Published online October 13, 2005

5868–5877 *Nucleic Acids Research, 2005, Vol. 33, No. 18*
doi:10.1093/nar/gki901

Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis

Alexander Meissner[1], Andreas Gnirke[2], George W. Bell[1], Bernard Ramsahoye[3], Eric S. Lander[1,2] and Rudolf Jaenisch[1,*]

[1]Whitehead Institute for Biomedical Research and Massachusetts Institute of Technology, Nine Cambridge Center, Cambridge, MA 02142, USA, [2]Broad Institute of MIT and Harvard, Cambridge, MA, USA and [3]John Hughes Bennett Laboratory, Division of Oncology, School of Molecular and Clinical Medicine, University of Edinburgh, Western General Hospital, Edinburgh, UK

Received August 23, 2005; Revised and Accepted September 28, 2005

ABSTRACT

We describe a large-scale random approach termed reduced representation bisulfite sequencing (RRBS) for analyzing and comparing genomic methylation patterns. BglII restriction fragments were size-selected to 500–600 bp, equipped with adapters, treated with bisulfite, PCR amplified, cloned and sequenced. We constructed RRBS libraries from murine ES cells and from ES cells lacking DNA methyltransferases Dnmt3a and 3b and with knocked-down (kd) levels of Dnmt1 (*Dnmt*[1^{kd},$3a^{-/-}$, $3b^{-/-}$]). Sequencing of 960 RRBS clones from *Dnmt*[1^{kd},$3a^{-/-}$,$3b^{-/-}$] cells generated 343 kb of non-redundant bisulfite sequence covering 66 212 cytosines in the genome. All but 38 cytosines had been converted to uracil indicating a conversion rate of >99.9%. Of the remaining cytosines 35 were found in CpG and 3 in CpT dinucleotides. Non-CpG methylation was >250-fold reduced compared with wild-type ES cells, consistent with a role for Dnmt3a and/or Dnmt3b in CpA and CpT methylation. Closer inspection revealed neither a consensus sequence around the methylated sites nor evidence for clustering of residual methylation in the genome. Our findings indicate random loss rather than specific maintenance of methylation in *Dnmt*[1^{kd},$3a^{-/-}$,$3b^{-/-}$] cells. Near-complete bisulfite conversion and largely unbiased representation of RRBS libraries suggest that random shotgun bisulfite sequencing can be scaled to a genome-wide approach.

INTRODUCTION

DNA methylation is a key epigenetic modification that provides heritable information not encoded in the nucleotide sequence. 5-Methylcytosine is the only known covalent modification of DNA in vertebrates (1). Mammalian DNA methylation serves a wide range of functions including regulation of gene expression, genomic imprinting and X-chromosome inactivation. It contributes to genomic stability and serves as a defense mechanism against transposable elements (2–5). In addition, its role in disease states such as cancer becomes increasingly evident (6–10).

Three catalytically active DNA methyltransferases (Dnmts) have been described that are responsible for establishing and maintaining methylation patterns in mammals (4,11–13). Dnmt1 has been largely viewed as the maintenance enzyme, owing to its preference for hemimethylated DNA (2). Dnmt3a and Dnmt3b have no preference and are required for *de novo* methylation activity (14). During murine preimplantation development methylation levels decrease with some notable exceptions including imprinted genes and IAP elements (3,15). Around the time of implantation normal methylation levels are restored by the *de novo* methyltransferases and later maintained by Dnmt1.

Targeted gene disruption for each of the catalytically active Dnmts (1, 3a and 3b) results in a lethal phenotype demonstrating the essential role of DNA methylation in development (11,13). Interestingly, undifferentiated ES cells deficient for Dnmt1, Dnmt3a, Dnmt3b or Dnmt3a/3b do not display any obvious abnormalities (13,16). Normally in wild-type ES cells most CpG dinucleotides are methylated with the exception of many CpG islands.

The intense interest in the biological functions of DNA methylation and its role in diseases have led to numerous

[*]To whom correspondence should be addressed. Tel: +1 617 258 5186; Fax: +1 617 258 6505; Email: jaenisch@wi.mit.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors techniques to detect and compare DNA methylation [reviewed in (8,17)]. Global methods such as nearest neighbor analysis (NNA) and high-performance liquid chromatography are valuable to quantify the total 5-methylcytosine content of a DNA sample, but information on the position in the genome cannot be gained (18,19).

Digestion with methylation-sensitive (or methylation-dependent) restriction enzymes (MSREs) has been used to selectively enrich the methylated and unmethylated DNA fractions, respectively (20–24). Similarly, methylation-dependent restriction in a cloning host has been employed as a filter against methylation-rich sequences in clone libraries (25). Another, more recent genome-wide approach used immunoprecipitation with a methyl cytosine antibody rather than restriction digestion to enrich for the methylated fraction (26). The enriched genome fractions are analyzed by sequencing or by array-hybridization (20,21,26). MSRE-based methods are somewhat indirect in that they discriminate for or against methylation at the recognition site of the particular enzyme used and cannot directly reveal the methylation status of cytosines or CpG dinucleotides outside the restriction site.

In contrast, methylation-sensitive chemical reactions have no specific recognition sequence. Sodium bisulfite efficiently deaminates unmethylated cytosine to uracil without affecting 5-methyl cytosine. In recent years, PCR amplification and sequencing of bisulfite-converted genomic DNA has emerged as the gold standard for analyzing and comparing methylation patterns at specific loci (27).

Despite these technological advances, in the absence of systematic sequence-based methylation analyses, the genomic methylation landscape in mammals is still largely unexplored. Therefore, the potential diagnostic value of specific methylation differences remains largely untapped.

The human epigenome project (HEP) is aimed at generating a high-resolution DNA methylation map of the human genome (28,29). To achieve this goal the bisulfite sequencing technique has been scaled-up in a targeted fashion using locus-specific PCR primers.

Here we describe a random approach for large-scale high-resolution DNA methylation analysis termed reduced representation bisulfite sequencing (RRBS). To test the feasibility of the method, we compared wild-type ES cells and ES cells deficient for Dnmt1, Dnmt3a and Dnmt3b. Our data suggest that RRBS provides high-quality data suitable for future large-scale comparative epigenetic studies of DNA methylation in a given cell type or tissue. In addition our sequencing data confirm and complement previous studies on the role of DNA methyltransferases in murine ES cells.

