nature biotechnology

Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells

Madeleine P Ball[1,2,6], Jin Billy Li[1,2,6], Yuan Gao[3], Je-Hyuk Lee[1,2], Emily M LeProust[4], In-Hyun Park[5], Bin Xie[3], George Q Daley[5] & George M Church[1,2]

Studies of epigenetic modifications would benefit from improved methods for high-throughput methylation profiling. We introduce two complementary approaches that use next-generation sequencing technology to detect cytosine methylation. In the first method, we designed ~10,000 bisulfite padlock probes to profile ~7,000 CpG locations distributed over the ENCODE pilot project regions and applied them to human B-lymphocytes, fibroblasts and induced pluripotent stem cells. This unbiased choice of targets takes advantage of existing expression and chromatin immunoprecipitation data and enabled us to observe a pattern of low promoter methylation and high gene-body methylation in highly expressed genes. The second method, methyl-sensitive cut counting, generated nontargeted genome-scale data for ~1.4 million *Hpa*II sites in the DNA of B-lymphocytes and confirmed that gene-body methylation in highly expressed genes is a consistent phenomenon throughout the human genome. Our observations highlight the usefulness of techniques that are not inherently or intentionally biased towards particular subsets like CpG islands or promoter regions.

Methylation of cytosine at CpG dinucleotides is an important regulatory modification in many eukaryotic genomes[1]. CpG methylation can be inherited through "maintenance methylation" of hemimethylated sites in newly synthesized DNA and plays a role in gene transcription, embryogenesis and diseases such as cancer[2–4]. It is generally studied using one of three techniques. The first, bisulfite sequencing, converts all unmethylated cytosines to uracil, which is subsequently recognized as thymine[5]. Although it has emerged as the 'gold standard' for methylation profiling, bisulfite sequencing is complicated by the decrease in sequence specificity associated with conversion of most cytosines in the genome. The second approach involves restriction enzymes that preferentially cut DNA based on its methylation status, typically when the recognition site is unmethylated[6]. Although robust, this method is limited to profiling the enzyme's recognition sites. Finally, some recent studies have used antibody-based affinity purification to pull down methylated DNA[7–9]. As this approach is based on the methylcytosine density in a region, it is more effective for profiling regions with a higher density of potentially methylated CpG sites. Strategies involving bisulfite sequencing, methylation-sensitive restriction enzymes and affinity purification have all been combined with microarrays[7,9–15] or high-throughput sequencing[16–18] to create high-throughput methylation profiles. But most of these approaches have been limited by choice or by technology to CpG islands and/or gene promoters.

Although whole genome bisulfite sequencing has been used for genome-wide analysis of methylation in *Arabidopsis thaliana*[17,18], it remains prohibitively expensive for the much larger human genome. We therefore developed two complementary technologies, both involving massively parallel sequencing[19,20], for high-throughput profiling of cytosine methylation and applied them to eight human cell lines. We used the first, a targeted approach involving padlock (molecular inversion) probes designed to recognize locations in bisulfite-treated genomic DNA, to specifically capture and profile ~7,000 CpG sites in the ENCODE pilot project regions[21]. The second method, which uses the methylation sensitive enzyme *Hpa*II and is thus called methyl-sensitive cut counting (MSCC), was used to create a genome-scale methylation profile of a single cell line. MSCC uses a library derived from all locations cut by *Hpa*II to profile the methylation of ~1.4 million unique sites in the human genome.

Our methylation profiles consistently reveal a pattern of gene-body methylation in the highly expressed genes of human cell lines. This builds on growing evidence for gene-body methylation in mammals[9,10,22] and shows it to be a general feature of all highly expressed genes in human cell lines. Our results also support prior observations that genes with promoters containing an intermediate level of CpG density have the highest expression-related differences in promoter methylation.

RESULTS

Bisulfite padlock probe design, synthesis and processing

Our first technology, involving bisulfite padlock probes (BSPPs), is a targeted method that isolates selected locations for methylation profiling. Padlock probes are ~100-nucleotide DNA fragments

[1]Department of Genetics, Harvard Medical School, [2]Broad Institute of MIT and Harvard University, Cambridge, Massachusetts, USA. [3]Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia, USA. [4]Genomics Solution Unit, Agilent Technologies Inc., Santa Clara, California, USA. [5]Department of Medicine, Division of Pediatric Hematology Oncology, Children's Hospital Boston, and Dana-Farber Cancer Institute; Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Karp Family Research Building 7214, Boston, Massachusetts, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to G.M.C. [redacted] or J.B.L. (jli@genetics.med.harvard.edu).

Received 4 February; accepted 6 March; published online 29 March 2009; doi:10.1038/nbt.1533

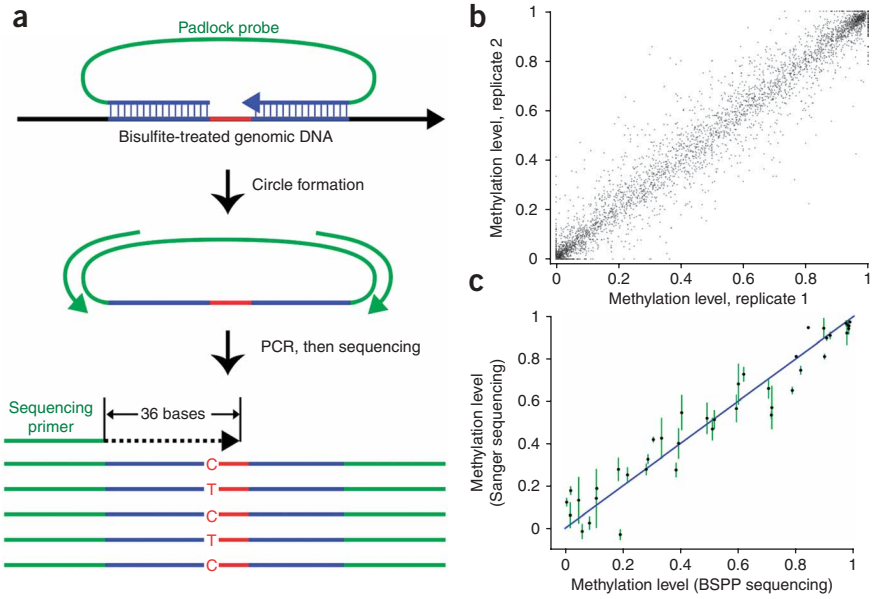

Figure 1 BSPP technology enables accurate measurement of methylation levels. (**a**) BSPP experimental scheme. Two hybridizing locus-specific arms (blue) are connected by a 50-bp common backbone sequence (green). We designed ~10,000 BSPPs to target CpG sites in bisulfite-treated DNA, with a CpG located at the 3′ end of the 10 bp polymerized span (red). Circles were formed by adding DNA polymerase, dNTPs and ligase, and were subsequently amplified using the backbone sequence for priming. Sequencing was then performed using an Illumina Genome Analyzer with a primer matching the backbone sequence; 28 bases of arm sequence were read before sequencing informative positions within the span (read lengths were 36 bases in total). (**b**) Correlation of methylation level in the technical replicates (Pearson coefficient $r = 0.965$). (**c**) Correlation of BSPP methylation with the methylation levels determined by bisulfite PCR followed by Sanger sequencing at 33 locations ($r = 0.966$). Error bars (green) represent the s.d. of methylation, as measured by Sanger sequencing.

