

Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing

Eun-Joon Lee^{1,2}, Lirong Pei^{1,2}, Gyan Srivastava³, Trupti Joshi³, Garima Kushwaha³, Jeong-Hyeon Choi^{1,4}, Keith D. Robertson^{1,2}, Xinguo Wang⁵, John K. Colbourne⁶, Lu Zhang⁷, Gary P. Schroth⁷, Dong Xu³, Kun Zhang⁸ and Huidong Shi^{1,2,*}

¹GHSU Cancer Center, ²Department of Biochemistry and Molecular Biology, Georgia Health Sciences University, Augusta, GA 30912, ³Department of Computer Science, Christopher S. Bond Life Sciences Center and Informatics Institute, University of Missouri, Columbia, MO 65211, ⁴Department of Biostatistics, Georgia Health Sciences University, Augusta, GA 30912, ⁵David H. Murdock Research Institute, Kannapolis, NC 28081, ⁶The Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47405, ⁷Illumina Inc. Hayward, CA 94545 and ⁸Department of Bioengineering, University of California at San Diego, La Jolla, CA 92093, USA

Received February 11, 2011; Revised June 15, 2011; Accepted July 6, 2011

ABSTRACT

We applied a solution hybrid selection approach to the enrichment of CpG islands (CGIs) and promoter sequences from the human genome for targeted high-throughput bisulfite sequencing. A single lane of Illumina sequences allowed accurate and quantitative analysis of ~1 million CpGs in more than 21 408 CGIs and more than 15 946 transcriptional regulatory regions. Of the CpGs analyzed, 77–84% fell on or near capture probe sequences; 69–75% fell within CGIs. More than 85% of capture probes successfully yielded quantitative DNA methylation information of targeted regions. Differentially methylated regions (DMRs) were identified in the 5'-end regulatory regions, as well as the intra- and intergenic regions, particularly in the X-chromosome among the three breast cancer cell lines analyzed. We chose 46 candidate loci (762 CpGs) for confirmation with PCR-based bisulfite sequencing and demonstrated excellent correlation between two data sets. Targeted bisulfite sequencing of three DNA methyltransferase (DNMT) knockout cell lines and the wild-type HCT116 colon cancer cell line revealed a significant decrease in CpG methylation for the *DNMT1* knockout and *DNMT1*, *3B* double knockout cell lines, but not in *DNMT3B* knockout cell line. We demonstrated the targeted bisulfite sequencing approach to be a powerful method to uncover novel aberrant methylation in the cancer epigenome.

Since all targets were captured and sequenced as a pool through a series of single-tube reactions, this method can be easily scaled up to deal with a large number of samples.

INTRODUCTION

DNA methylation plays an essential role in normal development and is involved in the pathogenesis of a variety of human diseases (1). The human genome is comprised of 3 billion bp of DNA, of which ~30 million are CpG dinucleotides, the main sites of DNA methylation. Given the GC frequency, however, the number of CpGs is much lower than expected. One unique feature of the mammalian genome is the presence of genomic regions containing dense clusters of CpGs, called 'CpG islands' (CGIs). CGIs are associated with over half of the known promoters. CpG methylation, occurring in CGIs, has functional consequences, such as transcriptional repression or aberrant activation (2). Many tissue-specific genes are known to be silenced by promoter methylation. Accumulating evidence has also indicated that hypermethylation of CGIs can be one of the most prevalent molecular events in human cancers (3,4).

The methylation status of DNA can be detected by several methods including: (i) methylation-sensitive restriction enzymes that only cut DNA at unmethylated recognition sites; (ii) sodium bisulfite which converts unmethylated cytosines to uracils but has no effect on methylated cytosines; (iii) methylated or unmethylated

*To whom correspondence should be addressed. Tel: +1 706 721 6000; Fax: +1 706 721 2928; Email: hshi@georgiahealth.edu
Present address:

Gyan Srivastava, Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.

DNA can be enriched by certain protein or antibodies (5). These procedures are usually combined with genome-screening techniques to examine DNA methylation on a whole- or subgenome scale (6–9). With the rapid advances in next-generation sequencing (NGS) technologies, whole-genome shotgun bisulfite sequencing was successfully used to map the complete methylomes of several human embryonic stem (ES) cell lines at the single-methyl-cytosine resolution (10,11). The single base-level analysis was instrumental for the discovery of non-CpG methylation in ES cells (11). However, in order to analyze a large number of clinical samples, technology that is quantitative, high throughput, cost-effective and both scalable and flexible with respect to coverage is extremely desirable (12).

Several genomic enrichment (or targeted capture) methods developed for whole-exome sequencing (13) were recently applied to targeted bisulfite sequencing (TBS) applications (12,14–16). Hodges and colleagues demonstrated that bisulfite-converted DNA can be hybridized to an oligonucleotide capture array to enrich regions of interest for TBS (16). Within 324 selected CGIs, 25 000 CpGs were successfully sequenced. Deng and colleagues designed approximately 30 000 padlock probes to assess methylation of approximately 66 000 CpG sites within 2020 CGIs on human chromosomes 12 and 20 (15); Ball and colleagues designed approximately 10 000 padlock probes to profile approximately 7000 CpGs within the ENCODE pilot project regions (14). Both studies demonstrated that the TBS approach based on padlock probes was highly specific and reproducible. However, due to the reduced sequence complexity after bisulfite treatment, the methods that directly enrich targets from bisulfite-converted DNA by hybridization may have difficulty selecting the optimal capture probes in the highly GC-rich regions. Nautiyal and colleagues developed a unique method for capturing target DNA via oligo-guided ligation before bisulfite conversion (12). The methylation status of more than 145 000 CpGs from 5472 promoters was determined using a microarray. One of the challenges of this methodology is the construction of the capture probe panels, which were prepared by single-plex PCR reactions (12).

In this study, we present an approach that combines solution-phase hybrid selection and massively parallel bisulfite sequencing to profile DNA methylation in targeted CGIs and promoter regions. The solution hybrid selection method was initially developed for capture-sequencing of exons (17). However, successful hybrid selection of genomic DNA coupled with bisulfite conversion has not been demonstrated. In addition, the high GC content present in the CGIs makes capturing these regions a challenge. We designed 51 551 highly specific capture probes and optimized the solution hybrid selection method for TBS in the CGIs and promoter regions across the human genome. Using one lane of Illumina sequences, we can quantitatively determine the methylation status of ~0.9–1 million CpGs in more than 21 408 CGIs and more than 15 946 transcriptional regulatory regions. Single-base resolution CpG methylation maps in the promoters of a large number of genes in three breast cancer cell lines were generated.

Also, differential methylation in the 5'-end of regulatory regions, as well as the intra- and intergenic regions, particularly in the X-chromosome was identified. The method developed was further tested using wild-type and DNA methyltransferase (DNMT) deficient colon cancer cell lines and the results revealed novel CpG methylation patterns in these cell lines.

MATERIALS AND METHODS

Preparation of biotin-labeled capture probes

The method for preparation of the biotinylated capture probe is the very similar to the method described by Gnirke *et al.* (17) with minor modifications. The oligonucleotide library was purchased from Agilent Technologies. It consisted of 51 551 oligonucleotides of the sequence 5'-AGGACCGGATCAACTN₁₆₀CATTGC GTGAACCGA-3' with N₁₆₀ indicating the capture probe sequences (17). The information on the probes (start and end positions of the probe sequences in the human genome), which can reproducibly capture target DNA is provided in Supplementary Table S1. The library was resuspended in 100 µl 0.1× TE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0). A 4 µl aliquot was used for the first PCR amplification in 100 µl containing 800 nmol dNTPs, 40 pmol each of PCR primer eMIP_CA_F (5'-TGCCTAGGACCGGATCAACT-3') and eMIP_CA_R (5'-GAGCTTCGGTTCACGCAATG-3') (18), and 5 U Pfu TurboCx DNA polymerase (Stratagene). The PCR conditions used were 5 min denature at 94°C, followed by 20 cycles of 20 s at 94°C, 30 s at 55°C, 30 s at 72°C and 5 min final extension at 72°C. The PCR products were purified using a PCR cleanup kit (Qiagen) and loaded on a 2% NuSieve 3:1 agarose gel (Lonza). A 200-bp band was cut from the gel and purified using a gel purification kit (Qiagen). The gel-purified DNA was diluted 1:1000 and kept at -80°C. To add the T7 promoter, 8 µl diluted PCR product was re-amplified in 200 µl aliquots containing 800 nmol dNTPs, 40 pmol of each of PCR primer combination eMIP_T7_F (5'-TAA TACTACTACTATAGGGAGGACCGGATCAACT-3')/eMIP_CA_R (5'-GAGCTTCGGTTCACGCAATG-3') and eMIP_T7_R (5'-TAATACGACTCACTATAGGGA GAGCTTCGGTTCACGCAATG-3')/eMIP_CA_F (5'-T GCCTAGGACCGGATCAACT-3') and 5 U Pfu TurboCx DNA polymerase (Stratagene) in two separate PCR reactions with the same conditions mentioned above. The 215-bp PCR products were purified using a PCR purification kit (Qiagen). To generate the biotinylated RNA probes, 1 µg of the purified PCR products was used as template in a 100 µl *in vitro* transcription reaction using a MAXIscript T7 kit, (Ambion). The template DNA was removed by incubating for 15 min with 1 µl Turbo DNase (Ambion). The reaction was stopped by adding 0.5 M EDTA and the biotinylated RNA probes were purified using a MEGAclear kit (Ambion). Biotinylated RNA probes were kept in the presence of 1 U/µl SUPERase-In RNase inhibitor (Ambion) at -80°C.

