are sufficient to exhaustively characterize highly methylated loci.

MiGS describes distribution of DNA methylation in the genome

MiGS data can also describe the distribution of DNA methylation on a genome-wide scale. On average, 37% of total sequencing reads from HCT116 and DICER^{ex5} mapped to a unique location. Eighteen percent were from repeat elements, mostly SINEs and rRNAs (Supplementary Figure S1). An additional 22% mapped only to the unassembled human genome and likely represent centromeric/sub-telomeric regions (Table 1). This pattern is consistent with previous findings that DNA methylation prevents unwanted transcription and maintains genome integrity by condensing these elements (31). Globally, HCT116 and DICER^{ex5} cells were virtually identical in the distribution of DNA methylation, confirming the previous report (23).

We further examined the genomic distribution of unambiguously mapped DNA methylation (Table 3). Promoter CpG island methylation has been the main focus of most DNA methylation studies, but our data indicate that the bulk of DNA methylation, 89% for HCT116 and DICER^{ex5} and 95% for DKO, occurs outside of 5′ promoter regions. Moreover, while a significant proportion of DNA methylation overlapped with annotated CpG islands, >60% of DNA methylation mapped outside of CpG islands. These observations confirm that DNA methylation has additional functions in the genome other than facilitating transcriptional silencing at proximal promoters. Such functions may be achieved through methylation of non-CpG island sequences as well as CpG islands.

Nonetheless, we observed preferential DNA methylation within CpG islands in all genomic contexts. CpG islands near transcription start sites are 4–7 times more likely to be methylated than non-CpG islands in the same context (Table 3). Interestingly, in all other genomic contexts, CpG islands are 40–151 times more likely to be methylated than non-CpG island sequences. This observation is consistent with the notion that promoter CpG islands are generally protected from DNA methylation to allow transcription initiation.

MiGS identifies biologically relevant DNA methylation

Since promoter DNA methylation is known to repress gene transcription, we first analyzed the distribution of methylated regions with regards to RefSeq genes. To assess whether DNA methylation at the 5′-end of genes has functional consequences, gene expression patterns after removal of DNA methylation, by either 5-aza-2′-deoxycytidine (DAC) treatment or genetic deletion of DNA methyltransferases in DKO cells, were examined (Table 4). After demethylation, genes with methylation at the 5′-end showed a larger increase in expression when compared with all other genes (Student's t-test,

Table 3. Distribution of DNA methylation

	Hg18[a]	Distribution of methylation[b]			Methylation at CpG islands[c]		
		HCT116	DICER^{ex5}	DKO	HCT116	DICER^{ex5}	DKO
5′-end	0.82%	11.29%	10.98%	4.83%	83.66% (7×)	83.54% (7×)	76.02% (4×)
3′-end	0.71%	2.59%	2.56%	2.14%	51.14% (40×)	50.82% (40×)	43.06% (29×)
Genic	41.20%	50.20%	50.26%	60.92%	40.47% (98×)	40.23% (97×)	31.81% (67×)
Intergenic	57.26%	35.92%	36.20%	32.11%	36.51% (151×)	35.88% (147×)	23.49% (82×)
Total	100%	100%	100%	100%	44.20% (90×)	43.67% (88×)	31.52% (52×)

[a] Proportion of the human genome (NCBI Build 36.1) located within 500 bp of a RefSeq gene transcription start site (5′-end), within 500 bp of a RefSeq gene transcription stop site (3′-end), in the body of a RefSeq gene (Genic), or between RefSeq genes (Intergenic).
[b] Distribution of methylated 100-bp windows according to their genomic location.
[c] Proportion of methylated windows that overlap with annotated CpG islands. The preferential enrichment of methylation at CpG islands is shown in brackets and is calculated by dividing the proportion of methylated windows in CpG islands by the proportion of methylated windows not in CpG islands in each genomic context.