designed to hybridize to genomic DNA targets in a horseshoe manner (**Fig. 1a**)[23–25]. After the gap between the two hybridized, locus-specific arms of a padlock probe is polymerized and ligated to form a circular strand of DNA, the circles generated can then be amplified using the common 'backbone' sequence that connects the two arms. This enables tens of thousands of probes to be used within a single reaction. The resulting libraries are then analyzed using massively parallel sequencing. We have used padlock probes to specifically amplify ~10,000 human exons[25], and have recently improved capturing efficiency >10,000-fold (Li *et al.*, unpublished data).

To apply padlock probes to profiling DNA methylation, we designed a probe set to target ~10,000 locations in bisulfite-treated human genomic DNA (**Supplementary Table 1** online). Bisulfite treatment converts all unmethylated cytosines to uracils, which are recognized as thymines[5]. To simplify the construction of sequencing libraries, probes were designed to target 10-nucleotide regions containing at least one CpG (**Fig. 1a**). Nonetheless, it should be possible to also capture larger regions[25]. Hybridizing arms flanking this span were designed to avoid CpGs because their methylation status (and sequence after bisulfite treatment) is unknown. This simplified the design, although it should also be possible to design probes that match all potential variants[26]. To avoid targeting unconverted DNA, one of the arms was required to have at least three non-CpG cytosines in the fifteen nucleotides closest to the span.

Probes were chosen from within the ENCODE pilot project regions, which represent ~1% of the human genome and have been characterized by expression and chromatin immunoprecipitation (ChIP) analysis[21]. Rather than targeting promoter regions or CpG islands, we chose ~10,000 probes that best satisfied our design criteria scattered over all regions (**Supplementary Table 2** online). Because we avoided CpGs in the hybridizing arms, these probes were actually biased against targeting CpG islands that are favored by most of previous methylation studies.

The probes, flanked with common primer sequences ('probe precursors'), were synthesized on a programmable microarray and cleaved into a single tube, as described previously[25]. After PCR amplification, the probe precursors were subject to enzymatic processing to trim both primer ends.

Performance of the BSPP assay

Our initial experiment used the BSPP set to investigate cytosine methylation in the Epstein-Barr virus (EBV)-transformed GM06990 B-lymphocyte cell line, also used in the ENCODE project[21]. The pool of ~10,000 BSPPs was hybridized with the bisulfite-converted genomic DNA of GM06990 in a single reaction. After the circles were formed and subsequently amplified (**Fig. 1a**), we gel-extracted DNA of the expected size (**Supplementary Fig. 1a** online) and cloned and sequenced individual library molecules to check the specificity of capture. Of the 79 isolates tested, 78 (99%) mapped to the intended target sites and were unique, illustrating the high specificity of padlock-probe technology despite the reduced genomic complexity after bisulfite conversion. We performed two technical replicates of target capture followed by Illumina Genome Analyzer (formerly Solexa) sequencing. Although the number of probe observations varied widely, ~7,700 (80%) and ~6,400 (68%) sites were covered with at least 1 and 10 reads, respectively, when ~3 million reads were derived from a single sequencing lane (**Supplementary Fig. 1b** and **Supplementary Table 3** online). Nevertheless, both the numbers of probe observations (**Supplementary Fig. 1c**) and the inferred methylation levels (**Fig. 1b**) were highly correlated. Because of this, if reduced variance is desired, probes can be empirically divided into separate pools, depending on their efficiencies[26]. To rule out the possibility of systemic bias, we performed traditional Sanger sequencing on 33 regions amplified from bisulfite-treated DNA (**Supplementary Table 4** online); the methylation levels determined by this method correlated well with the BSPP-determined methylation ($r = 0.966$; **Fig. 1c**).

Methylation levels were bimodally distributed, with most sites <20% or >80% methylated (**Supplementary Fig. 2** online). This is consistent with previous reports[27]. As CpGs in close proximity are known to be often co-methylated[27], we took advantage of the clonal feature of Illumina sequencing to investigate whether co-methylation occurs at the single-molecule level. Within probes spanning more than one CpG, sites with intermediate methylation levels (between 20% and 80%) displayed a positive correlation with the methylation states of neighboring CpGs on individual strands (**Supplementary Fig. 3** online).