Sequencing library construction

Genomic DNA was isolated from cancer cell lines using DNeasy Blood and Tissue Kit (Qiagen). The genomic DNA was randomly fragmented into 150–500-bp segments using a Bioruptor (Diagenode). The sheared DNA was repaired, and an adenine was added to the 3'-end of the DNA fragments. Pre-annealed Illumina sequencing adaptors containing 5-methyl-cytosine were ligated to both ends of the DNA fragments according Illumina standard protocols (Illumina). The sequencing library was then purified with an Ampure XP Kit (Agencourt). About 20–30 µg of DNA each from breast cancer cell lines MCF10A, MCF7 and MDA-MB-231 were used to establish the protocols, which yielded roughly 12 µg of ligated fragment library that is enough for six capture reactions for each cell line. To make the assay consistent, we attempted to maintain a constant ratio of library DNA to the amount of capture probes used during the hybridization reaction. To investigate if the amount of input DNA affects sequencing results, we performed bisulfite sequencing experiments using DNA obtained from 6, 3, 2 or 1 capture reactions. The details about the amount of captured DNA used in each experiment are listed in Table 1.

Hybrid selection of target DNA

The hybridization procedure was performed as previously described (17). Briefly, 5 µg human Cot-1 DNA (Invitrogen), 5 µg Salmon sperm DNA (Invitrogen) and 1 µg pair-end DNA library were mixed and adjusted to 7 µl by vacuum drying. The mixture was denatured at 95°C for 5 min, and annealed at 65°C for 5 min. Then, 13 µl of pre-warmed 2× hybridization solution (10× SSPE, 10× Denhardt's, 10 mM EDTA and 0.2% SDS) and 6 µl of a freshly prepared mixture containing 1 µg

biotinylated RNA probe and 20 U SUPERase-In were added at 65°C. The hybridization mixture was incubated for 48 h at 65°C. After the hybridization was finished, the DNA:RNA hybrid was captured with 100 µl of M-280 Streptavidin Dynabeads (Invitrogen) that had been washed three times and resuspended in binding buffer (200 µl 1 M NaCl, 10 mM Tris-HCl, pH 7.5 and 1 mM EDTA). The bead/capture solution was rotated for 30 min at the room temperature and then was pulled down and washed once at room temperature for 15 min using 500 µl of 1× SSC/0.1% SDS. The bead/capture solution was again pulled down and washed three times at 65°C for 10 min using 500 µl of 0.1× SSC/0.1% SDS. For elution of the capture DNA, a portion of 0.2 M NaOH solution was added and incubated at room temperature for 10 min. The beads were pulled down and the supernatant was transferred to a new tube containing 70 µl of neutralization buffer (1 M Tris-HCl pH 7.5). The captured DNA was desalted and purified using a Zymo DNA Clean & Concentrator Kit (Zymo Research) and eluted in 20 µl. The neutralization step can be omitted, and the DNA in 0.2 M NaOH solution can be directly used for bisulfite treatment. For each sample, six capture reactions were performed with probes targeting the Watson strand, and six reactions were performed for the Crick strand. The captured DNA was pooled in a strand specific manner and used in two separate bisulfite treatment reactions.

Bisulfite conversion, PCR amplification and Illumina sequencing

Bisulfite treatment was performed on the captured DNA using EZ DNA Methylation Kit (Zymo Research, Orange, CA, USA) according to the manufacturer's

Table 1. Summary of the targeted bisulfite sequencing experiments

Cell lines	MCF10A	MDA-MB-231	MCF7-1 ^a	MCF7-2 ^a	MCF7	HCT116	1KO	3BKO	DKO	Normal breast	MCF10A	MDA-MB-231
Number of captures	6	6	6	6	3	3	3	3	3	2	1	1
Total number of raw reads (M ^b)	27.38	29.15	29.22	30.95	33.79	35.60	38.56	33.66	37.24	22.14	29.86	24.21
Total number of mapped reads (M)	22.40	24.13	24.08	25.61	27.44	27.86	32.02	27.22	29.57	17.87	22.86	17.40
Mapping rate (%)	77.39%	79.55%	79.04%	79.00%	81.22%	78.26%	83.04%	80.87%	79.76%	80.68	76.56%	71.86%
Total number of CpGs measured (M)	1.08	0.96	0.91	1.08	0.63	0.54	0.42	0.48	0.65	0.31	0.41	0.16
Average number of reads per CpG	86	137	146	112	96	229	204	234	129	89	25	87
Number of CpGs on/near probe (M)	0.84	0.75	0.76	0.84	0.36	0.45	0.31	0.40	0.41	0.17	0.12	0.03
On-target rate (%) ^c	77.78%	78.94%	83.51%	77.78%	77.78%	83.33	73.80%	83.33%	63.07%	55.28%	29.27%	18.75%
Number of CGIs mapped with reads	22 370	22 484	21 862	22 161	17 714	19 505	17 655	18 601	19 476	11 549	13 922	5107
Number of CpGs mapped on CGI (M)	0.77	0.69	0.70	0.75	0.34	0.43	0.30	0.38	0.39	0.12	0.12	0.03
Probes that mapped with more than 10 reads	44 567	45 266	43 709	44 266	23 681	26 222	23 306	24 849	25 747	14 892	17 996	5663
Probe success rate (%)	86.45%	87.81%	84.79%	85.87%	45.93%	50.87%	45.21%	48.20%	49.94%	28.89%	34.91%	10.99%

^aThese are technical replicates.

^bM: million.

^cNumber of CpG on/near probe divided by the total number of CpG.

protocol. The bisulfite-modified DNA was used immediately for PCR or stored at -80°C . For the first round of PCR amplification, the bisulfite-converted DNA pooled from six capture reactions was amplified in $4 \times 50 \mu\text{l}$ aliquots by 10 cycles of PCR using the following reaction mixture, 2.5 U of uracil-insensitive Pfu TurboCx Hotstart DNA polymerase (Stratagene), $5 \mu\text{l}$ $10 \times$ Pfu TurboCx buffer, 2.5 mM dNTPs, $25 \mu\text{M}$ of each Pair End (PE) Primers 1 and 2. Cycling conditions for the reaction were: 95°C for 2 min, 98°C for 30 s, then 10 cycles of 98°C for 15 s, 60°C for 30 s and 72°C for 4 min, final extension at 72°C for 10 min. In order to completely remove the primer dimers, the PCR products were first purified using a Zymo DNA Clean & Concentrator Kit (Zymo Research) followed by a second purification with an Ampure XP Kit (Agencourt). For the second round of PCR amplification, the first PCR product was amplified in $50 \mu\text{l}$ volume using 14 cycles of PCR with the following reaction mixture, 2.5 U Phusion DNA polymerase (Finnzyme), $10 \mu\text{l}$ $5 \times$ Phusion buffer HF, 1.5 mM dNTPs, 25 mM of each Pair End (PE) Primers 1 and 2. Cycling conditions for the reaction were: 98°C for 15 s, 65°C for 30 s and 72°C for 4 min, final extension at 72°C for 5 min. The PCR products were purified using a Zymo DNA Clean & Concentrator Kit (Zymo Research) as well as an Ampure XP Kit (Agencourt). The PCR products were sequenced using an Illumina GAIIx according to manufacturer's protocol.

Amplicon sequencing using 454-GS-FLX

Bisulfite-PCR primers for 46 genes were designed using the PRIMEGENSv2 software (19) and synthesized by IDT in a 96-well format. The amplicon information is listed in Supplementary Table S2. Bisulfite-modified genomic DNA from MDA-MB-231 and MCF7 was used as the template for PCR amplification of the 46 genes. The individual PCR products were pooled together, purified, end-repaired and ligated to the bar-coded sequencing adaptors using 454 library construction kits and sequenced according to the manufacturer's protocols (Roche 454 Life Sciences).