Table 4. Gene expression changes of methylated genes upon demethylation

	N	Chemical demethylation		Genetic demethylation	
		Mean change[a]	*P*-value	Mean change[b]	*P*-value
5′-end methylation	3031	1.57	$<2.2 \times 10^{-16}$	1.81	2.2×10^{-16}
No methylation	16 695	1.12		1.03	
5′-end CpG island methylation	2465	1.63	1.1×10^{-3}	1.91	3.6×10^{-4}
5′-end non CpG island methylation	566	1.33		1.45	
5′-end non CpG island methylation	566	1.33	2.4×10^{-3}	1.45	4.4×10^{-7}
No methylation	16 695	1.12		1.03	
5′end methylation	3031	1.57	$<2.2 \times 10^{-16}$	1.81	$<2.2 \times 10^{-16}$
Genic or 3′-end methylation	5294	0.99		0.95	

[a] Average of ratios of gene expression in DAC-treated HCT116 cells over those in HCT116 cells.
[b] Average of ratios of gene expression in DKO cells over those in HCT116 cells.

two-tailed, $P < 2.2 \times 10^{-16}$). Interestingly, demethylation of genes silenced by promoter CpG island hypermethylation resulted in a higher increase of transcripts when compared with demethylation of genes with hypermethylated promoters that do not satisfy the definition of a CpG island ($P = 0.0011$ for DAC treatment and $P = 0.0004$ for DKO). Although promoter CpG islands are protected from methylation, our data suggest their hypermethylation leads to efficient transcriptional silencing. Promoters with methylated CpGs, but not CpG islands, may be incompletely silenced. This hypothesis would be consistent with the smaller changes in expression observed when these genes are demethylated.

According to our data, ~90% of the uniquely mapped DNA methylation occurred outside of promoter regions (Table 3). To begin understanding the biological role of these marks, we compared them with CTCF binding sites and microRNA (miRNA) coding sequences. CTCF binding is blocked by DNA methylation (32), while miRNA transcription can be silenced by DNA methylation (33). Indeed, 16% of the detected DNA methylation overlap with CTCF binding sites and 0.2% overlap with miRNA coding regions (data not shown). These percentages are likely underestimated since the annotations for both CTCF binding sites and miRNA coding regions are still incomplete. Nonetheless, they correspond to a 4-fold enrichment of DNA methylation at CTCF binding sites (Fisher's exact test, two-tailed, $P < 2.2 \times 10^{-16}$) and an 8-fold enrichment surrounding miRNA coding sequences ($P < 2.2 \times 10^{-16}$). Although CTCF binding and miRNA regulation cannot account for all of the non-promoter DNA methylation we detected, these comparisons indicate functional relevance of non-promoter DNA methylation and underlie the advantage of studying DNA methylation without sequence context bias.

DISCUSSION

The functional genome is encoded and regulated by both the nascent DNA sequences and epigenetic modifications, such as DNA methylation. Therefore, in the post-genomic era, there is a demand for genome-wide characterization of epigenetic modifications to facilitate the full understanding of the utility and regulation of our genetic material. However, the current methodologies of large-scale DNA methylation profiling are insufficient for comprehensively surveying of DNA methylation in multiple samples simultaneously. Bisulfite sequencing requires the sequencing of the entire genome to map DNA methylation patterns in each sample. Even with massively parallel sequencing technologies, this is too expensive for most laboratories and is impractical for the study of multiple samples. Other types of assays rely on the enrichment of methylated DNA sequences followed by detection of methylated regions by hybridizing the DNA on tilling arrays (17,18). Such array-based methods restrict the discovery of DNA methylation to the probes present, which often focus on gene promoters, CpG islands, and genic regions. Unsatisfied by these current methods, we devised a technique combining precipitation of densely methylated genomic DNA by recombinant MBD with massively parallel sequencing of captured fragments to efficiently map DNA methylation patterns in colon cancer cells.

We observed that an average of 38% of genomic DNA methylation occurs in repeat elements and centromeric/sub-telomeric regions. This is consistent with the notion that DNA methylation is important for condensing these genomic regions to provide structure and protect against unwanted transcription. Furthermore, a large fraction of uniquely mapped DNA methylation resides outside of gene promoters (>88%) and CpG islands (>55%). These would likely be missed by most custom arrays, which typically focus on CpG islands and/or gene promoters. The inclusion of DNA methylation data in all genomic contexts is another advantage of MiGS over existing technologies.

Furthermore, by comparing these genome-wide DNA methylation patterns to gene expressions, CTCF binding sites, and miRNA coding regions, we were able to assign functional roles for the detected DNA methylation. We observed that gene promoter methylation robustly correlates with transcriptional silencing, regardless of whether the sequence qualifies as a CpG island. We also found that non-promoter DNA methylation overlaps with CTCF binding sites and miRNA coding regions and therefore, is likely to be biologically relevant. These information would, again, be missed by many array-based detection methods.