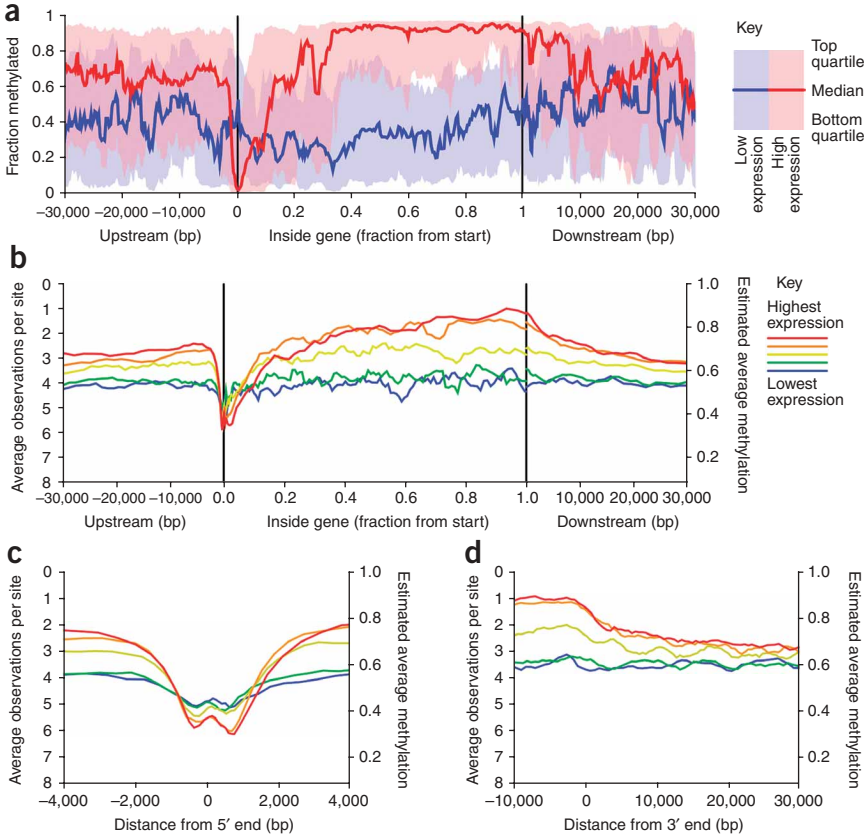

Figure 2 Changes in methylation associated with position relative to genes differ between highly and weakly expressed genes. (**a**,**b**) A phenomenon of gene-body methylation is seen in both BSPP (**a**) and MSCC (**b**) data. Running median levels of BSPP-determined methylation versus gene position show a difference between highly and weakly expressed genes in the ENCODE pilot regions of the GM06990 cell line. Running averages of MSCC *Hpa*II counts versus gene position for all genes in the PGP1 lymphoblast cell line show differences between different gene expression levels. Genes were categorized into five equally sized groups based on expression level. The contribution of each MSCC data point was normalized for local CpG density, *Msp*I control counts and, for sites within the gene, for gene length. (**c**) Expression-related differences are seen in the running average of MSCC *Hpa*II counts versus distance at the transcription start site. (**d**) Expression-related differences are also seen in the running average of MSCC *Hpa*II counts versus distance from the transcriptional ends of genes.

Methylation versus transcription and histone modification

To explore the relationship between gene expression and methylation in the promoter and elsewhere in the gene, we used ENCODE project gene expression data for the same cell line (GM06990) to split genes into two equal groups: "highly expressed" and "weakly expressed" genes. For each group we plotted median cytosine methylation against gene position (**Fig. 2a**). In the highly expressed genes, we saw a pattern of low methylation in the promoter region and considerable methylation in the rest of the gene body. The weakly expressed genes were moderately methylated in both promoter and gene-body regions.

Cytosine methylation may interact with other epigenetic features, such as histone modifications. To look for correlations between DNA methylation and histone modification, we compared available ChIP data[21] with our methylation data obtained from the same cell line. Cytosine methylation was correlated with H3K36 methylation and anticorrelated with H3K27 methylation (**Supplementary Fig. 4** online). These correlations probably reflect the distribution of our probes, half of which target regions within gene bodies, whereas only 5% target regions within 1 kb of transcription start sites. The correlations are consistent with the gene-body pattern of the histone modifications: H3K36 methylation is higher in gene bodies of highly expressed genes, whereas H3K27 is high in gene bodies of weakly expressed genes[28].

BSPP profiling of cell lines from the Personal Genome Project

Our methylation profiling methods have, in part, been developed as a pilot for studying epigenomics within the context of the Personal Genome Project (PGP), a program through which researchers will deeply explore the relationship between genotype and phenotype through collection of multi-faceted biological information from individuals registered within the project[29]. To explore how methylation patterns vary between different cell types and different individuals, we applied the ENCODE BSPP set to several cell lines from the PGP: PGP1 and PGP9 EBV-transformed B-lymphocytes, PGP1 and PGP9 fibroblasts, and induced pluripotent stem (iPS) cells derived from PGP1 and PGP9 fibroblasts. Consistent with previous studies[27], the methylation patterns of lymphoblast lines derived from different individuals were highly correlated ($r = 0.85$, **Supplementary Fig. 5a** online), whereas the correlation between fibroblast and lymphoblast cells from the same individual was much lower ($r = 0.63$, **Supplementary Fig. 5b**). The PGP1 and two independent PGP9-derived iPS cell lines were hypermethylated in the ENCODE regions of ~400 genes, compared to the fibroblast line from which they were derived (**Supplementary Figs. 2f–h** and **5c**,**d**). This may be a general phenomenon, although because we surveyed a limited set of locations and cell culturing can affect global methylation levels[16], further investigation is warranted. The phenomenon of gene-body methylation in highly expressed genes was also observed in the PGP lymphoblast and fibroblast cell lines (**Supplementary Fig. 6** online).

MSCC assay

In contrast with the BSPP approach, our second technology, MSCC, profiles methylation across the whole genome. MSCC queries the sensitivity of all CCGG sites within the genome to *Hpa*II, a methylation-sensitive restriction enzyme that cuts unmethylated CCGG sequences. Methylation-sensitive restriction enzymes typically have a recognition site that contains a CpG dinucleotide and are blocked from cutting if that site is methylated[6]. The MSCC assay is not limited to using *Hpa*II and could be used with other methylation-sensitive restriction enzymes to profile other, nonoverlapping genome-scale sets of CpGs (**Supplementary Table 5** online). These sets could be combined to created denser genome-scale profiles.

With MSCC, no choice is made for which sites are targeted—all uniquely identifiable *Hpa*II sites are profiled. *Hpa*II sites have a distribution similar to the distribution of all CpG dinucleotides (**Supplementary Table 2**), making them a good target for relatively unbiased genome-scale profiling. By generating a library of tag