METHODS

RRBS library construction and sequencing

Mouse ES DNA (50–100 μg) was digested to completion by overnight incubation with 1000 U of BglII and electrophoresed on a 1.8% agarose gel. Marker lanes were stained with SYBR Green (Invitrogen). A narrow slice containing the 500–600 bp fraction was excised from the unstained preparative portion of the gel. DNA was recovered by electroelution, phenol extraction and ethanol precipitation as described elsewhere (30). Typical yields were 300–600 ng of size-selected BglII fragments as measured by PicoGreen fluorescence (Invitrogen). The size-selected BglII fragments (1–2 pmol) were ligated to 700 pmol BglII adapter pre-annealed from oligodeoxynucleotides 5′-AGTTATTCCGGACTGTCGAAGCTGAATGCCATGG-3′ and 5′-pGATCCCATGGCATTCAGCTTCGACAGTCCGGAAT-3′ in 70 μl containing 2400 U T4 DNA ligase (New England Biolabs) for 16 h at 14°C. Excess adapter was removed by ultrafiltration (Millipore Montage) followed by preparative electrophoresis in 2% agarose and electroelution, yielding 50–100 ng of adapter-ligated material.

Adapter-ligated, size-selected BglII fragments (50 ng) were bisulfite-treated using the reagents and protocol of the CpGenome DNA modification kit (Chemicon) with the following modifications: the DNA was alkali-denatured for 20 min at 55°C; the total reaction volume was increased from 650 to 750 μl and contained 0.22 g urea (31); and the mixture was incubated for 24 h at 55°C. After alkaline desulfonation and final desalting, single-stranded uracil-containing reaction products were eluted in 40 μl of TE buffer and converted to double-stranded DNA by PCR with primers 5′-TTGGATTGTTGAAGTTGAATG-3′ and 5′-AAACTATCAAAACTAAATACCATAAAATC-3′ designed to amplify molecules carrying bisulfite-modified adapter sequences at both ends. For each bisulfite reaction, eight 50 μl PCRs were performed, each containing 2.5 μl bisulfite-treated DNA, 25 pmol of each PCR primer and 2.5 U PfuTurboCx Hotstart DNA polymerase (Stratagene). Thermocycling included eight cycles of ‘touchdown’ (32) at annealing temperatures from 55 to 52°C (two cycles at each temperature) followed by 10 cycles at an annealing temperature of 51°C. Denaturation (94°C), annealing and extension (72°C) times were 10 s, 30 s and 3 min, respectively. PCR products were cleaned-up by ultrafiltration followed by preparative electrophoresis on a 2% agarose gel. Typical yields were between 50 and 100 ng for each library. Gel-purified PCR product (4 ng) was incubated for 5 min with 1 μl pCR BluntII TOPO vector and cloned by electroporation of *Escherichia coli* TOP10 (Invitrogen). The cloning efficiency was ~2000 colonies per ng of PCR product.

Plasmid DNA was isolated by standard protocols, and cloned inserts were sequenced using 2.7 pmol M13 reverse primer and 2 μl BigDye3.1 mix (Applied Biosystems) in 10 μl sequencing reactions (25 cycles). Caused by preferential cloning in one orientation, ~80% of the sequences were the G-poor strand. Most inserts that had been cloned in the other orientation (C-poor strand) sequenced poorly, with peak heights and sequence quality suddenly dropping after 300–400 bases.

Data analysis

In silico digestion of the mouse genome (NCBI Build 33, May 2004) was performed at BglII sites, followed by selection of fragments ranging from 440 to 640 bases. Cytosines were converted to thymine, with upper/lower case used to differentiate converted from original thymines. Each strand was converted separately. Sequencing reads were mapped to the genome by using NCBI BLAST (without query filtering) to search the database of size-selected and converted BglII fragments. The top BLAST hit determined the most probable genome location of each read and also permitted identification of original and converted cytosines over the high-scoring segment pair length. The repeat content of the *in silico* reduced representation and the sets of sequencing reads were compared with that of the whole genome using RepeatMasker (http://www.repeatmasker.org). Locations of all sequence reads relative to selected genomic landmarks were determined by comparing fragment coordinates to those of the RefSeq and Ensembl transcript sets and CpG islands from UCSC.

The expected number of redundant RRBS sequences and the sequence overlap between two DNA samples were calculated by composite Poisson statistics in 5 bp bins across the range of insert sizes. D_i is the number of BglII fragments in the reference genome that fall into bin i. N_{ai} is the number of successful sequencing reads from DNA sample a that fall into bin i. $\sum\left(N_{ai}^2/2D_i\right)$ double-hits in sample a are expected by random sampling (33). The expected number of BglII fragments sequenced at least once in sample a and in sample b is $\sum\left[\left(1-\mathrm{e}^{-N_{ai}/D_i}\right)\times\left(1-\mathrm{e}^{-N_{bi}/D_i}\right)\times D_i\right]$.

ES cell manipulation

Lentiviral infections of ES cells were performed as described previously (34). ES cells were cultivated on irradiated mouse embryonic fibroblasts in DME containing 15% fetal calf serum, leukemia inhibiting factor, penicillin/streptomycin, L-glutamine and non-essential amino acids. All ES cells were depleted of feeder cells for two passages on 0.2% gelatine before isolating DNA.

Southern blot and methylation analysis

To assess the levels of DNA methylation, genomic DNA was digested with HpaII, and hybridized to pMR150 as a probe for the minor satellite repeats (35), or with an IAP-probe (36). For the methylation status of imprinted genes, a combined bisulfite restriction analysis (COBRA) assay was performed with the CpGenome DNA modification kit (Chemicon) using PCR primers and conditions described previously (37). PCR products were gel purified, digested with BstUI or HpyCH4 IV and resolved on a 2% agarose gel. NNA was done as described previously (19).

RESULTS

Reduced representation bisulfite sequencing

RRBS is analogous to the reduced representation shotgun sequencing used for single nucleotide polymorphism (SNP) discovery (33). The method is based on size selection of restriction fragments to generate a 'reduced representation' of the genome of a strain, tissue or cell type.

For this study, we digested genomic DNA with BglII and purified fragments between 500 and 600 bp in size on an agarose gel. Based on the available mouse genome sequence, BglII digestion is expected to generate 21 939 BglII fragments in this size range comprising ~12 Mb (0.5%) of the genome. Size-selected BglII fragments were equipped with end adapters, denatured and treated with bisulfite to convert all unmethylated cytosines to uracil. Bisulfite-converted DNA remains single-stranded as the two strands are no longer complementary. Primers specific for the converted adapter sequence and a proofreading thermostable DNA polymerase were used to synthesize the second strand and to PCR amplify the bisulfite-converted material. Blunt-end PCR products were cloned in a plasmid vector and sequenced (Figure 1).