With only two lanes of sequencing per sample, we detected both known DNA methylation and identified numerous novel methylated regions in these cells with low false positive and false negative rates. Although additional sequencing could identify more methylated loci, >95% of the highly methylated loci, which include the majority of genes previously described, can be captured with merely two lanes of GAII sequencing. By comparison, more than 100 lanes of sequencing on the same instrument are required to obtain the same information using bisulfite sequencing. One disadvantage of MiGS, as compared to bisulfite sequencing, is that the description of methylation is not at single base pair resolution; rather, the resolution depends on the fragment size of the sonicated DNA. Thus, fragments determined to be methylated by MiGS can contain individual CpG sites that are not methylated. However, MiGS can focus bisulfite sequencing efforts to fragments that are significantly methylated. By doing so, the cost of analyzing methylation on a genome-wide scale, even at single base pair resolution, is substantially reduced. Our data support that MiGS is a sensitive, specific, thorough, and cost-effective method for studying genome-wide DNA methylation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank J. McPherson and T. Hudson for their help with GAII sequencing. We also thank W. G. Nelson, V, for his gift of the pFBC6H-MBD2_MBD plasmid used for expressing the recombinant MBD protein and B. Vogelstein for DICER^{ex5} and DKO cells.

FUNDING

Funding for open access charge: Cleveland Clinic Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Yen,R.W., Vertino,P.M., Nelkin,B.D., Yu,J.J., el-Deiry,W., Cumaraswamy,A., Lennon,G.G., Trask,B.J., Celano,P. and Baylin,S.B. (1992) Isolation and characterization of the cDNA encoding human DNA methyltransferase. *Nucleic Acids Res.*, **20**, 2287–2291.
2. Gruenbaum,Y., Cedar,H. and Razin,A. (1982) Substrate and sequence specificity of a eukaryotic DNA methylase. *Nature*, **295**, 620–622.
3. Bird,A. (2007) Perceptions of epigenetics. *Nature*, **447**, 396–398.
4. Holliday,R. and Pugh,J.E. (1975) DNA modification mechanisms and gene activity during development. *Science*, **187**, 226–232.
5. Cedar,H., Stein,R., Gruenbaum,Y., Naveh-Many,T., Sciaky-Gallili,N. and Razin,A. (1983) Effect of DNA methylation on gene expression. *Cold Spring Harb. Symp. Quant. Biol.*, **47 (Pt 2)**, 605–609.
6. Reik,W., Collick,A., Norris,M.L., Barton,S.C. and Surani,M.A. (1987) Genomic imprinting determines methylation of parental alleles in transgenic mice. *Nature*, **328**, 248–251.
7. Riggs,A.D. (1975) X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.*, **14**, 9–25.
8. Feinberg,A.P. and Vogelstein,B. (1987) Alterations in DNA methylation in human colon neoplasia. *Semin. Surg. Oncol.*, **3**, 149–151.
9. Spruck,C.H. III, Rideout,W.M. III and Jones,P.A. (1993) DNA methylation and cancer. *[Review]. EXS.*, **64**, 487–509.
10. Virmani,A.K., Rathi,A., Sathyanarayana,U.G., Padar,A., Huang,C.X., Cunnigham,H.T., Farinas,A.J., Milchgrub,S., Euhus,D.M., Gilcrease,M. *et al.* (2001) Aberrant methylation of the adenomatous polyposis coli (APC) gene promoter 1A in breast and lung carcinomas. *Clin. Cancer Res.*, **7**, 1998–2004.
11. Tsuchiya,T., Tamura,G., Sato,K., Endoh,Y., Sakata,K., Jin,Z., Motoyama,T., Usuba,O., Kimura,W., Nishizuka,S. *et al.* (2000) Distinct methylation patterns of two APC gene promoters in normal and cancerous gastric epithelia. *Oncogene*, **19**, 3642–3646.
12. Ibanez de Caceres,I., Battagli,C., Esteller,M., Herman,J.G., Dulaimi,E., Edelson,M.I., Bergman,C., Ehya,H., Eisenberg,B.L. and Cairns,P. (2004) Tumor cell-specific BRCA1 and RASSF1A hypermethylation in serum, plasma, and peritoneal fluid from ovarian cancer patients. *Cancer Res.*, **64**, 6476–6481.
13. Rice,J.C., Massey-Brown,K.S. and Futscher,B.W. (1998) Aberrant methylation of the BRCA1 CpG island promoter is associated with decreased BRCA1 mRNA in sporadic breast cancer cells. *Oncogene*, **17**, 1807–1812.
14. Bian,Y.S., Osterheld,M.C., Fontolliet,C., Bosman,F.T. and Benhattar,J. (2002) p16 inactivation by methylation of the CDKN2A promoter occurs early during neoplastic progression in Barrett's esophagus. *Gastroenterology*, **122**, 1113–1121.
15. Holst,C.R., Nuovo,G.J., Esteller,M., Chew,K., Baylin,S.B., Herman,J.G. and Tlsty,T.D. (2003) Methylation of p16(INK4a)

promoters occurs in vivo in histologically normal human mammary epithelia. *Cancer Res.*, **63**, 1596–1601.