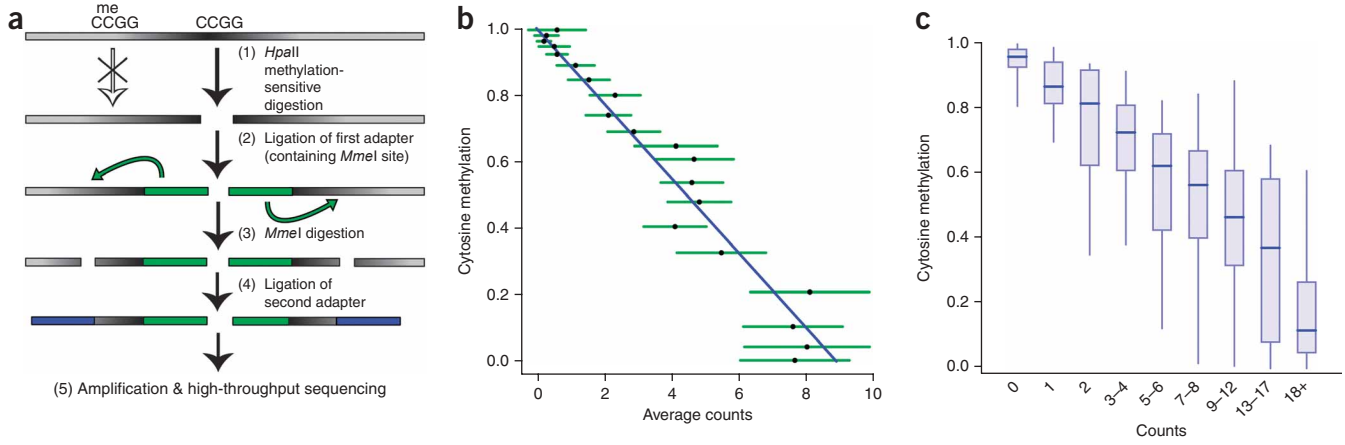

Figure 3 MSCC technology enables accurate estimates of methylation. (**a**) Scheme to generate a MSCC library. (1) After *Hpa*II digestion cuts genomic DNA specifically at all unmethylated CCGG sites, (2) the first adaptor containing an *Mme*I recognition site is ligated, (3) *Mme*I digestion cuts within the unknown genomic sequence to produce an 18–19 bp tag, (4) a second adaptor is added by ligation and (5) the library is amplified and sequenced. The number of reads for a given site is correlated with the amount of digestion that occurs there and thus is an indication of methylation level. (**b**) BSPP methylation versus MSCC counts. Data were grouped according to the BSPP-determined methylation levels into 20 bins, with each bin containing an equal number of data points. The mean number of counts (black points) is linearly related to the mean methylation of a bin (best fit line is shown in blue). Green error bars represent the 95% confidence interval based on the s.e.m. for that bin. (**c**) Binned total MSCC counts from paired tag sites show how well individual sites predict methylation. Horizontal bars represent median methylation as determined by BSPP, boxes represent the quartiles and whiskers mark the 5th and 95th percentiles.

fragments from all cut locations and then using massively parallel sequencing to gather millions of observations of these, we can infer the methylation level by the number of times a site is observed (**Fig. 3a**). Sites with many and no reads are inferred to have low and high methylation levels, respectively. A control library was also constructed by replacing *Hpa*II with a methylation-insensitive isoschizomer, *Msp*I. However, sequencing an *Msp*I control is a large additional cost and our data indicate that results from the *Hpa*II library alone are highly correlated with methylation at individual sites.

The human genome contains 2.3 million *Hpa*II sites and each of these, if cut, can generate two possible library tags. Of the ~4.6 million possible tag sequences, we considered about half (~2.3 million) to be sufficiently unique for use in profiling: they have more than a single one-nucleotide difference when compared with any other possible sequence. These tags combine to cover a total of 1,417,432 individual CpG sites (**Supplementary Table 6** online). Of these, 888,455 (63%) produce two unique tags ('paired tags') and 528,977 (37%) produce a single unique tag. These combine to a total of 1,417,432 CpG sites that are profiled with this method (**Supplementary Table 6** online). Nearly half of these sites occur within genes (>18,000 genes have at least one site within them), 3.4% are within 1 kb of the transcription start site (>10,000 genes have at least one site in this region) and 13.5% are within CpG islands (90% of CpG islands have at least one site within them) (**Supplementary Table 2**).

MSCC profiling accurately measures methylation

We produced an MSCC *Hpa*II library and *Msp*I control library for the PGP1 EBV-transformed B-lymphocyte cell line, for which we also had BSPP and gene expression data. Libraries were sequenced using an Illumina Genome Analyzer and matched to a list of all possible tag sequences (**Supplementary Table 7** online). We performed two technical replicates of the *Hpa*II library that, although subject to variance according to the Poisson distribution, showed a high correlation in the number of observations for each site ($r = 0.82$, **Supplementary Fig. 7** online). The availability of BSPP data for the same sample enabled us to compare the methylation levels determined by BSPPs with MSCC *Hpa*II data for 381 sites (726 individual tags) (**Supplementary Fig. 8** online). When data are binned according to the BSPP-determined methylation levels, the average number of

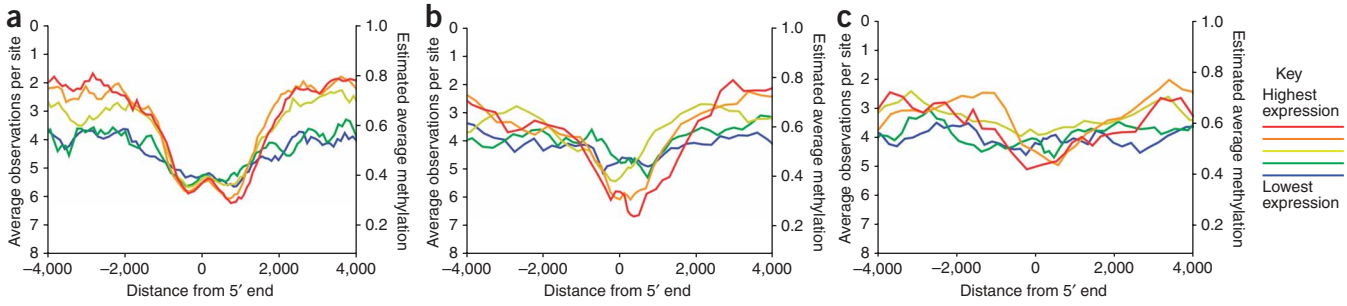

Figure 4 The effects of promoter CpG density and methylation in genes with different levels of expression. (**a**) High CpG promoters (65% of all promoters) tend to have low methylation regardless of expression. (**b**) Intermediate CpG promoters (16% of promoters) tend to have low levels of methylation in highly expressed genes and high levels of methylation in weakly expressed genes. (**c**) Low CpG promoters (28% of promoters) tend to be highly methylated regardless of gene expression.

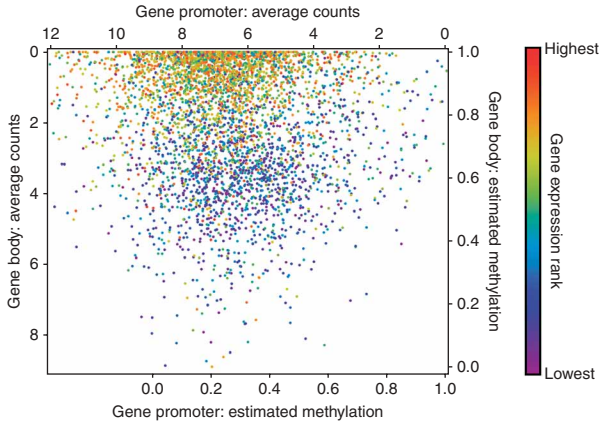

Figure 5 Methylation profiles of individual genes. Individual genes are plotted according to the average number of MSCC *Hpa*II counts found in the promoters (horizontal axis, –400 to +1,000 relative to start) and gene bodies (vertical axis, between the gene end and +2,000 relative to start). The color of each point reflects the expression level of that gene. Points were plotted in a random order to avoid artifacts produced by nonrandom overlaps. Only genes with at least ten data points in each region were used in the analysis.

counts for each bin is linearly related to its methylation level (**Fig. 3b**). We used this to estimate average methylation levels when counts for multiple sites are averaged. BSPP methylation data can also be used to estimate methylation levels for individual sites based on MSCC *Hpa*II counts (**Fig. 3c** and **Supplementary Fig. 9** online).