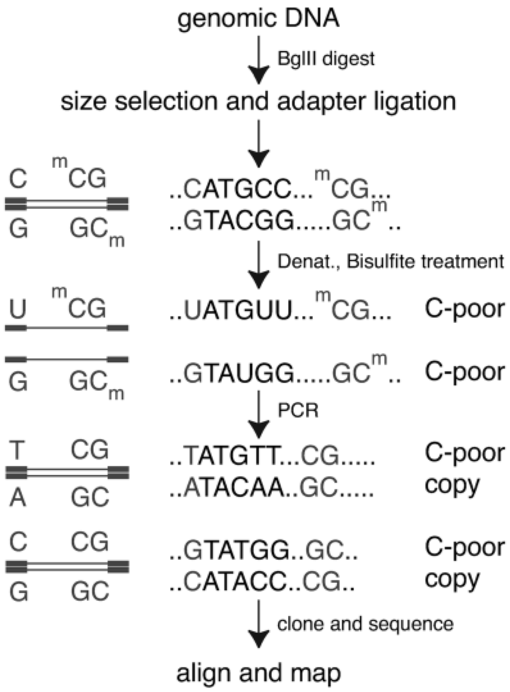

Figure 1. Reduced representation bisulfite sequencing. Genomic DNA is digested to completion using a restriction enzyme (here BglII). After size-selection an adapter is added. The DNA is denatured, and unmethylated cytosines are bisulfite-converted to uracil. The two resulting C-poor strands are no longer complementary to each other. Primers specific for the converted adapter sequence are used to fill-in the second (G-poor) strand and for PCR amplification. PCR products are cloned and sequenced. Sequences generated from RRBS libraries are projected onto the genome by searching against a reduced representation database of BglII fragments that had been size-selected and bisulfite-converted *in silico*.

For analysis of the bisulfite sequences and to identify the corresponding genomic sequence we searched RRBS reads against a reduced representation database of the mouse genome that contained both strands of BglII fragments that had been size-selected and bisulfite-converted *in silico*. When aligned to the original genome sequence, a 5-methylcytosine is thus displayed as a matching C in the bisulfite sequence, and C to T transitions indicate unmethylated cytosines.

Even though bisulfite sequencing is a widespread technique, some concerns persist. Since bisulfite converts single-stranded but not double-stranded DNA, incomplete denaturation or re-annealing leads to incomplete conversion. This complicates the data analysis, as it is not always possible to determine whether an unconverted cytosine represents bona fide methylation or an experimental artifact.

Another potential problem is depurination, strand breakage and DNA degradation caused by the harsh reaction conditions, which lower the yield of full-length BglII fragments significantly. It has been estimated that >90% of the input DNA is lost to DNA degradation during the first hour of a bisulfite reaction (38). However, to maximize the conversion rate, the reaction is usually carried out overnight, necessitating extensive PCR amplification before cloning or sequencing to compensate

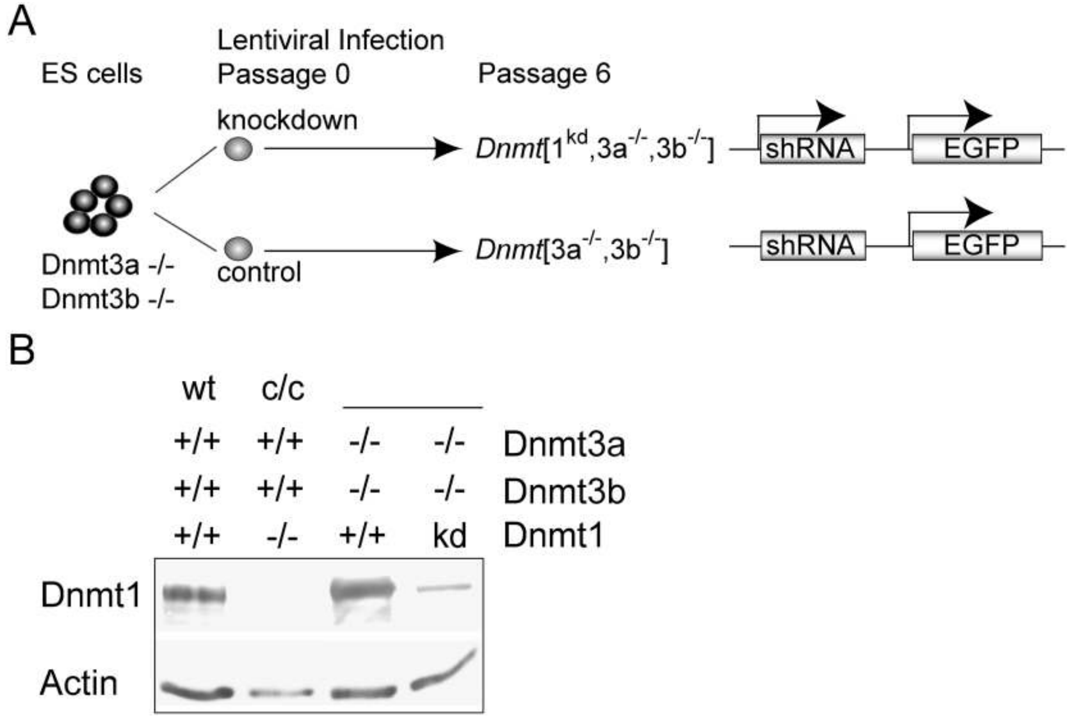

Figure 2. Generation of Dnmt1, Dnmt3a and Dnmt3b deficient ES cells. (**A**) *Dnmt*3a/3b homozygous double knockout ES cells have been described earlier (43). The knockdown virus is expressing a *Dnmt*1 shRNA, whereas the control is not. The infection was termed Passage 0. After the infection ES cells were passaged four times on feeders followed by two additional passages under feeder-free conditions (Passage 6). Number of viral integrations were determined by Southern blotting and clones with single integration were selected (data not shown). (**B**) Western blot analysis. The status of the different Dnmts is indicated above. The knockdown ES cells showed a significant reduction in Dnmt1 levels compared with their sister clone. c/c is a previously reported *Dnmt*1 null ES line (16).

for the inevitable loss of DNA. Moreover, since most proofreading enzymes stall at uracil residues in the template strand, non-proofreading *Taq* polymerase is usually prescribed for second-strand synthesis and PCR amplification which can lead to PCR-induced sequencing errors.