ChIP and ChIP-seq

ChIP experiments were carried out using a Magna ChIP kit (Millipore) following the manufacturer's suggested protocols. For each ChIP experiment, 5×10^6 MDA-MB-231 cells were cross-linked with 1% formaldehyde at 37°C for 10 min. The anti-trimethyl-H3K4 (Upstate, no. 04-745) antibody was used. The ChIP-seq libraries were prepared from 10 ng of both Input and ChIP DNA samples using a ChIP-seq sample preparation kit (Illumina) according to the manufacturer's protocol. The libraries were sequenced on a GAII sequencer (Illumina) at 36-bp read length. The raw sequencing reads were processed and aligned against the human genome using ELAND program (Illumina) and the peaks were identified using MACS (20).

Bisulfite sequences alignment and methylation analysis

The sequence analysis was done using an in-house bioinformatics analysis pipeline in which the raw bisulfite sequencing reads generated using Illumina sequencing were first converted to 'unmethylated reads' via the conversion of all Cs to Ts *in silico*. The unmethylated, converted sequencing reads were then aligned to the two copies of *in silico* converted human reference genome (hg18) using SOAP2 (15). The results of both mappings were combined, and any reads mapped against both copies of converted reference were resolved by identifying their strand based on the TCAG ratio and the mapping positions used accordingly. The clonal reads were not removed when the sequencing reads were mapped to the genome. The methylation level $[(\text{number of methylated}) / (\text{number of methylated} + \text{number of unmethylated})]$ was calculated after measuring the amount of methylation (CG \rightarrow CG) and unmethylation (CG \rightarrow TG) in the CpG position. The results were stored and can be visualized using the UCSC Genome Browser. Cluster analyses and statistical analyses were performed using Partek Genomic Suites (Partek) and R. The genomic annotation was performed using in-house Perl scripts. The 454-sequencing reads were analyzed using BMapper (<http://sourceforge.net/projects/bsmapper>). The raw and processed sequencing data were submitted to NCBI gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo>). The accession number is GSE26826 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26826>).

RESULTS

The overall experimental approach is illustrated in Figure 1a. A pool of single-stranded DNA oligonucleotides was synthesized by Agilent Technologies' oligo library synthesis (OLS) services. Each oligonucleotide consisted of a target-specific 160-mer sequence flanked by 15 bases of a universal primer sequence on each side to allow PCR amplification. After initial PCR, a T7 promoter was added on each side of the PCR product in two separate PCR reactions. The double-stranded PCR products were converted into two sets of complementary biotin-labeled RNA probes by *in vitro* transcription (17), which were used to capture Watson and Crick strands of the genomic DNA in two separate capture reactions, respectively. The captured DNA was pooled from several hybridization reactions, treated with sodium bisulfite, amplified by PCR and subsequently sequenced using an Illumina GA IIx sequencer (Figure 1a).

We designed 51 551 highly specific capture probes using the following criteria: (i) one capture probe (160 bp) for every 500 bp window within a given CGI across the genome. CGIs <500 bp only have one probe per island; (ii) one capture probe per transcription start sites (TSS) for all RefSeq genes. We implemented these criteria in the PRIMEGENS-v2 software to search for unique 160-bp segments within the targeted regions (19,21). A highly stringent criterion for cross hybridization was used and only highly specific probes were selected. A total of 35 746 probes were successfully designed to target 23 441

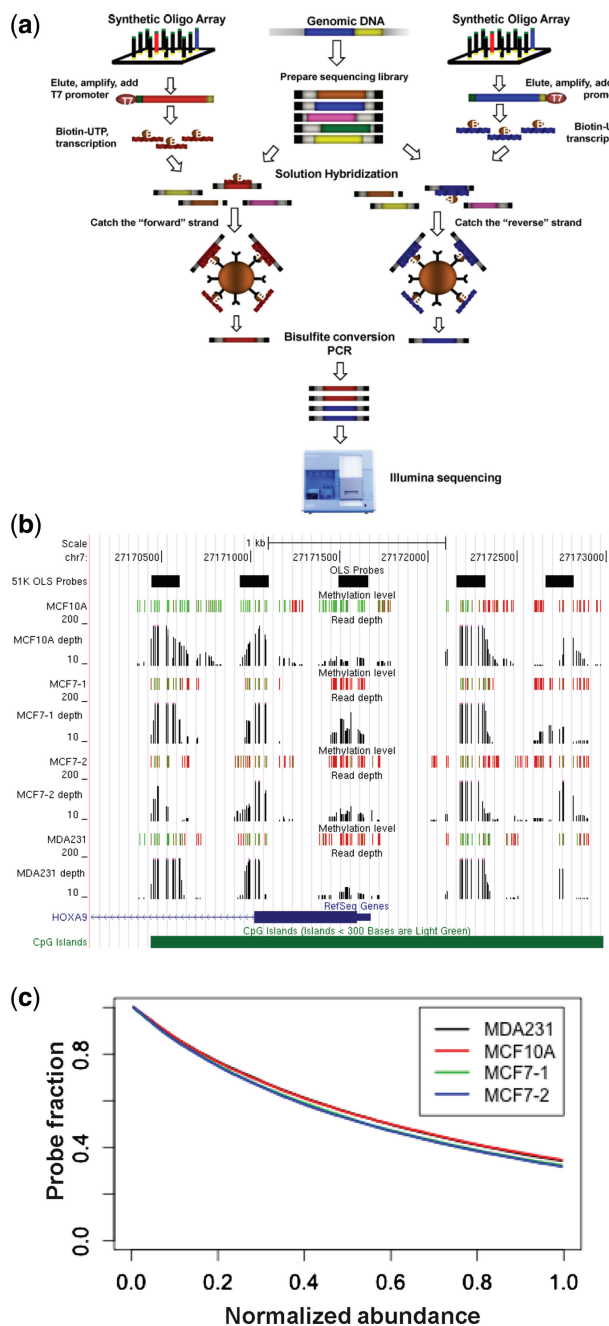


Figure 1. Targeted bisulfite sequencing. (a) Illustrated are the steps involved in the preparation of biotinylated RNA capture probes (top left and right), whole-genome fragment input library (top middle) and hybrid selection-enriched output library (middle left and right). The captured DNA was treated with sodium bisulfite, amplified by PCR and sequenced using an Illumina GAIIX sequencer. (b) Examples of targeted bisulfite sequencing using solution hybrid selection. The tracks shown from the top to bottom are: sequences of the capture probes; the DNA methylation level at each CpG site derived from the bisulfite-sequencing reads; the sequencing depth at each CpG site; RefSeq genes; and annotated CGIs in the UCSC genome browser. Red and green colors indicate methylated and unmethylated-CpG sites, respectively. MCF7-1 and MCF7-2 are technical replicates. For each sample, two tracks (methylation level and read depth) are shown. (c). Distribution of normalized abundance for the captured targets shared among all samples. The x-axis is the normalized abundance of each captured target, which is calculated by dividing the counts of the target by the average counts of all targets. The y-axis is the fraction of probes with coverage equal to or greater than the normalized coverage.

CGIs (~83% of the 28 226 CGIs in the human genome). Next, we designed 15 805 probes to capture the regions crossing the TSS (± 300 bp) of 19 369 RefSeq genes after excluding genes that already contained the CGI that spanned around the TSS. Overall, the capture probes targeted over 8.2 million bases in the genome and overlap with >80% of CGIs and the promoters of known genes in the human genome.

Targeted bisulfite sequencing was performed on three breast cancer cell lines, MCF7, MDA-MB-231 and MCF10A. We obtained 23–30 million single-end 76-bp sequencing reads for each cell line. The bisulfite sequences were analyzed using an in-house bioinformatics analysis pipeline. We mapped the raw reads to the *in silico* bisulfite-converted human genome, and an average of ~79% unique mapping rate was achieved. As an example, the typical bisulfite sequencing results and sequencing coverage profiles in the promoter CGI of *HOXA9* are shown in Figure 1b. The average sequencing depth per CpG is between 86 and 146 across cell lines (Table 1). The fraction of probes with at least half of the average abundance is 52–54% (Figure 1c), which is comparable to previous studies (15,17). Methylation levels for CpG sites covered with at least 10 sequencing reads were extracted providing accurate quantification for ~1 million CpGs. Overall, 77–84% of the CpGs measured in the samples fell on or near capture probe sequences; 69–75% fell within CGIs. More than 85% of capture probes successfully yielded quantitative DNA methylation information of the targeted regions (Table 1). As expected, ~40% of CpGs were located in 5'-end transcriptional regulatory regions including promoters and 5'-UTRs, while <18% of CpGs were associated with repetitive sequences (Supplementary Figure S1). In all three cell lines (Supplementary Table S1), 21 408 CGIs and 15 946 transcriptional regulatory regions were captured and sequenced. Bisulfite treatment efficiency can be determined by calculating the C to T conversion (rate for all cytosine bases other than those in CpG dinucleotides (this includes CpA, CpC or CpT dinucleotides). The bisulfite conversion rate was estimated to be 98.3, 98.3 and 98.7% for MCF7, MCF10A and MDA-MB-231, respectively.