MSCC counts have more noise for more highly digested *Hpa*II sites (that is, sites with lower methylation). As a result, MSCC is more accurate at distinguishing moderately methylated sites from highly methylated sites than it is for distinguishing moderately from weakly methylated sites, although deeper sequencing coverage should improve accuracy (**Supplementary Table 8** online). In addition, preliminary data suggest that the accuracy can be improved by sequencing an 'inverse library' of methylated CCGG sites, which is constructed by dephosphorylating *Hpa*II-digested fragment ends, digesting with *Msp*I and then ligating an *Mme*I-containing adaptor to generate sequencing tags (**Supplementary Fig. 10** online). In the following analyses, however, we used only the MSCC *Hpa*II data generated from three lanes of Illumina sequencing.

Comparison of MSCC methylation with gene expression

Compared to BSPP, which analyzed several thousand data points covering ~400 genes, the MSCC technology covered 1.4 million sites in >18,000 genes distributed over the entire genome. This allowed us to examine the relationship between gene expression and cytosine methylation more thoroughly. We split genes into five equal groups based on their expression levels and plotted the running average of MSCC observations versus gene position for each (**Figs. 2b–d**). We observed a similar pattern of low promoter methylation and considerable gene-body methylation in highly expressed genes as we did in the BSPP assays (**Fig. 2b** and **Supplementary Fig. 6a**). Previous studies have indicated that gene expression may require low promoter methylation extending several hundred bases into the gene[30]. Consistent with this, we observed that highly expressed genes have low methylation, extending to around +1 kb downstream of the transcription start site with a valley at around 600–700 bp (**Fig. 2c**). Our data also show another valley upstream of the transcription start site. The pair of valleys is similar to the double peaks observed in H3K4 methylation and other histone modifications[28] and may be related to recent findings of bidirectional transcription at gene promoters[31]. At the 3′ end, highly expressed genes appear to have increased methylation running up to and dropping off after the end of the gene (**Fig. 2d**).

Previous experiments have indicated that the relationship between promoter methylation and gene expression is related to the CpG density of the promoter[32]. To examine this, we divided promoters into three types according to CpG content: high CpG promoters, low CpG promoters and intermediate CpG promoters. Our results, which are consistent with previous observations, provide quantitative averaged profiles of methylation versus position for each promoter type (**Fig. 4**). Although subtle expression-related differences exist, on average, high CpG promoters have less methylation (**Fig. 4a**), whereas low CpG promoters display more methylation (**Fig. 4c**), regardless of gene expression level. Our results also show that the largest expression-related differences in promoter methylation are found in intermediate CpG promoters (**Fig. 4b**).

To explore how methylation information was correlated with gene expression on the level of individual genes, we compared the gene promoter methylation and gene-body methylation of individual genes. According to these two metrics, genes formed two clusters that corresponded to high and low expression levels (**Fig. 5**). This figure shows that the average gene-body methylation differences observed between highly expressed and weakly expressed genes reflect a consistent phenomenon rather than a subset of genes containing hypermethylated or hypomethylated gene bodies.

DISCUSSION

The rapid development of cheaper, massively parallel sequencing technologies[33] is opening the way for new strategies to study biologically important processes[34–36], including the regulatory roles of epigenetic modifications like DNA methylation[16–18]. These methods are contributing to the emergence of high-throughput sequencing as a ubiquitous convergent platform. The digital aspect of sequencing makes techniques inherently more quantitative, accurate and reproducible than more traditional molecular techniques. BSPP and MSCC are two complementary methods that take advantage of the power of new sequencing technologies to profile cytosine methylation at single-base resolution in targeted and genome-scale surveys, respectively.

In contrast to other restriction enzyme–based profiling methods[10,11,15], MSCC does not involve selection of fragments within a specified size range. As a consequence, by design, its efficiency and accuracy for profiling each site is not influenced by local sequence characteristics, such as recognition-site density. MSCC also takes advantage of high-throughput sequencing technology, which is rapidly becoming cheaper, and becomes more accurate when more sequencing reads are obtained. Although reduced representation bisulfite sequencing (RRBS) also involves massively parallel sequencing to profile methylation at a subset of sites[16], it cannot be designed to target specific segments (as BSPP can) and is biased by profiling a scattered set of genomic segments with high CpG density. In contrast, MSCC is less biased, as *Hpa*II sites are more evenly distributed throughout the genome.

Gene-body methylation has been observed in *Arabidopsis thaliana*[12,13,17,18], where it is associated with active genes. Methylation in the gene bodies of certain mammalian genes has long been known to be positively correlated with elevated expression[37], and there is now growing evidence that this may be a general phenomenon. Gene-body methylation has been observed in the active human X chromosome when compared to its inactive counterpart[10], hypomethylated sites in the gene body have been associated with weakly expressed genes in

cancer cell lines[22], and methylation of CpG-rich sites in gene bodies has been associated with highly expressed genes in human B cells[9]. A general phenomenon of gene-body methylation in highly expressed genes is strongly supported by our data obtained using both the BSPP and MSCC assays. Gene-body methylation has been hypothesized to suppress spurious initiation of transcription within active genes in *Arabidopsis*[12,13] and a similar function may exist in mammals[1].

CpG islands and promoters have been the preferred target of many studies and have, in the past, guided the design of many methylation profiling experiments[14,16,32,38]. Our observations concerning gene-body methylation, differential methylation in intermediate CpG promoters[32], and other evidence for differential methylation in regions outside CpG islands and promoters[15], underscore the importance of less biased profiling methods in uncovering aspects of methylation that might otherwise have been missed. As DNA sequencing costs drop, tools like BSPP and MSCC could soon be broadly applied to study the epigenomic changes associated with developmental shifts, environmental changes, and disease states.