These limitations are less worrisome for single-copy loci, but could be significant in a genome-wide setting, where no preselection against fast-re-annealing repetitive sequences is made and where amplification bias and skewed sequence representation creates serious sampling problems.

Indeed, our preliminary attempts were plagued by DNA degradation, incomplete conversion and poor efficiency of PCR amplification, most probably caused by the re-annealing of repetitive sequences including the common adapter sequence at the ends of each DNA molecule. Moreover, certain sequences were clearly overrepresented in the resulting libraries indicating amplification bias during the PCR. These initial problems were largely remedied by performing the bisulfite reaction in the presence of urea as suggested by Paulin *et al*. (31) and by fine-tuning experimental parameters such as DNA concentration, time and temperature of the bisulfite reaction, and number of PCR cycles for the double-strand rescue and amplification by a proofreading thermostable DNA polymerase engineered to accept uracil in the template strand (39).

To test if our optimized protocol was sufficient to achieve complete genome-wide bisulfite conversion without compromising library complexity and representation, we wished to construct and sequence RRBS libraries from genomic DNA that was largely free of methylation. To this end we generated ES cells deficient in all three major DNA methyltransferases.

ES cells deficient for Dnmt1, Dnmt3a and Dnmt3b

We combined knockouts for the *de novo* Dnmts (Dnmt3a and Dnmt3b) with RNAi-induced knockdown of Dnmt1 (Figure 2A) using a lentivirus-based system for stable short hairpin RNA (shRNA) expression (34). The Dnmt1 knockdown resulted in a significant albeit not complete loss of Dnmt1 protein compared with the *Dnmt*[$3a^{-/-}$,$3b^{-/-}$] control cells (Figure 2B).

To determine whether the decrease in Dnmt1 levels led to efficient demethylation, we analyzed the methylation status of minor satellite repeats and IAP elements in a number of control and knockdown ES cell lines by MSRE analysis. Significant repeat demethylation was observed when Dnmt1 was knocked down, and the methylation levels in the *Dnmt*[1^{kd},$3a^{-/-}$,$3b^{-/-}$] ES cells closely resembled the digest of genomic DNA with *Msp*I which cuts irrespective of the methylation status (Figure 3A and B). Loss of methylation at these repeat elements appears to be primarily caused by the lack of Dnmt1 and largely independent of the *de novo* Dnmts at these early passages.

Using a COBRA assay (40) we observed loss of imprinting at four imprinted genes following Dnmt1 knockdown as compared with the controls (Figure 3C). Taken together, these experiments showed that Dnmt1 knockdown resulted in significant loss of methylation at specific genes and repeat elements.