Next, the amount of input DNA was decreased in order to determine the effects on the performance of TBS method based on solution capture. When half of the amount of captured DNA (from three capture reactions) was used for bisulfite treatment and subsequent Illumina sequencing, the number of CpGs sequenced was reduced by ~45% (Table 1). The bisulfite sequence mapping rate and the on-target rate did not change significantly, indicating that the specificity of the solution hybrid capture was not affected by the amount of input DNA. However, the sensitivity of the TBS approach was adversely affected. When DNA from only a single capture reaction was used, a significant decrease in the number of probes that were able to capture the target DNA was observed. Using genomic DNA from a normal breast tissue sample (~10 μ g of DNA), we were able to measure the methylation status of ~0.3 million CpGs and 11 549 CGIs.

To validate the measurement accuracy of the targeted bisulfite sequencing, we first compared the methylation

level of the same CpG sites on two opposite strands. Since we captured the forward and reverse strands separately (Figure 1a), and CpG methylation is symmetric on the two DNA strands, the accuracy of the assay can be determined by comparing the methylation level of these CpG sites on the two strands (15). For 41 257 such CpG sites that were covered by more than 100 sequencing reads, the Pearson correlation coefficient (R) was 0.955 (Figure 2a), despite the fact that there was little correlation in the sequencing coverage at these CpG sites between the two strands (Figure 2b). To confirm the targeted bisulfite sequencing results, we quantified the methylation levels of 762 CpG sites from 46 randomly selected PCR products using 454 sequencing (Figure 3a–d and Supplementary Table S2). Correlation between the two assays was very good, but not perfect, at the single-CpG level ($R = 0.818$ and 0.826 for MCF7 and MDA-MB-231, respectively). From Figure 4 and Supplementary Table S2, it is apparent that 454 amplicon sequencing often gave a higher methylation measurement. In the MCF7 cell line, >80% of 219 CpG sites containing different methylation measurements were found to have greater methylation

levels when measured by 454 sequencing. Similar results were obtained for the MDA-MB-231 cell line. The higher methylation measurements indicated by 454 sequencing could be due to selective amplification of methylated DNA (which is more GC rich than unmethylated DNA) in the initial PCR reaction. However, when compared at the amplicon level, the correlation between the two assays was excellent ($R = 0.924$ and 0.944 for MCF7 and MDA-MB-231, respectively; Figure 4). To test the reproducibility of the capture system, we compared the methylation measurements from two independent technical replicates. Excellent correlation was observed (Figure 2c). Therefore, the validation experiments demonstrate that the capture sequencing approach is accurate and reproducible.

Using this method, we generated genome wide, single-base resolution DNA-methylation maps in three of the most commonly used breast cancer cell lines which have been cited extensively in the literature. The overall methylation levels of CpGs displayed a similar bimodal distribution among the three cell lines, which is consistent with previous reports (Supplementary Figure S2). To evaluate the DNA methylation status of each targeted loci, we first

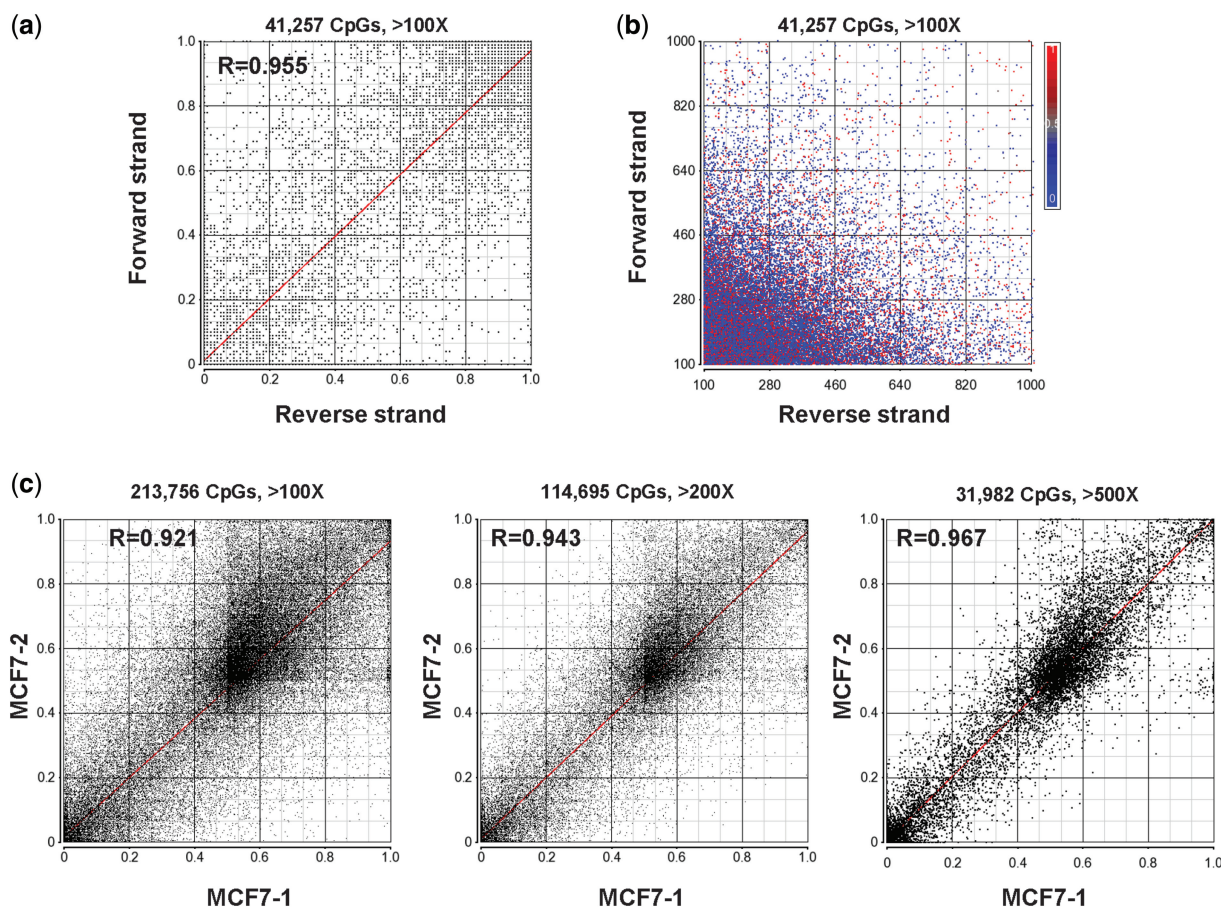


Figure 2. Accuracy and reproducibility of the targeted bisulfite-sequencing approach. (a) The scatter plot shows a comparison of methylation measurements from both strands of the same CpG sites. The methylation levels of the forward strand were plotted against the levels of the reverse strand on 41 257 CpGs covered by sequencing reads mapped to both strands. Representative data from MCF7-2 was used. (b) The scatter plot shows the comparison of sequencing coverage between forward and reverse strands for the same CpG sites. The same set of CpGs are used in both panels (a and b). (c) Comparison of the methylation levels between two technical replicates of MCF7. The same DNA sample from MCF7 was processed on a different day and sequenced in different batches. The data was plotted with increases in read depth per CpG. An increased Pearson correlation coefficient was observed, indicating that the read depth affects the consistency of the methylation data slightly.