METHODS

Cell lines, RNA and genomic DNA, expression profiling, and bisulfite treatment. Genomic DNA of GM06990 (a HapMap/ENCODE sample) was obtained from Coriell Cell Repository. With the approval of Harvard Medical School's Institutional Review Boards, blood and skin biopsies were obtained from donors of the Personal Genome Project. The EBV-transformed B-lymphocyte cell lines and the derivative genomic DNA for donors PGP1 (GM20431) and PGP9 (GM21833) were generated and acquired from Coriell Cell Repository. Genomic DNA obtained directly from Coriell was used for methylation analysis of these lines, cultured cell lines were used for gene expression profiling. The primary fibroblast lines for PGP1 and PGP9 were generated by and obtained from Brigham Women's Hospital. The cultured cell line was used for both genomic DNA and gene expression profiling.

The PGP1 iPS line and two PGP9 iPS cell lines were derived by infecting primary human fibroblasts of PGP1 and PGP9 with highly concentrated retroviral OCT3, KLF4, SOX2 and c-MYC particles[39]. The infected cells were trypsinized onto a feeder layer after 4 d and maintained in hES medium (KO-DMEM (Invitrogen), 20% KO-SR (Invitrogen), 1× l-glutamine (Gibco), 1× MEM NEAA (Gibco), 1× penicillin/streptomycin (Gibco), 55 μM mercaptoethanol and 10 ng/ml bFGF). The iPS colonies were identified by their characteristic morphology after 3–4 weeks.

Immortalized lymphocytes were cultured in RPMI-1640 medium (Invitrogen) with 10% FBS (Invitrogen) and 2 mM L-glutamine. Primary fibroblasts were cultured in DMEM/F12 medium (Invitrogen) with 15% FBS and 10 ng/μl EGF. Human iPS cell lines were grown on a feeder layer of mouse embryonic fibroblasts (Global Stem) in hES media, and mechanically separated from mouse cells before DNA/RNA extraction.

Genomic DNAs and total RNAs were extracted with AllPrep DNA/RNA/Protein Mini Kit (Qiagen). RNA gene expression profiling was done using Illumina's bead array technology through the service provided by Harvard Partner Center for Genetics and Genomics. Bisulfite treatment was performed using the EZ DNA Methylation-Gold Kit (Zymo Research). Typical yield was 50–75% after bisulfite conversion.

Bisulfite padlock probe design and synthesis. Files for genomic sequence for the ENCODE pilot project regions were obtained from UCSC. Potential locations were chosen from nonrepetitive sequences containing 10 bases with a 5′ CpG flanked by least 20 bases of CpG-free flanking sequence on each side. Flanking arm sequences were designed for either bisulfite-treated strand, up to 28 bases in length, avoiding CpGs and targeting a melting temperature range of 50–55 °C. The 'ligation arm' was required to contain at least three non-CpG cytosines, and a guanine content of at least 20% was required of both arms. Probes were then selected to optimize uniqueness measurements based on 15-mer frequencies and BLAST searches for near matches. To avoid self-hybridization, no overlap between probes was allowed in the final set. Final probe sequences were 106 bp in length: two arms 28 bp long (random sequence to 28 bp, if necessary) and a 50-bp common backbone sequence. The final set of 9,552 probe sequences and locations as well as number of observations and methylation estimates from each sample is provided in **Supplementary Table 1**.

Probes were synthesized using a programmable microarray (Agilent) as 150-bp oligos containing common end sequences. These were cleaved off and collected in a single tube with an estimated concentration of 0.18 fmol/species. To amplify, we took 1% of the oligos and performed real-time PCR, monitoring the amplicon in a 100-μl reaction assembled with Platinum *Taq* supermix, 50 pmol of each primer, and 0.5× SYBR green. One of the primers was designed to contain phosphorothioates between the first four 5′ bases and a 3′ uracil. The other primer contained the sequence GATC at the 3′ end. The PCR program was: 95 °C for 5 min, 15 cycles of 95 °C for 30 s/58 °C for 1 min/72 °C for 1 min, and finally 72 °C for 5 min. The PCR product was purified with Qiagen PCR purification kit and quantified. Using a 96-well plate, a total of 9.6 ml PCR reaction was set up with 25 fmol template along with Platinum *Taq* supermix, 4.8 nmol of each primer, and 0.5× SYBR Green. The same PCR program was used. PCR products were purified by phenol:chloroform followed by Qiagen PCR purification kit and a total of 37 μg of DNA was obtained.

The PCR product was split into eight reactions with 10 units of lambda exonuclease (New England Biolabs (NEB)) in 1× lambda exonuclease reaction buffer and incubated at 37 °C for 45 min then 75 °C for 15 min. After being purified with QiaQuick columns the single-stranded (ss)DNA was quantified with Nanodrop to be 33 ng/μl in 200 μl total. This was split into four tubes, each of which was assembled with 50 μl of ssDNA (33 ng/μl), 6 μl of 10× *Dpn*II reaction buffer, and 2 μl of 100 μM 'guide oligo' designed to hybridize to the 3′ end of the ssDNA and ending in GATCNN. The mixture was heated to 95 °C for 5 min, followed by a ramp to 60 °C at 0.1 °C/s, 60 °C for 10 min, then 37 °C for 1 min. Into each tube, 5 μl of *Dpn*II (10 units/μl) (NEB) and 5 μl of USER enzyme (1 unit/μl) (NEB) were added and these were incubated at 37 °C for 3 h. The final product was loaded into 6% TBE Urea precast polyacrylamide gels (Invitrogen) and the desired band was cut and purified. The final concentration of padlock probes was quantified on a gel to be 9 ng/μl, which is 257 nM (27 pM for each of 9,552 species).

CpG padlock capturing and sequencing library construction. We assembled 1 μg (~0.5 amol of haploid) bisulfite-treated genomic DNA in a 15 μl reaction with 1× Ampligase buffer and 33.5 ng (~1 pmol) of probes. The reaction was incubated at 95 °C for 10 min, ramped to and held at 64 °C for 5 h, 65 °C for 5 h, 60 °C for 24 h. At 60 °C we added the gap filling and sealing mix: 2 μl of Ampligase storage buffer containing 0.5 pmol of dNTPs, 2 units Taq Stoffel fragment (Applied Biosystems), and 2.5 units Ampligase (Epicenter). The reaction was then incubated at 60 °C for 2 h, then cycled five times with 95 °C for 2 min/60 °C for 5 h. The temperature was then lowered to 37 °C and 2 μl of exonuclease I (20 units/μl) (USB) and 2 μl exonuclease II (200 units/μl) (USB) were added. The reaction was incubated at 37 °C for 2 h followed by 94 °C for 5 min.