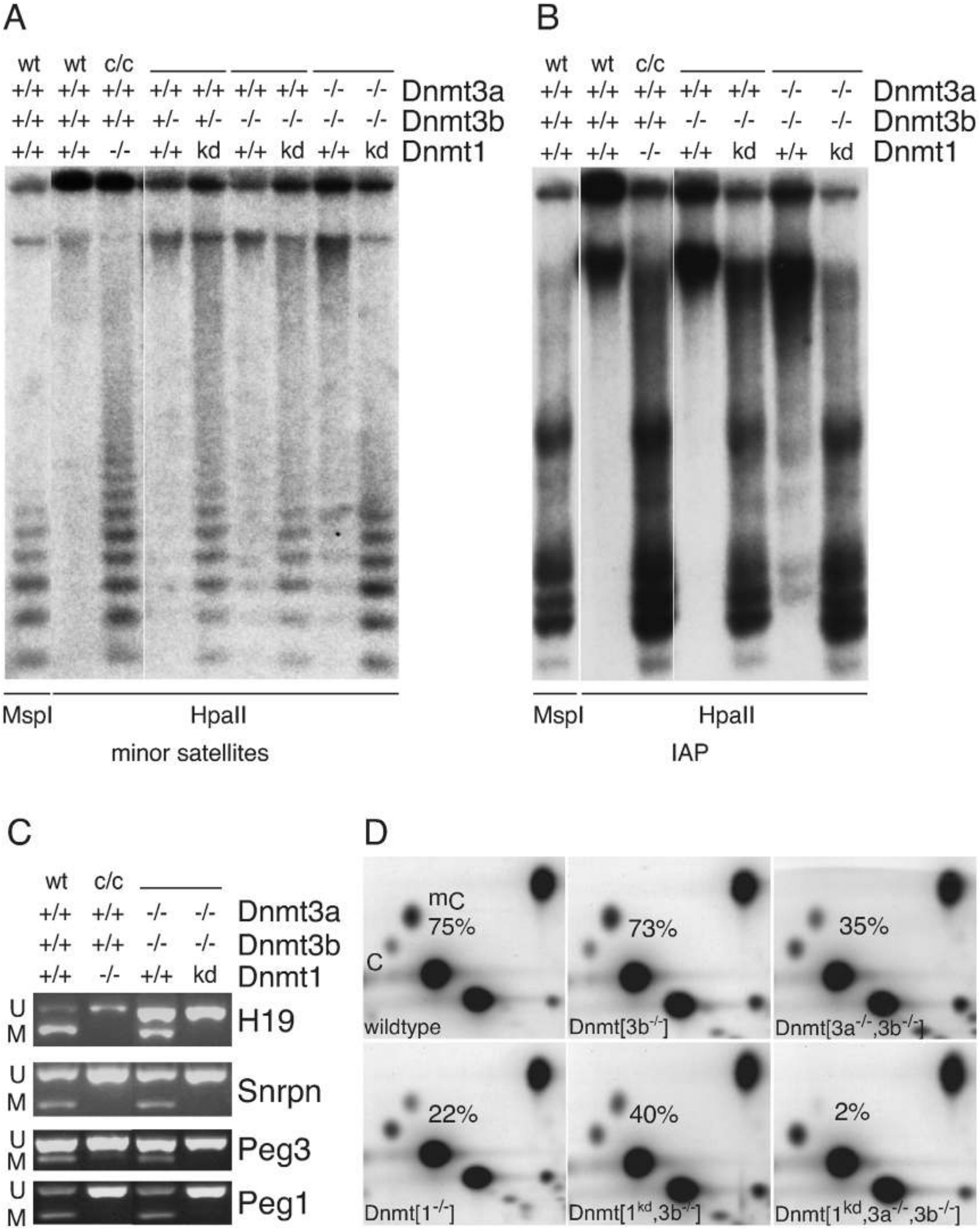

Figure 3. Methylation status of the Dnmt-deficient ES cells. All knockdown and control ES cells were analyzed at Passage 6 after infection. (**A**) Minor satellite repeat methylation. HpaII digests of genomic DNA were hybridized to minor satellite probe pMR150. MspI is an isoschizomere of HpaII and cuts irrespective of the methylation status (i.e. appearance of a ladder in HpaII lane indicates loss of methylation). The status of the different Dnmts is shown above the Southern blot. All knockdown and control ES cell lines were generated as described in Figure 2. Each knockdown line contains a single lentiviral integration (data not shown). (**B**) IAP methylation. HpaII-digested genomic DNA was hybridized to an IAP probe. (**C**) COBRA analysis for imprinted genes. Genomic DNA was bisulfite treated and after PCR amplification of H19, Snrpn, Peg1 and Peg3 a restriction digest was performed to analyze the methylation status of the differentially methylated regions (U = unmethylated, M = methylated). The second (smaller) fragment of the methylated and digest product is not shown. (**D**) Total mCpG quantification by NNA. The spots corresponding to CpG and mCpG are indicated in the upper left panel. The per cent mCpG/(CpG+mCpG) are displayed in each panel (estimated error 5%).

To better quantify these results, we used NNA, which allowed to determine the global amounts of CpG methylation in wild-type and mutant ES cells. We detected ∼2% residual CpG methylation in the *Dnmt*[1^{kd},$3a^{-/-}$,$3b^{-/-}$] cells compared with 22% in the *Dnmt*1 null ES cells and 75% in wild-type ES cells (Figure 3D). Dnmt3b heterozygous and homozygous ES cells displayed wild-type methylation levels in the presence of Dnmt1 and showed similar loss of methylation within six passages of Dnmt1 knockdown (Figure 3D and data not shown) confirming the potency of the shRNA.

To test the RRBS approach and to determine whether specific sequences were retaining methylation we generated and sequenced BglII RRBS libraries from wild-type and Dnmt-deficient ES cells.

Sequencing of RRBS libraries

In preliminary experiments we noticed that sequencing RRBS clones with reverse primer had a significantly higher success rate and produced longer reads on average than sequencing with forward primer. We therefore sequenced the RRBS clones single-pass using reverse primer. Only clones with high-quality sequence across the entire length of the insert were used for the final methylation analysis. Table 1 summarizes the sequencing statistics from 960 RRBS clones from Dnmt-deficient cells and 192 clones from wild-type ES cells.

Although blunt-ended PCR products can insert in either orientation into the cloning vector, only a minority had inserts in the orientation that resulted in the C-poor sequence, i.e. the strand that has been modified by bisulfite (153 out of 876 RRBS reads from the *Dnmt*[1^{kd},$3a^{-/-}$,$3b^{-/-}$] library). The vast majority of the clones produced the complementary G-poor reads. Notably, the sequence quality was also significantly different for the two orientations. Almost all G-poor reads were high-quality across the entire insert whereas peak heights and quality of many C-poor reads dropped after a few hundred bases, leaving relatively few complete C-poor sequences for the methylation analysis. Preferential cloning in one orientation and high drop-out rate for C-poor strands were more pronounced in the *Dnmt*[1^{kd},$3a^{-/-}$,$3b^{-/-}$] library which has an extremely asymmetric base distribution. Of the sequenced inserts 96% from this library consisted solely of three bases, i.e. either A, G and T or A, C and T owing to complete absence of methylated cytosine in the corresponding genome loci. Directional cloning and sequencing bias has been observed before with bisulfite-treated DNA (38) and is therefore not a RRBS specific phenomenon.

Table 1. Sequencing and methylation statistics

ES cell line	Dnmt[1^{kd}, $3a^{-/-}$,$3b^{-/-}$]	wild-type
Colonies picked	960	192
Bisulfite sequencing reads[a]	876	186
Insert in plus orientation[b]	153	50
Plus read complete[c]	38	23
Insert in minus orientation[d]	723	136
Minus read complete[c]	719	134
Complete bisulfite sequencing reads[c]	757	157
Genome hits	676	148
Non-redundant genome hits[e]	609	147
Total bp of non-redundant genome hits	342 556	80 692
Cytosines in aligned genome sequence	66 212	15 296
5-Methylcytosine (mC)	38 (0.06%)	707 (4.6%)
CpG in aligned genome sequence	3458	594
mCpG	35 (1.0%)	533 (90%)
CpA in aligned genome sequence	23 046	5601
mCpA	0 (0%)	135 (2.4%)
CpT in aligned genome sequence	25 505	5924
mCpT	3 (0.01%)	39 (0.7%)
CpC in aligned genome sequence	14 203	3177
mCpC	0 (0%)	0 (0%)

[a] Excludes growth failures, sequencing failures, mixed clones, vector-only clones and a total of nine reads that showed no bisulfite conversion at all.
[b] Sequenced strand is the bisulfite-converted C-poor strand.
[c] High-quality sequence across entire length of BglII fragment.
[d] Sequenced strand is the G-poor complementary strand of the bisulfite-converted strand.
[e] Includes sequences that are duplicated in the genome. BglII fragments that were hit more than once were counted only once.

Of the complete RRBS reads from *Dnmt*[1^{kd},$3a^{-/-}$,$3b^{-/-}$] cells (89%) found a near-perfect match in the reduced representation reference-sequence database and could be placed with high confidence on the mouse genome. The rate of genome alignments for sequences from wild-type ES was slightly higher (94%). Overall, the success rate of full-length, mapped bisulfite sequence was 72% of all clones picked. A schematic of the distribution of RRBS sequences along the mouse chromosomes is available in the Supplementary Data (Supplementary Figures S1 and S2). In addition we have developed a genome browser that allows a more comprehensive view of the genomic environment of the RRBS libraries and the data generated (for a sample screenshot see Supplementary Figure 3).