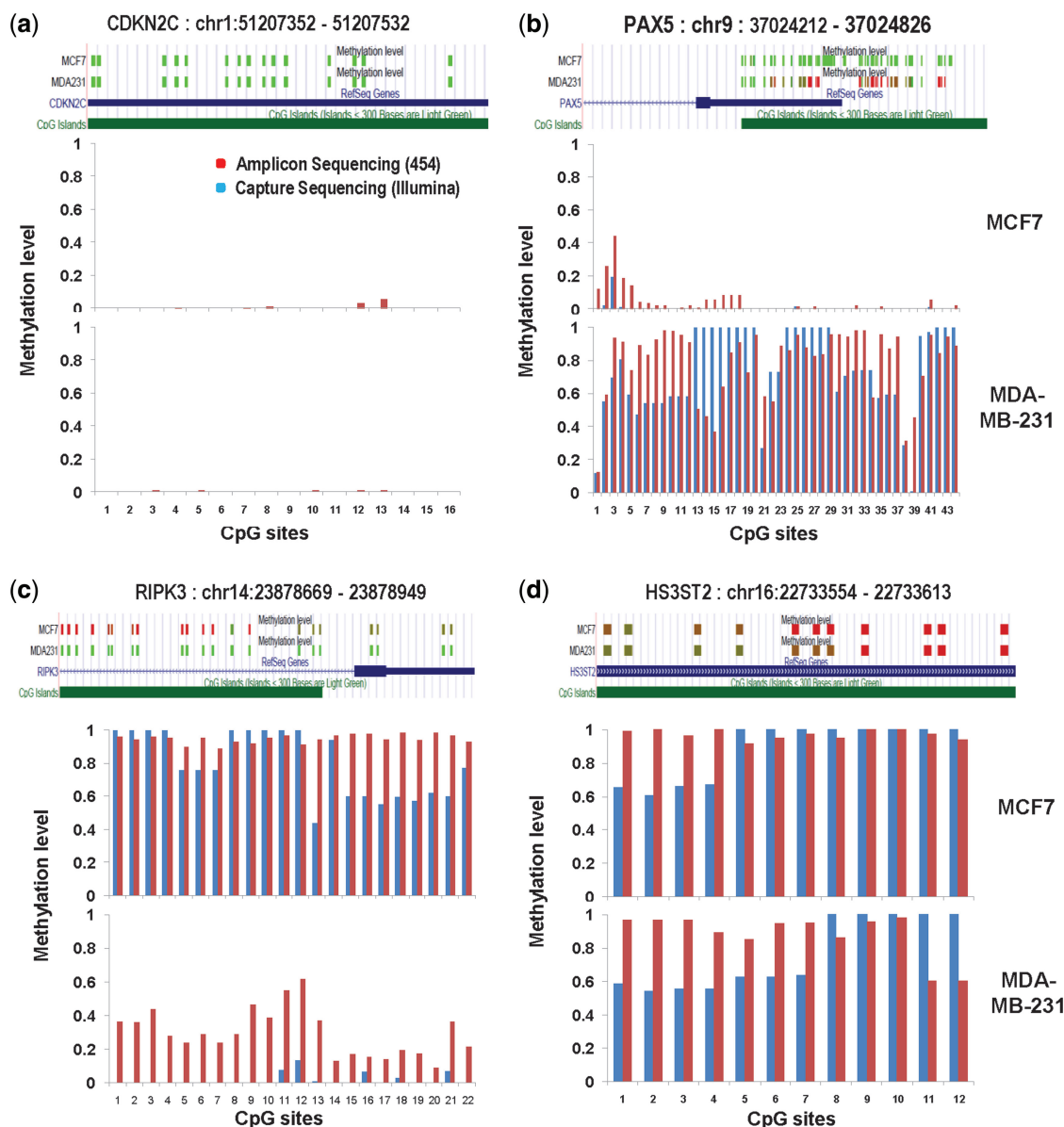


Figure 3. Confirmation of the targeted bisulfite-sequencing results by 454 amplicon bisulfite sequencing in MCF7 and MDA-MB-231 cells. (a–d) Differential DNA-methylation patterns in the CpGs of four representative amplicons are shown. The browser tracks shown in each panel from top to bottom are: the DNA methylation level at each CpG site derived from the bisulfite-sequencing reads; Ref-Seq genes; and annotated CGIs in the UCSC genome browser. Red and green colors indicate methylated and unmethylated CpG sites, respectively. The bar graphs show the comparison between two methods for the same four candidate genes. The x-axis is the CpG position in each amplicon, y-axis indicates the percentage of methylation.

grouped the CpGs based on the 500-bp windows centered by the capture probes, and then took an average of the methylation values of all CpGs within the 500-bp window. The average methylation value was assigned to each target loci. The methylation levels of 42 460 loci and their annotations are included in Supplementary Table S1. We split these targeted loci into three main categories: 5'-end of known genes (27 996 loci), intragenic regions including the body and 3'-end of genes (8921 loci), and intergenic regions (5472 loci). The distribution of the average methylation levels of target loci displayed a similar trend to the distribution at the single-CpG level (Figure 5a).

However, the patterns of distribution seemed to depend on their functional annotations, and are very similar between the three cell lines, except the X-chromosome. A significantly higher number of unmethylated loci were observed in the X-chromosome of MDA-MB-231 cells when compared to other cell lines (Figure 4a). Cluster analysis grouped the MDA-MB-231 and MCF10A cell lines together when the methylation data from 5'-end, intra- and intergenic loci was used. However, cluster analysis using methylation data from the X-chromosome-grouped MCF7 and MCF10A together (Figure 5b), separating them from the triple negative and aggressive

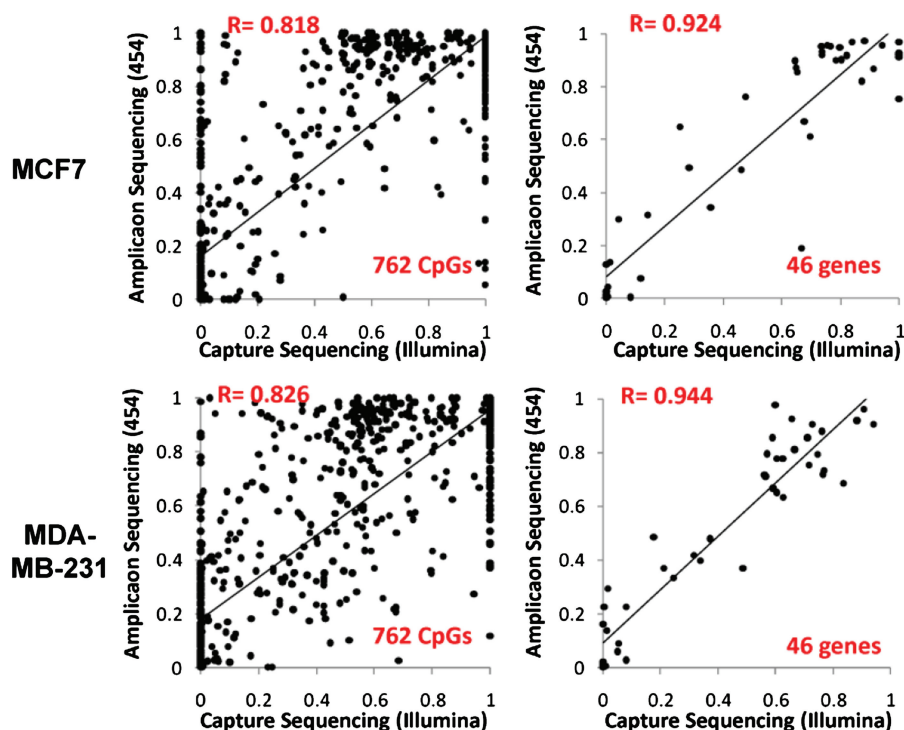


Figure 4. Scatter plots comparing the methylation levels of 762 CpG sites from 46 randomly selected genes determined by targeted bisulfite sequencing, as well as 454 amplicon bisulfite sequencing. For a base-by-base comparison (left panel), the methylation results from two platforms show good correlation in MCF7 ($R = 0.818$) and MDA-MB-231 ($R = 0.826$) cell lines, respectively. On a gene-by-gene basis (right panel), the average methylation levels of 46 genes were highly correlated between two methods ($R = 0.924$ and 0.944 for MCF7 and MDA-MB-231, respectively).

MDA-MB-231 cell line. Overall, differentially methylated regions (DMRs) can be identified in the 5'-end regulatory regions, as well as the intra- and intergenic regions, particularly in the X-chromosome among the three cell lines (Figure 5b and Supplementary Figure S3).

We identified 1155, 2133 and 985 methylated genes with average methylation levels equal or above 0.75 in MCF10A, MCF7 and MDA-MB-231 cell lines, respectively (Figure 5c). Surprisingly, MDA-MB-231 cells seem to have the fewest number of methylated genes among the three cell lines. These methylated genes were classified according to their gene ontology (GO) category (Supplementary Tables S3–S5). The single-CpG resolution methylation maps of many known tumor suppressor genes were established in the three cell lines. Supplementary Figure S4 reveals the methylation patterns in the CGIs of tumor suppressor genes such as *SFRP1*, *SFRP2*, *SOX17* and *WIF-1* which are all inhibitors of WNT-signaling pathway (Supplementary Figure S4A) as well as several microRNAs that are embedded in the CGIs (Supplementary Figure S4B). Supplementary Figure S5 illustrates differential methylation patterns in all four HOX gene clusters. In addition, the DNA methylation patterns correlated particularly well with histone modification data (Figure 5d). For instance, CpG methylation was observed in the promoter of *RASSF1A* tumor suppressor gene (the long transcript of *RASSF1*), but not in the promoter of its alternative splice variant, *RASSF1C*

(the short form of *RASSF1* transcript) (Figure 5d). This is consistent with our previous finding that DNA methylation functioned in concert with histone modifications to regulate gene expression of these specific transcripts (22). *GATA2* is another example, which has not been characterized previously. For both genes, the histone H3 lysine 4 tri-methylation (H3K4Me3) was associated with the unmethylated promoters. Particularly for *GATA2*, the CGI seems to be split in half and only the unmethylated region is clearly overlapping with H3K4Me3.