The circularized probes were amplified using primers matching the backbone sequences in two 100 μl reactions containing 10 μl of the above reaction product, 50 μl of 2× iQ SYBR Green supermix (Bio-Rad) and 40 pmol each primer. Real-time PCR was used to monitor the reaction, which used this program: 96 °C 3 min, 5 cycles of 96 °C for 15 s, 60 °C for 30 s, 72 °C for 30 s, then 13 cycles of 96 °C for 15 s, 72 °C for 1 min, 72 °C for 1 min, then 72 °C for 5 min. A 6% TBE polyacrylamide gel was used to purify the band containing the final library molecules.

BSPP library sequencing and analysis. Libraries were diluted to 10 nM and each was sequenced with one lane of an Illumina Genome Analyzer. Reads were matched with BLAST to a custom database containing the predicted reads, with CpG cytosines replaced with N, and accepted only if they had no mismatches in the 10-bp span (except the masked CpG cytosines) and not more than three mismatches elsewhere. Methylation was determined by the number of 'C' reads out of all reads for a given location.

To validate methylation levels determined by padlock probes we designed primers targeting 33 of the profiled locations in bisulfite-treated DNA, performed PCR amplification and Sanger sequencing of the PCR product. The methylation level of each site was determined using the ratio of T peak at the target location compared to neighboring non-CpG T peaks, with peak

height determined using PeakPicker software[40]. This is similar to the principle applied in the commercially available software ESME[41]. Because we performed multiple sequencing reactions and from both directions, multiple estimates were combined to get the average and s.d. values we plotted for each site.

RNA expression data were gathered using the ENCODE project PolyA+ RNA signal track downloaded from UCSC. Using scores for regions annotated as exons by RefGene, median values were taken to represent gene expression level. To construct average gene graphs, each methylation data point was assigned position information according to its location relative to nearby genes: a fractional value if within a gene, or bp if upstream or downstream. The running median and quartiles were plotted.

Histone modification data were acquired from Sanger ChIP data downloaded from UCSC. To look for correlations, raw ChIP scores versus methylation were plotted along with the running median and quartiles. Gene profiles of histone modifications were also created as done for methylation data.

MSCC library creation. Two custom adapters were created for MSCC, each composed of two oligonucleotides ordered from IDT. Adapter A contains an 5′ *Mme*I recognition site and 5′ CG overhang, adapter B contains a 3′ NN overhang.

To construct the MSCC *Hpa*II library, 2 µg of PGP1 lymphocyte genomic DNA was assembled into a 100-µl reaction with 20 units HpaII (NEB) in 1× NEBuffer 1, incubated at 37 °C for 2 h, then 65 °C for 20 min. To this was added 1.66 µl of 10 µM adaptor A, 12 µl 10 mM ATP and 120 units T4 DNA ligase (NEB). This was incubated at 16 °C for 4 h, then 65 °C for 15 min. Ethanol precipitation was performed and DNA was resuspended to 50 µl with a reaction mixture containing 8 units *Bst* DNA polymerase fragment (NEB), 200 µM dNTP and 1× thermopol buffer (NEB). This was incubated at 50 °C for 20 min, then 85 °C for 20 min. Ethanol precipitation was performed again, and the pellet was resuspended to 50 µl with a reaction mixture containing 2 units MmeI (NEB), 50 µM SAM and 1× NEBuffer 4. This was incubated at 37 °C for 2 h, then 80 °C for 20 min. To this was added 1.66 µl of 10 µM adaptor B, 6 µl 10 mM ATP and 3 µl T4 DNA ligase, and the mixture was incubated at 16 °C for 4 h, then 65 °C for 15 min.

The mixture was run on a 6% nondenaturing TBE polyacrylamide gel (Invitrogen) and the target band at ~140 bp was purified. PCR was then performed on ~80% of this purified sample using primers matching the sequences of adaptor A and adaptor B. The assembled mixture was 100 µl containing 500 nM of each primer, 200 µM dNTPs, 1× HF buffer and 2 units iProof (Bio-Rad) and run with the cycle: 98 °C for 30 s, 8 cycles of 98 °C for 10 s/67 °C for 15s/72 °C for 15 s, then 72 °C for 5 min. PCR product was purified with QiaQuick PCR clean-up kit.

The *Msp*I control library was constructed in the same manner as the *Hpa*II library, with the following changes: (i) in the first step 40 units of *MspI* (NEB) were used in place of *HpaII* and NEBuffer 2 was used instead of NEBuffer 1; and (ii) no amplification was done after gel purification.

The inverse library was constructed in this manner: *Hpa*II digestion was performed as done in the *Hpa*II library. After this, 10 units Antarctic Phosphatase (NEB) and 11 µl 10× Antarctic Phosphatase Buffer (NEB) were added to the mixture, which was then incubated at 37 °C for 1 h, and 65 °C for 15 min. DNA was purified with phenol:chloroform followed by ethanol precipitation. The DNA was then resuspended and treated in the same manner as the *Msp*I control library.

MSCC sequencing and read placement. In total, three lanes of sequencing were performed using an Illumina Genome Analyzer: two for the first technical replicate and one for the second technical replicate. These reads each contained sequence from the adapters and an 18–19 bp 'tag' derived from genomic sequence.

To match sequences, a list of all possible tags was created from all CCGG sites in the human genome (hg18, downloaded from UCSC). Tags were considered unique (later used for profiling) if no identical or single-mismatch tags existed, the neighboring *Hpa*II site was at least 40 bp distant and there were no conflicting *Mme*I recognition sites. An in-house program was used to find all tag matches within 0-, 1- or 2-base distances. Reads were accepted if they were an exact match and no single mismatches could be made, or if there was no exact match, a single mismatch and no double mismatches existed. Counts or observations are the number of times a particular location was matched by a read (**Supplementary Table 7**).

MSCC data analysis. To validate MSCC data we compared it with BSPP data collected for a set of 381 shared CpG locations (726 total tags) to get "counts versus methylation" information. These data points were binned according to methylation to form 20 bins with 36 or 37 data points each and the average counts versus average methylation was plotted. We expect average counts to be linearly related to methylation with the equation: methylation = a * counts – 1. A best fit for this equation to the average data points was produced with a = –0.1124. This was used to infer methylation when plotting average counts information.