Fifty-six loci were hit by more than one RRBS sequence from the *Dnmt*[1^{kd},$3a^{-/-}$,$3b^{-/-}$] library. Ten of these potentially represent sequences that occur more than once in the genome. The remaining 46 appear to be unique loci that have indeed been cloned and sequenced twice. This is more than the 23 double-hits expected by random sampling of an ideal library, possibly indicating a slight cloning or sequencing bias. Consistent with random cloning, the much smaller number of wild-type RRBS sequences produced only one double-hit. Eleven fragments were sequenced in both cell lines, compared with eight sequence overlaps expected given the number and size distribution of successful reads from each library (Figure 4). The total length of non-redundant and mapped RRBS sequences was 342 556 bp for *Dnmt*[1^{kd},$3a^{-/-}$,$3b^{-/-}$] and 80 692 bp for wild-type ES cells.

To determine whether these RRBS libraries were generally representative we compared the GC content, the representation

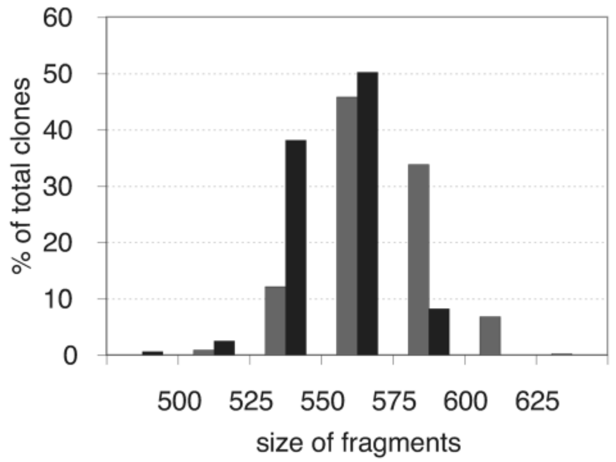

Figure 4. Size distributions of the sequenced clones from each library. RRBS reads from wild-type ES cells (black) had a mean of 553 bp and an SD of 17 bp. *Dnmt*[1^{kd},$3a^{-/-}$,$3b^{-/-}$] reads were (570 ± 20) bp in size (grey bars). The size distributions of the two libraries were overlapping but not identical.

of CpG islands, transcripts, promoter regions and different classes of repeat elements between the entire mouse genome (41), the 500–600 bp BglII fraction thereof and the genome sequences hit by the RRBS clones (Table 2). While reducing the representation introduced a noticeable bias, in particular a reduction of repeats, bisulfite conversion, PCR amplification, cloning and sequencing did not. The GC content of loci covered by RRBS sequences ranged from 32 to 63%, indicating satisfactory performance of our protocol over a wide range of GC content. Likewise, the distribution of the sequenced clones in the genome did not show conspicuous hot or cold spots (Supplementary Figures S1 and S2). Taken together, our data suggest that RRBS libraries are sufficiently random and representative of the genome fraction used to make them.

Reducing the complexity by size fractionation of a limit digest with BglII (recognition site AGATCT) is expected to bias somewhat against GC-rich regions of the genome. Pooling two single digests with compatible enzymes such as BglII and BamHI (GGATCC) before the size selection would sample the genome more evenly and increase the complexity of the RRBS libraries.

Table 2. Fraction (in per cent) of various types of sequences in the mouse reference genome, the 500–600 bp BglII reduced representation thereof (RR genome) and RRBS sequences from Dnmt-deficient and wild-type ES cells

	Genome[a]	RR genome	RRBS *Dnmt*	RRBS *wt*
GC content	42.0	41.5	43.7	43.1
CpG islands[b]	0.4	0.1	0.1	0.0
ENSEMBL genes	34.3[c]	35.0[c]	41.9[d]	35.3[d]
Promoter	5.0[e]	5.0[e]	7.0[f]	4.7[f]
SINEs	8.2	2.7	2.6	2.3
LINEs	19.2	10.2	10.7	11.3
LTR elements	9.9	2.9	2.9	4.3
MER DNA elements	0.9	0.2	0.1	0.2

[a] Repeat and GC content were taken from Ref. (41).

[b] CpG islands were taken from the mm6 mouse genome assembly on the UCSC genome browser.

[c] Fraction of genome sequence that falls within gene bounds of non-overlapping ENSEMBL gene models.

[d] Fraction of RRBS sequences with significant hits to the ENSEMBL gene fraction of the genome.

[e] Fraction of genome sequence that falls within 5 kb upstream of the transcription start site of ENSEMBL gene models.

[f] Fraction of RRBS sequences with significant hits to regions 5 kb upstream of transcription start sites.

Comparison of wild-type and Dnmt-deficient ES cells

The RRBS sequences revealed the methylation status of 66 212 cytosines in *Dnmt*[1^{kd},3a$^{-/-}$,3b$^{-/-}$] ES cells (Table 1, bottom half). Only 38 of these were inferred to be methylated, 35 of them in CpG and three in CpT dinucleotide context. Considering the non-random distribution of mC among the four dinucleotides, it unlikely that all of them were caused by incomplete bisulfite conversion or PCR or sequencing errors. Moreover, the 35 mCpGs are ~1% of all bisulfite-sequenced CpGs, which is close to the 2% mCpG level determined by NNA (Figure 3D). By comparison, 90% of CpGs were methylated in wild-type ES cells. We also observed a considerable difference in the level of non-CpG methylation [(mCpA+mCpT)/C], which was >250-fold reduced in the Dnmt-deficient ES cells.

In the *Dnmt*[1^{kd},3a$^{-/-}$,3b$^{-/-}$] RRBS sequences, 25 020 bases were covered 2- or 3-fold, comprising 4669 cytosines including 217 CpGs. Overlapping RRBS sequences agreed for most loci. In two cases, only one sequenced *Dnmt*[1^{kd},3a$^{-/-}$,3b$^{-/-}$] clone displayed a methylcytosine. At another discordant site, the two reads agreed at one mCpG but disagreed at another.

To address the issue of heterogeneity, we selected 10 loci with mCpGs and 10 loci without methylation and designed specific PCR primers to bisulfite re-sequence them in a targeted fashion. Multiple clones were sequenced for each locus in wild-type, *Dnmt*[3a$^{-/-}$,3b$^{-/-}$] and the *Dnmt*[1^{kd},3a$^{-/-}$,3b$^{-/-}$] cells. In all but one case, at least one re-sequenced clone matched the previously determined mCpG pattern precisely, and the overall level of methylation for each region was similar in all cases (Figure 5 and data not shown). Thus, as a rule, a single clone from the RRBS library provides a good indication of the general methylation pattern at any given site. This is in line with the predominantly bimodal methylation profiles observed previously [reviewed in Ref. (42)]. For example, >80% of the loci in the HEP survey of the MHC were either hypermethylated or hypomethylated (29).