To further test the robustness of the targeted bisulfite sequencing method, we analyzed a group of *DNMT* knockout cell lines derived from colon cancer cell line HCT116 (23). These are isogenic cell lines with a single gene defect (23–25). Using targeted bisulfite sequencing, we analyzed 0.54, 0.42, 0.48 and 0.65 million CpGs in *DNMT1* knockout (1KO), *DNMT3B* knockout (3BKO) and *DNMT1*, *3B* double knockout (DKO) cell lines. Similar to previous reports (23,24,26,27), the targeted bisulfite sequencing data revealed a significant decrease in the number of methylated CpG sites in the *DNMT1*, *3B* double knockout (DKO) cell line, but only modest decreases in the single knockout, 1KO and 3BKO, cell lines (Figure 6a). Cluster analysis of the 129022 common CpGs shared by all four cell lines revealed an interesting pattern in which demethylation increased gradually from the 3BKO to DKO cell lines (Figure 6b). Approximately 30.5% of the 129022 CpGs were fully methylated

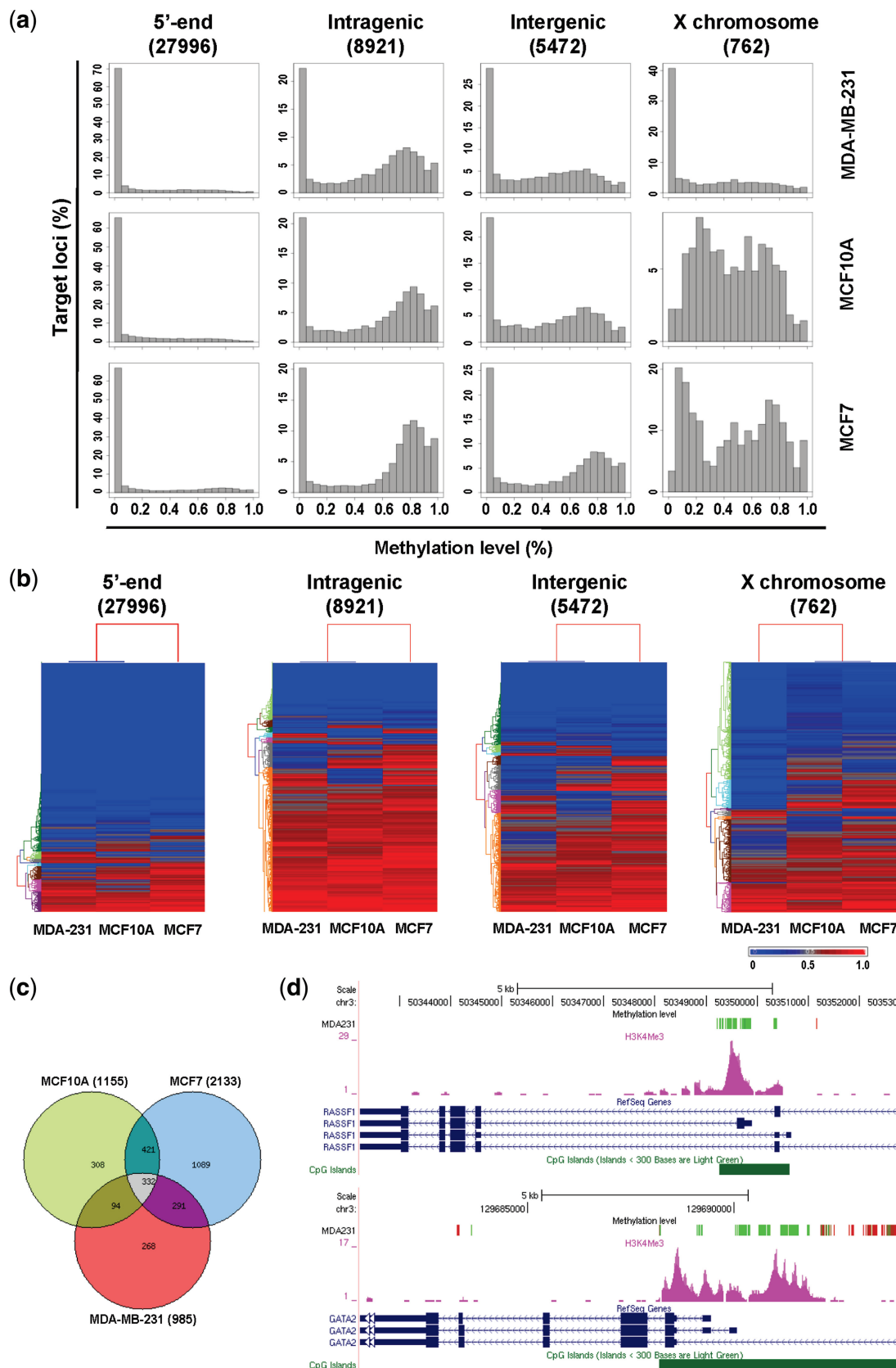


Figure 5. DNA-methylation analysis in three breast cancer cell lines. (a) The histograms show the distribution of the average methylation levels of target loci among different functional annotations (*n*, number of target loci). The target loci were annotated into three major categories: 5'-end of known genes (from -2000 bp to +500 bp of TSS), intragenic (-500 bp of TSS to the end of the transcript), and inter-genic regions. The methylation levels are bimodally distributed, except for loci in the X-chromosome of MCF7 and MCF10A cells which were unimodally distributed. (b) Clustering

(continued)

(methylation levels > 0.95) in the parental HCT116 cells. Further analysis of the 38 650 fully methylated CpGs (mCpGs) in HCT116 showed that 88.3% of mCpGs displayed some degree of decreased methylation in at least one DNMT knockout cell line. Almost half of the mCpGs (17 021 CpGs) had a methylation value of less than 0.5 (approximately ≥ 2 -fold reduction in methylation level) in at least one DNMT knockout cell line. We divided the 17 021 mCpGs into four different groups based on the patterns of methylation in four colon cancer cell lines (Figure 6c). Fifty-four percent of mCpGs were hypomethylated in both the 1KO and DKO cell lines (Class 2). Of mCpGs, 34% were hypomethylated only in the DKO cell line (Class 4). Of mCpGs, 9% were hypomethylated in all three knockout cell lines (Class 1) and only 3% of mCpGs had demethylation in both the 3BKO and DKO cell lines (Class 3). Figure 6d illustrates some of the demethylation patterns observed in the four cell lines. The demethylated CpG sites were widely distributed across the genome, affecting promoter and 5'-UTR (24%, 4019 CpGs), gene body including promoter, intron and 3'-UTR (33%, 5649 CpGs) and intergenic regions (43%, 7353) (Figure 6c). The distribution of demethylated CpGs in the aforementioned four classes did not change significantly among promoter, gene body and intergenic regions. Only 386 demethylated CpGs were associated with repetitive sequences that were widely distributed across the subclass of repetitive sequences.

DISCUSSION

For the first time, we have demonstrated that the solution hybrid-capture approach can be successfully used for targeted bisulfite sequencing. The solution hybrid selection approach has recently become a popular method for target enrichment. One unique feature of this protocol is to convert the long oligonucleotides into biotin-labeled RNA bait for capture of target DNA. This approach increases the specificity of the hybridization capture, and makes the capture process highly flexible. However, an earlier study indicated that targets having higher GC content resulted in a much lower coverage (17). This systematic bias poses a challenge for capturing CGI sequences. This challenge was overcome by carefully selecting unique capture probes sequences. We demonstrated that over 85% of the 51 551 probes successfully captured their target DNA. Using a single lane of Illumina sequences (20–30 million reads), we can reproducibly quantify 0.9–1 million CpGs in the human genome. This number is 6–36-fold higher than the previously published results that use other sequence capture methods. Since the number of capture probes can be easily expanded, the coverage of the TBS approach

described here can be significantly increased. We also showed that the TBS approach based on solution hybrid selection was highly reproducible. The accuracy and precision of this approach was further confirmed by bisulfite sequencing analysis of 46 candidate loci using an independent method. The current capture probe set focused mainly on the promoter sequences and CGIs; however, it can be easily expanded to include CGI shore, enhancer and other regulatory regions.

Although whole-genome bisulfite sequencing is ideal for methylome analysis, it remains cost-prohibitive for analyzing a large number of clinical samples. Meissner *et al.* (28) developed a reduced representation bisulfite sequencing (RRBS) method for sub-genome bisulfite sequencing using NGS and demonstrated its utility for analyzing clinical samples. However, the analysis is only limited to the *MspI* fragments in the genome. The solution hybrid selection combined with bisulfite sequencing on NGS platforms has clear advantages due to its flexible and expandable probe design as well as the simple sample preparation process which could be fully automated. Multiple captured DNA samples can be bar-coded and sequenced as a pool to meet the increased throughput of NGS.