Positions relative to genes for each MSCC site were calculated as before, using the RefGene list from UCSC. For multiple possible starts/ends, only the first entry was used. Using expression data genes were split into five equally sized groups based on gene expression levels. Running averages of MSCC counts were made for each graph: an interval of 5,000 data points for **Figure 2b**, an interval of 5,000 data points and 500 bp minimum window size for **2c** and **2d** and 500 data points with 500 bp minimum and 2,000 bp maximum windows for **Figure 3**. Counts were normalized for local CpG density (surrounding 200 bp), for *Msp*I control library counts, and, for the in-gene locations in **Figure 2b**, for gene length. To exclude data from gene promoter regions, only genes at least 15 kb in length were used to generate **Figure 2d**.

To analyze promoters based on CpG density, promoters were split into three types based on CpG density. Looking within the interval of –0.5 kb to +2 kb relative to transcription start (based on refGene annotation): high CpG promoters contain a 500-bp interval with a GC content of at least 0.55 and a CpG observed/expected ratio of at least 0.75, low CpG promoters contained no 500-bp interval with a CpG observed/expected ratio of at least 0.48, and all remaining promoters were defined as intermediate CpG promoters. Of 17,546 promoters analyzed, 11,445 (65%) were defined as HCP, 2,849 (16%) were defined as ICP, and 3,252 were defined as LCP (28%).

Methylation profiles for individual genes were created by finding average MSCC counts in the promoter region (defined as –400 to +1,000 bp) and in the gene body (defined as +3,000 bp to the end). Only genes with at least 10 MSCC data points in each region were plotted.

Accession numbers. Short reads (with quality scores) have been uploaded to the SRA (short read archive) databases at NCBI under accession number SRA008183.

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

ACKNOWLEDGMENTS

We thank Kun Zhang for discussion throughout this work; Wei Lin for help with computational design; Andrew Chess and Ravid Straussman for discussion and critical reading of the manuscript; Harvard Biopolymers Facility for Solexa sequencing; and Harvard Partners Center for Genetics and Genomics for gene expression profiling. This work was supported by the NHGRI-Centers of Excellence in Genomic Science (to G.M.C.).

AUTHOR CONTRIBUTIONS

M.P.B., J.B.L. and G.M.C. conceived the study, designed the research and wrote the manuscript. M.P.B. and J.B.L. performed experiments and data analysis. Y.G. and B.X. carried out initial Solexa sequencing. J.-H.L. helped with culturing cell lines and isolating DNA/RNA. E.M.L. synthesized the padlock oligos. I.-H.P. and G.Q.D. generated the iPS cell lines.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/

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

1. Suzuki, M.M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).

2. Goll, M.G. & Bestor, T.H. Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.* **74**, 481–514 (2005).
3. Feinberg, A.P. & Tycko, B. The history of cancer epigenetics. *Nat. Rev. Cancer* **4**, 143–153 (2004).
4. Jiang, Y.H., Bressler, J. & Beaudet, A.L. Epigenetics and human disease. *Annu. Rev. Genomics Hum. Genet.* **5**, 479–510 (2004).
5. Clark, S.J., Harrison, J., Paul, C.L. & Frommer, M. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.* **22**, 2990–2997 (1994).
6. Bird, A.P. & Southern, E.M. Use of restriction enzymes to study eukaryotic DNA methylation: I. The methylation pattern in ribosomal DNA from *Xenopus laevis*. *J. Mol. Biol.* **118**, 27–47 (1978).
7. Keshet, I. *et al.* Evidence for an instructive mechanism of *de novo* methylation in cancer cells. *Nat. Genet.* **38**, 149–153 (2006).
8. Cross, S.H., Charlton, J.A., Nan, X. & Bird, A.P. Purification of CpG islands using a methylated DNA binding column. *Nat. Genet.* **6**, 236–244 (1994).
9. Rauch, T.A., Wu, X., Zhong, X., Riggs, A.D. & Pfeifer, G.P. A human B cell methylome at 100-base pair resolution. *Proc. Natl. Acad. Sci. USA* **106**, 671–678 (2009).
10. Hellman, A. & Chess, A. Gene body-specific methylation on the active X chromosome. *Science* **315**, 1141–1143 (2007).
11. Khulan, B. *et al.* Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome Res.* **16**, 1046–1055 (2006).
12. Zhang, X. *et al.* Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell* **126**, 1189–1201 (2006).
13. Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T. & Henikoff, S. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.* **39**, 61–69 (2007).
14. Bibikova, M. *et al.* High-throughput DNA methylation profiling using universal bead arrays. *Genome Res.* **16**, 383–393 (2006).
15. Irizarry, R.A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41**, 178–186 (2009).
16. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).
17. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
18. Cokus, S.J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
19. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
20. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
21. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
22. Shann, Y.J. *et al.* Genome-wide mapping and characterization of hypomethylated sites in human tissues and breast cancer cell lines. *Genome Res.* **18**, 791–801 (2008).
23. Nilsson, M. *et al.* Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **265**, 2085–2088 (1994).
24. Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21**, 673–678 (2003).
25. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
26. Deng, J. *et al.* Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat. Biotechnol.* advance online publication, doi:10/1038/nbt.1530 (29 March 2009).
27. Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* **38**, 1378–1385 (2006).
28. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
29. Church, G.M. The personal genome project. *Mol. Syst. Biol.* **1**, 0030 (2005.).
30. Appanah, R., Dickerson, D.R., Goyal, P., Groudine, M. & Lorincz, M.C. An unmethylated 3′ promoter-proximal region is required for efficient transcription initiation. *PLoS Genet.* **3**, e27 (2007).
31. Buratowski, S. Transcription. Gene expression–where to start? *Science* **322**, 1804–1805 (2008).
32. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* **39**, 457–466 (2007).
33. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
34. Schuster, S.C. Next-generation sequencing transforms today's biology. *Nat. Methods* **5**, 16–18 (2008).
35. Mardis, E.R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).
36. Kahvejian, A., Quackenbush, J. & Thompson, J.F. What would you do if you could sequence everything? *Nat. Biotechnol.* **26**, 1125–1133 (2008).
37. Jones, P.A. The DNA methylation paradox. *Trends Genet.* **15**, 34–37 (1999).
38. Illingworth, R. *et al.* A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.* **6**, e22 (2008).
39. Park, I.H., Lerou, P.H., Zhao, R., Huo, H. & Daley, G.Q. Generation of human-induced pluripotent stem cells. *Nat. Protoc.* **3**, 1180–1186 (2008).
40. Ge, B. *et al.* Survey of allelic expression using EST mining. *Genome Res.* **15**, 1584–1591 (2005).
41. Lewin, J., Schmitt, A.O., Adorjan, P., Hildmann, T. & Piepenbrock, C. Quantitative DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplificates. *Bioinformatics* **20**, 3005–3012 (2004).