Four representative examples are shown in Figure 5. For the two loci on chromosome 4 and 15, respectively, all clones,

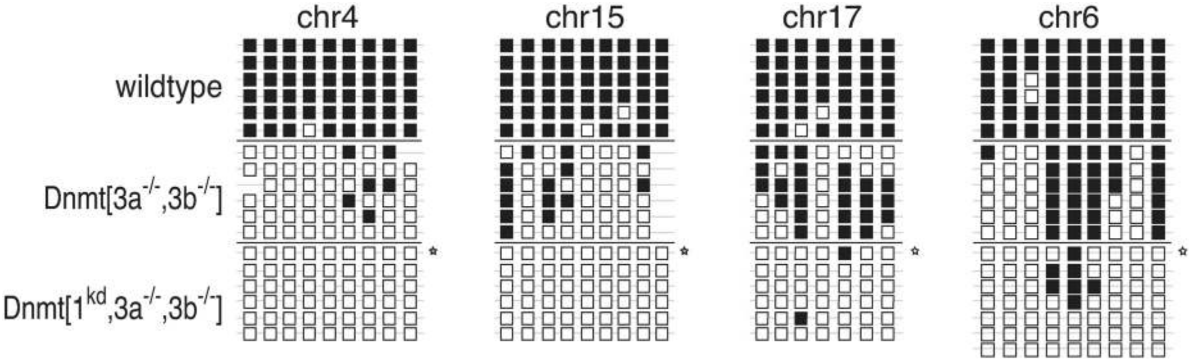

Figure 5. Targeted bisulfite sequencing of specific loci. Ten loci for which RRBS sequencing indicated mCpGs in Dnmt-deficient cells and 10 loci that were devoid of methylation were bisulfite re-sequenced using specific primers in wild-type (top), 3a/b double knockout (middle) and *Dnmt*[1^{kd},3a$^{-/-}$,3b$^{-/-}$] cells (bottom). Shown are two examples of each set. Each row represents a single sequenced molecule. Filled squares are methylated CpGs and empty ones indicate unmethylated sites. The asterisk indicates the original clone sequenced from the library.

including the clone from the RRBS library, indicated complete absence of methylation in *Dnmt*[1^{kd},$3a^{-/-}$,$3b^{-/-}$] cells. The sister cell line with normal Dnmt1 levels (Figure 2) was also considerably demethylated at these sites compared with wild-type ES cells. The two other loci maintained more mCpGs in the methylation-impaired cell lines. The two CpGs on chromosome 17 that were most consistently methylated in *Dnmt*[$3a^{-/-}$,$3b^{-/-}$] cells showed also residual methylation in the *Dnmt*[1^{kd},$3a^{-/-}$,$3b^{-/-}$] cells. One of these two mCpGs was detected in the RRBS clone. Targeted resequencing detected methylation at the second CpG. This pattern is consistent with passive random loss of CpG methylation in *Dnmt*[1^{kd},$3a^{-/-}$,$3b^{-/-}$] cells.

DISCUSSION

Large-scale random bisulfite sequencing

In this study we explored the feasibility of large-scale shotgun bisulfite sequencing for genome-wide analysis of DNA methylation. We have shown that bisulfite sequencing libraries can be made that are largely unbiased and representative and display few false-positive methylcytosines caused by incomplete cytosine to uracil conversion or PCR and sequencing errors.

Insert sizes of the libraries were kept very small (500–600 bp) for two reasons. First, the bisulfite reaction requires relatively high temperatures (50–60°C) and a low pH (pH 5), conditions that are known to cause depurination and strand breakage; smaller molecules are less prone to damage and require fewer PCR cycles to recover intact for cloning than larger ones, thereby minimizing the risk of a skewed representation. Second, larger-insert clones would require sequencing of both strands; however, C-poor strands sequenced poorly in our hands.

We used limit digestion with BglII and size fractionation to reduce the complexity of the DNA. The resulting RRBS libraries cover a small but reproducible fraction of the genome and are therefore suitable for large-scale comparative methylation studies across different strains, tissues or cell types. Based on the overall success rate (72%) and insert-size distributions encountered during this pilot study (Figure 4), we expect that for a pair-wise comparison, sequencing 100 × 384 RRBS clones from each DNA sample will produce 4.0 Mb of high-quality overlapping bisulfite sequence with 2- to 3-fold coverage in each library of fragments within 1 SD of the mean size. Assuming that improvements in sequencing of C-poor strands (85% success rate) and better libraries with congruent insert-size distributions can be made, the same sequencing effort would yield ~5.8 Mb of pair-wise comparative sequence which, of course, is still only a tiny fraction of the genome.

At this level of genome coverage, differential methylation at most individual sites in the genome, including many functionally important ones, will escape detection. However, we expect the coverage to be sufficient to generate methylation variable position markers for future bisulfite SNP 'epigenotyping' (17). A genome-wide set of comparative bisulfite sequences may prove useful to train computer algorithms for predicting methylation patterns. RRBS sequencing may be sufficient to detect genomic imprints (or the loss thereof), tissue-specific regulated methylation domains or long-range methylation gradients along a chromosome. We also envision RRBS applications in epigenetic cancer profiling and biomarker discovery.

Methylation patterns in methylation-impaired ES cells

Despite the essential role of each known DNA methyltransferase during mouse development (11,13), DNA methylation and the enzymes responsible for its establishment and maintenance appear to be largely dispensable in undifferentiated ES cells. *Dnmt*1 knockout ES cells retain ~20% CpG methylation probably a result of continuous *de novo* methylation by Dnmt3a and Dnmt3b. Although early passage Dnmt3a/b double mutant ES cells show almost wild-type levels of CpG methylation (43,44), they progressively lose methylation with <1% remaining after 75 passages (44). This gradual loss may reflect the infidelity of the maintenance enzyme Dnmt1.