Several other groups have also developed methods for targeted genome-level bisulfite sequencing (12,14–16). Our assays have several unique features when compared with these methodologies. First, because the solution hybrid selection procedure captures the DNA before bisulfite conversion, the methylation status of the target sequence does not factor into the capture probe design. Second, the solution hybrid-capture approach is highly scalable and flexible with respect to coverage while maintaining the high level of specificity. Finally, the capture probes can be obtained commercially through Agilent's eArray service. One limitation of our current protocol is the requirement of a relatively high amount of DNA for the initial capture reaction. We reason that the following factors might have been involved: (i) the targeted regions are small (0.25% of the human genome), and the probe density is rather low, one probe per 500-bp window. If one probe failed to capture the DNA, there is no nearby probe to compensate. (ii) A significant amount of DNA was lost during the library preparation and bisulfite modification steps. We sequenced the capture DNA from a single-capture reaction before bisulfite treatment and found that the coverage was similar and distributed more evenly than the bisulfite sequencing of pooled DNA from multiple capture reactions (data not shown). We predict that the amount of DNA required can be reduced by designing a larger and more densely distributed probe set, and by further optimization of the library construction (29), hybridization (30) and bisulfite treatment procedures. Carrier DNA can also be added to the bisulfite treatment

Figure 5. Continued

of three breast cancer cell lines using the methylation data of target loci. The target loci were separated into three groups based on functional annotation and analyzed separately. (c) Venn diagram showing the overlap between genes with at least one significantly methylated 5'-end locus (methylation value > 0.75) in MCF10A, MCF7 and MDA-MB-231 cells. (d) DNA methylation and H3K4Me3 profiles in two representative promoters (*RASSF1* and *GATA2*) in MDA-MB-231 cells. The tracks shown from top to bottom are: the DNA methylation level at each CpG site; ChIP-seq results of H3K4Me3 (the track in pink color); RefSeq genes; and annotated CGIs in UCSC genome browser. Red and green colors indicate methylated and unmethylated CpG sites, respectively.

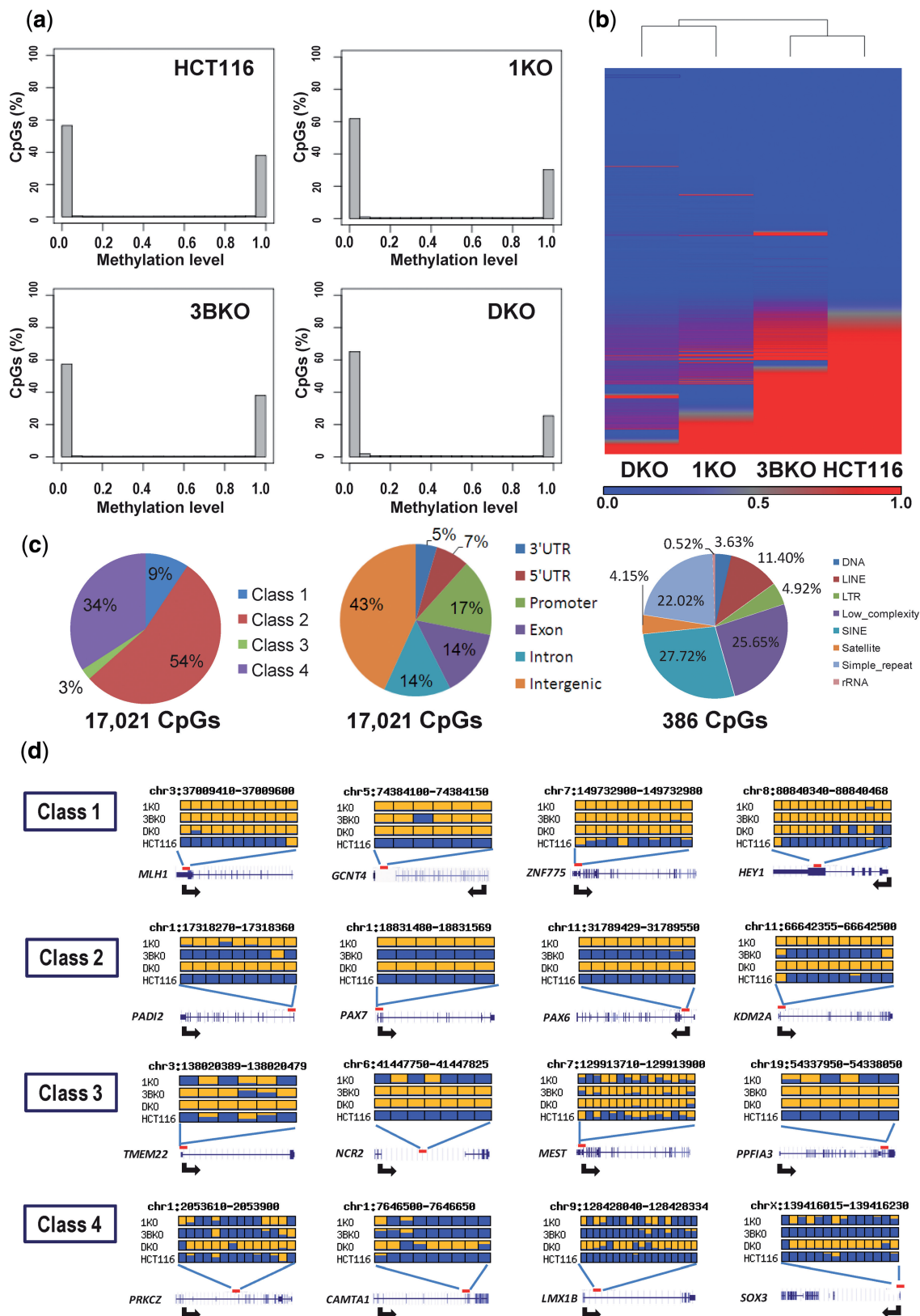


Figure 6. Targeted bisulfite-sequencing analysis of DNMT knockout colon cancer cell lines. (a) The histograms show distribution of CpG methylation levels in DNMT knockout and wild-type HCT116 cell lines. (b) Cluster analysis of 129 022 common CpGs shared by four cell lines. Blue indicates a low level of methylation, while red indicates a high level of methylation. (c) Distribution of hypomethylated CpGs in HCT116 according to the patterns of demethylation among cell lines (left panel); functional genome annotation (middle); and association with repetitive sequences (right panel). (d) Patterns of demethylation in DNMT knockout colon cancer cell lines (1KO, 3BKO and DKO) and HCT116 cell line. Yellow, no methylation; blue, methylation; the definition of each class was indicated in the text.

reaction to reduce the potential loss of DNA during bisulfite treatment and the subsequent purification process (31). To analyze primary tumor samples with a limited amount of DNA, a bar-code could be assigned to multiple tumor samples, the samples could then be pooled together, and finally the capture reaction could be performed.

Using the targeted bisulfite-sequencing approach, we generated single-base resolution methylation maps in thousands of CGI and non-CGI promoters in breast cancer cell lines, MCF10A, MCF7 and MDA-MB-231. Interesting methylation patterns, such as localized shifts between hyper- and hypomethylated states, were observed in some promoters. For instance, although the promoter region of *HOXA9* is methylated in all three cell lines, the first exon is clearly differentially methylated (Figure 1b). The biological significance of this observation is unclear and may deserve further investigation. However, we have shown that DNA methylation may regulate the alternative promoter usage and function in concert with histone methylation (Figure 5d). We mapped a large number of differentially methylated genes including microRNAs (Supplementary Figure S4B). Differential methylation was identified not only in the 5'-end regulatory regions, but also in the intra- and intergenic regions (Figure 5b). Intriguingly, the cluster analyses of DNA methylation in 5'-end regulatory regions, and inter- and intragenic regions all grouped MCF10A and MDA-MB-231 together. This is consistent with the classification based on gene expression, by which MDA-MB-231 and MCF10A were classified as 'Basal B' subtype, while MCF7 was classified as 'Luminal' subtype (32). Although MCF10A is a non-tumorigenic breast epithelial cell line, we have found a large number of methylated genes in this cell line. Interestingly, it seems that DNA methylation patterns in the X-chromosome can distinguish MCF10A from MDA-MB-231 cell line. A significantly higher number of under-methylated loci were observed in MDA-MB-231 cells. Although we could not determine if this observation is due to loss of X-inactivation based on our data, it certainly deserves further investigation in the future.