16. Frommer,M., McDonald,L.E., Millar,D.S., Collis,C.M., Watt,F., Grigg,G.W., Molloy,P.L. and Paul,C.L. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl Acad. Sci. USA*, **89**, 1827–1831.
17. Hatada,I., Fukasawa,M., Kimura,M., Morita,S., Yamada,K., Yoshikawa,T., Yamanaka,S., Endo,C., Sakurada,A., Sato,M. *et al.* (2006) Genome-wide profiling of promoter methylation in human. *Oncogene*, **25**, 3059–3064.
18. Jacinto,F.V., Ballestar,E. and Esteller,M. (2008) Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. *Biotechniques*, **44**, 35, 37, 39 passim.
19. Rauch,T.A. and Pfeifer,G.P. (2009) The MIRA method for DNA methylation analysis. *Methods Mol. Biol.*, **507**, 65–75.
20. Irizarry,R.A., 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.
21. Barr,H., Hermann,A., Berger,J., Tsai,H.H., Adie,K., Prokhortchouk,A., Hendrich,B. and Bird,A. (2007) Mbd2 contributes to DNA methylation-directed repression of the Xist gene. *Mol. Cell Biol.*, **27**, 3750–3757.
22. Yegnasubramanian,S., Lin,X., Haffner,M.C., DeMarzo,A.M. and Nelson,W.G. (2006) Combination of methylated-DNA precipitation and methylation-sensitive restriction enzymes (COMPARE-MS) for the rapid, sensitive and quantitative detection of DNA methylation. *Nucleic Acids Res.*, **34**, e19.
23. Ting,A.H., Suzuki,H., Cope,L., Schuebel,K.E., Lee,B.H., Toyota,M., Imai,K., Shinomura,Y., Tokino,T. and Baylin,S.B. (2008) A requirement for DICER to maintain full promoter CpG island hypermethylation in human cancer cells. *Cancer Res.*, **68**, 2570–2575.
24. Rhee,I., Bachman,K.E., Park,B.H., Jair,K.W., Yen,R.W., Schuebel,K.E., Cui,H., Feinberg,A.P., Lengauer,C., Kinzler,K.W. *et al.* (2002) DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature*, **416**, 552–556.
25. Cummins,J.M., He,Y., Leary,R.J., Pagliarini,R., Diaz,L.A. Jr, Sjoblom,T., Barad,O., Bentwich,Z., Szafranska,A.E., Labourier,E. *et al.* (2006) The colorectal microRNAome. *Proc. Natl Acad. Sci. USA*, **103**, 3687–3692.
26. Lee,B.H., Yegnasubramanian,S., Lin,X. and Nelson,W.G. (2005) Procainamide is a specific inhibitor of DNA methyltransferase 1. *J. Biol. Chem.*, **280**, 40749–40756.
27. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
28. Levy,S., Sutton,G., Ng,P.C., Feuk,L., Halpern,A.L., Walenz,B.P., Axelrod,N., Huang,J., Kirkness,E.F., Denisov,G. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.
29. Smit,A.F.A., Hubley,R. and Green,P. Repeat Masker Open 3.0 <http://www.repeatmasker.org>. (1996–2006).
30. Schuebel,K.E., Chen,W., Cope,L., Glockner,S.C., Suzuki,H., Yi,J.M., Chan,T.A., Van Neste,L., Van Criekinge,W., van den Bosch,S. *et al.* (2007) Comparing the DNA hypermethylome with gene mutations in human colorectal cancer. *PLoS Genet.*, **3**, 1709–1723.
31. Bird,A. (1992) The essentials of DNA methylation. *Cell*, **70**, 5–8.
32. Bell,A.C. and Felsenfeld,G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **405**, 482–485.
33. Lujambio,A., Calin,G.A., Villanueva,A., Ropero,S., Sanchez-Cespedes,M., Blanco,D., Montuenga,L.M., Rossi,S., Nicoloso,M.S., Faller,W.J. *et al.* (2008) A microRNA DNA methylation signature for human cancer metastasis. *Proc. Natl Acad. Sci. USA*, **105**, 13556–13561.