We further validated the ability of targeted bisulfite sequencing to detect changes in DNA methylation by using DNA from wild-type and DNMT deficient HCT116 colon cancer cell lines. These DNMT deficient cell lines have been used as DNA-demethylation controls in numerous studies (26,27,33,34). The targeted bisulfite sequencing correctly identified a gradual increase in the number of DNA hypomethylation events from the 3BKO, to 1KO, to DKO cell lines. This result is in close agreement with previous findings that utilized HPLC-based global methylation analysis (23,24). The new sequencing results indicated the critical role of DNMT1 in maintaining global DNA methylation, as hypomethylated CpG sites in the DKO cell line were most likely to be hypomethylated in the 1KO cell line as well. We have also identified *DNMT3B*-specific demethylation patterns in a small group of genomic loci (Figure 6d). These candidate loci may be further investigated in the future in order to dissect the role of each individual DNMT enzyme in establishing CpG methylation patterns. In summary, the single-base resolution DNA-methylation maps generated in the

study may become a valuable resource for gene-specific epigenetic studies in these cell lines.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Bill Dynan for critical review of the article and Mr James Wilson for his assistance in editing the article. The DNMT knockout cell lines are generous gifts of Dr Bert Vogelstein.

FUNDING

Funding for open access charge: National Institutes of Health (grants CA134304 to H.S. and DA025779 to H.S. and K.Z., respectively, in part); Department of Defense (grant W81XWH-07-1-0560 to D.X. and H.S., in part). H.S. is a Georgia Cancer Coalition Distinguished Cancer Scientist.

Conflicts of interest statement. E.J.L., L.P., G.S., T.J., G.K., J.H.C., X.W., K.R., J.C., D.X., K.Z., and H.S. declare no conflict of interest. L.Z. and G.P.S. were employed by Illumina Inc. at the time these experiments were carried out.

REFERENCES

- Robertson,K.D. (2005) DNA methylation and human disease. *Nat. Rev. Genet.*, **6**, 597–610.
- Huang,T.H. and Esteller,M. (2010) Chromatin remodeling in mammary gland differentiation and breast tumorigenesis. *Cold Spring Harb. Perspect. Biol.*, **2**, a004515.
- Jones,P.A. and Baylin,S.B. (2002) The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.*, **3**, 415–428.
- Laird,P.W. (2005) Cancer epigenetics. *Hum. Mol. Genet.*, **14**(Spec No. 1), R65–R76.
- Laird,P.W. (2003) The power and the promise of DNA methylation markers. *Nat. Rev. Cancer*, **3**, 253–266.
- Callinan,P.A. and Feinberg,A.P. (2006) The emerging science of epigenomics. *Hum. Mol. Genet.*, **15**(Spec No. 1), R95–R101.
- Taylor,K.H., Shi,H. and Caldwell,C.W. (2010) Next generation sequencing: advances in characterizing the methylome. *Genes*, **1**, 143–165.
- Bibikova,M. and Fan,J.-B. Genome-wide DNA methylation profiling. *Wiley Interdisciplin. Rev. Syst. Biol. Med.*, **2**, 210–223.
- Laird,P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
- Laurent,L., Wong,E., Li,G., Huynh,T., Tsigirgos,A., Ong,C.T., Low,H.M., Kin Sung,K.W., Rigoutsos,I., Loring,J. *et al.* (2010) Dynamic changes in the human methylome during differentiation. *Genome Res.*, **20**, 320–331.
- Lister,R., Pelizzola,M., Downen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.
- Nautiyal,S., Carlton,V.E., Lu,Y., Ireland,J.S., Flaucher,D., Moorhead,M., Gray,J.W., Spellman,P., Mindrinos,M., Berg,P. *et al.* (2010) High-throughput method for analyzing methylation of CpGs in targeted genomic regions. *Proc. Natl Acad. Sci. USA*, **107**, 12587–12592.
- Teer,J.K., Bonnycastle,L.L., Chines,P.S., Hansen,N.F., Aoyama,N., Swift,A.J., Abaan,H.O., Albert,T.J., Margulies,E.H., Green,E.D. *et al.* (2010) Systematic comparison of three genomic

- enrichment methods for massively parallel DNA sequencing. *Genome Res.*, **20**, 1420–1431.
14. Ball, M.P., Li, J.B., Gao, Y., Lee, J.H., LeProust, E.M., Park, I.H., Xie, B., Daley, G.Q. and Church, G.M. (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.*, **27**, 361–368.
 15. Deng, J., Shoemaker, R., Xie, B., Gore, A., LeProust, E.M., Antosiewicz-Bourget, J., Egli, D., Maherali, N., Park, I.H., Yu, J. *et al.* (2009) Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat. Biotechnol.*, **27**, 353–360.
 16. Hodges, E., Smith, A.D., Kendall, J., Xuan, Z., Ravi, K., Rooks, M., Zhang, M.Q., Ye, K., Bhattacharjee, A., Brizuela, L. *et al.* (2009) High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Res.*, **19**, 1593–1605.
 17. Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, **27**, 182–189.
 18. Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F. *et al.* (2007) Multiplex amplification of large sets of human exons. *Nat. Methods*, **4**, 931–936.
 19. Srivastava, G.P., Guo, J., Shi, H. and Xu, D. (2008) PRIMEGENS-v2: genome-wide primer design for analyzing DNA methylation patterns of CpG islands. *Bioinformatics*, **24**, 1837–1842.
 20. Zhang, Y., Liu, T., Meyer, C., Eeckhoutte, J., Johnson, D., Bernstein, B., Nussbaum, C., Myers, R., Brown, M., Li, W. *et al.* (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
 21. Srivastava, G.P., Kushwaha, G., Shi, H. and Xu, D. (2010) In Fung, G. (ed.), *A Practical Guide to Bioinformatics Analysis*. iConcept Press Ltd, Annerley, Australia.
 22. Yan, P.S., Shi, H., Rahmatpanah, F., Hsiau, T.H., Hsiau, A.H., Leu, Y.W., Liu, J.C. and Huang, T.H. (2003) Differential distribution of DNA methylation within the RASSF1A CpG island in breast cancer. *Cancer Res.*, **63**, 6178–6186.
 23. 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.
 24. Rhee, I., Jair, K.W., Yen, R.W., Lengauer, C., Herman, J.G., Kinzler, K.W., Vogelstein, B., Baylin, S.B. and Schuebel, K.E. (2000) CpG methylation is maintained in human cancer cells lacking DNMT1. *Nature*, **404**, 1003–1007.
 25. Egger, G., Jeong, S., Escobar, S.G., Cortez, C.C., Li, T.W., Saito, Y., Yoo, C.B., Jones, P.A. and Liang, G. (2006) Identification of DNMT1 (DNA methyltransferase 1) hypomorphs in somatic knockouts suggests an essential role for DNMT1 in cell survival. *Proc. Natl Acad. Sci. USA*, **103**, 14080–14085.
 26. Paz, M.F., Wei, S., Cigudosa, J.C., Rodriguez-Perales, S., Peinado, M.A., Huang, T.H. and Esteller, M. (2003) Genetic unmasking of epigenetically silenced tumor suppressor genes in colon cancer cells deficient in DNA methyltransferases. *Hum. Mol. Genet.*, **12**, 2209–2219.
 27. Sandoval, J., Heyn, H.A., Moran, S., Serra-Musach, J., Pujana, M.A., Bibikova, M. and Esteller, M. (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, **6**, 692–702.
 28. Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nussbaum, C., Jaffe, D.B. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.
 29. Adey, A., Morrison, H., Asan, X., Xun, X., Kitzman, J., Turner, E., Stackhouse, B., MacKenzie, A., Caruccio, N., Zhang, X. *et al.* (2011) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.*, **11**, R119.
 30. Fisher, S., Barry, A., Abreu, J., Minie, B., Delorey, T., Fennell, T., Allen, A., Ambrogio, L., Berlin, A., Blumenstiel, B. *et al.* (2011) A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.*, **12**, R1.
 31. Gu, H., Bock, C., Mikkelsen, T.S., Jager, N., Smith, Z.D., Tomazou, E., Gnirke, A., Lander, E.S. and Meissner, A. (2010) Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat. Methods*, **7**, 133–136.
 32. Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.P., Tong, F. *et al.* (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, **10**, 515–527.
 33. Gius, D., Cui, H., Bradbury, C.M., Cook, J., Smart, D.K., Zhao, S., Young, L., Brandenburg, S.A., Hu, Y., Bisht, K.S. *et al.* (2004) Distinct effects on gene expression of chemical and genetic manipulation of the cancer epigenome revealed by a multimodality approach. *Cancer Cell*, **6**, 361–371.
 34. Lujambio, A., Ropero, S., Ballestar, E., Fraga, M.F., Cerrato, C., Setien, F., Casado, S., Suarez-Gauthier, A., Sanchez-Cespedes, M., Git, A. *et al.* (2007) Genetic unmasking of an epigenetically silenced microRNA in human cancer cells. *Cancer Res.*, **67**, 1424–1429.