Our data showed that ES cells lacking DNA methyltransferases Dnmt3a and 3b and with greatly reduced levels of Dnmt1 were viable with 1–2% CpG methylation remaining after only six passages. The extremely low rate of false-positive methylcytosines allowed us to identify and inspect some of the rare sites that retained methylation. There were no obvious hotspots of residual mCpGs in the genome (Supplementary Figures). Also, there was no correlation between the numbers of CpGs and the residual methylation at a given site. The distance to CpG islands or to known genes appeared to be random. None of the loci was notably conserved across species. Finally, no specific motif was detected upstream and downstream of the residual mCpG dinucleotides (data not shown). Our findings provide no evidence of specific maintenance of residual mCpG by yet another DNA methyltransferase. Rather, *Dnmt*[1^{kd},$3a^{-/-}$,$3b^{-/-}$] cells seem to lose residual CpG methylation in a random fashion over time.

Only 3 of the 25 505 sequenced CpT dinucleotides were inferred to be methylated in Dnmt-deficient cells. No methylated CpA was detected. By comparison, wild-type cells showed 0.7% CpT and 2.4% CpA methylation in agreement with previous observations (45,46). Previous experiments have also shown that the presence of Dnmt1 is not required for non-CpG methylation (46). In contrast, non-CpG methylation becomes almost undetectable in ES cells lacking Dnmt3a and Dnmt3b (45). Both global nearest neighbor and our bisulfite-sequencing data therefore suggest that the *de novo* DNA methyltransferases 3a and/or 3b are responsible for asymmetric CpA and CpT methylation in murine ES cells.

CONCLUSION

In this pilot study we have employed a combination of RNAi-induced knockdown and complete knockout of DNA methyltransferases to generate murine ES cells that were almost devoid of DNA methylation. These cells had 1–2% residual CpG methylation left after a few passages, and non-CpG methylation was >250-fold reduced compared with wild-type ES cells.

Unamplified, nearly methylation-free genomic DNA is an ideal substrate to optimize and test conditions for genome-wide bisulfite conversion, PCR amplification and library construction for future genomic shotgun bisulfite sequencing of mammalian genomes. We have shown that essentially complete bisulfite conversion can be achieved without undue adverse effects on library complexity and sequence representation.

Large-scale random bisulfite sequencing complements existing directed bisulfite sequencing strategies, which are well suited to analyze a limited number of gene promoters and regulatory sequence elements in a large number of samples. One advantage of sequencing clone libraries in a random fashion is that no target-specific PCR or sequencing primers are needed. Once the library is made, the method is amenable to automation and is scaleable. Since the bisulfite reads are not assembled but merely aligned to the reference genome sequence, we expect this method to work well in combination with highly parallel sequencing technologies that produce single reads of ~100 bases in length (47). Finally, in principle, bisulfite-converted libraries can be constructed from randomly sheared DNA for future whole-genome bisulfite sequencing.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Laurie Jackson Grusby, Konrad Hochedlinger and Katrin Plath for critical reading of the manuscript. This work was funded in part by NIH grants 5R01CA87869 (RJ) and HG03067-02 (ESL). AM was supported by the Boehringer Ingelheim Fonds. Funding to pay the Open Access publication charges for this article was provided by Whitehead Institute for Biomedical Research.

Conflict of interest statement. None declared.

REFERENCES

1. Jeltsch,A. (2002) Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. *Chembiochem*, **3**, 274–293.
2. Robertson,K.D. and Wolffe,A.P. (2000) DNA methylation in health and disease. *Nature Rev. Genet.*, **1**, 11–19.
3. Jaenisch,R. (1997) DNA methylation and imprinting: why bother? *Trends Genet*, **13**, 323–329.
4. Bestor,T.H. (2000) The DNA methyltransferases of mammals. *Hum. Mol. Genet.*, **9**, 2395–2402.
5. Jaenisch,R. and Bird,A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genet.*, **33** (Suppl), 245–254.
6. Jones,P.A. and Baylin,S.B. (2002) The fundamental role of epigenetic events in cancer. *Nature Rev. Genet.*, **3**, 415–428.
7. Robertson,K.D. (2002) DNA methylation and chromatin—unraveling the tangled web. *Oncogene.*, **21**, 5361–5379.
8. Laird,P.W. (2003) The power and the promise of DNA methylation markers. *Nature Rev. Cancer*, **3**, 253–266.
9. Gaudet,F., Hodgson,J.G., Eden,A., Jackson-Grusby,L., Dausman,J., Gray,J.W., Leonhardt,H. and Jaenisch,R. (2003) Induction of tumors in mice by genomic hypomethylation. *Science*, **300**, 489–492.
10. Feinberg,A.P. (2004) The epigenetics of cancer etiology. *Semin. Cancer Biol.*, **14**, 427–432.
11. Li,E., Bestor,T.H. and Jaenisch,R. (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, **69**, 915–926.
12. Okano,M., Xie,S. and Li,E. (1998) Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature Genet.*, **19**, 219–220.
13. Okano,M., Bell,D.W., Haber,D.A. and Li,E. (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, **99**, 247–257.
14. Okano,M. and Li,E. (2002) Genetic analyses of DNA methyltransferase genes in mouse model system. *J. Nutr.*, **132**, 2462S–2465S.
15. Lane,N., Dean,W., Erhardt,S., Hajkova,P., Surani,A., Walter,J. and Reik,W. (2003) Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *Genesis*, **35**, 88–93.
16. Lei,H., Oh,S.P., Okano,M., Juttermann,R., Goss,K.A., Jaenisch,R. and Li,E. (1996) *De novo* DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development*, **122**, 3195–3205.
17. Murrell,A., Rakyan,V.K. and Beck,S. (2005) From genome to epigenome. *Hum. Mol. Genet.*, **14**, R3–R10.
18. Ramsahoye,B.H. (2002) Measurement of genome wide DNA methylation by reversed-phase high-performance liquid chromatography. *Methods*, **27**, 156–161.
19. Ramsahoye,B.H. (2002) Nearest-neighbor analysis. *Methods Mol. Biol.*, **200**, 9–15.
20. Lippman,Z., Gendrel,A.V., Colot,V. and Martienssen,R. (2005) Profiling DNA methylation patterns using genomic tiling microarrays. *Nature Methods*, **2**, 219–224.
21. Lippman,Z., Gendrel,A.V., Black,M., Vaughn,M.W., Dedhia,N., McCombie,W.R., Lavine,K., Mittal,V., May,B., Kasschau,K.D. *et al.* (2004) Role of transposable elements in heterochromatin and epigenetic control. *Nature*, **430**, 471–476.
22. Yamada,Y., Watanabe,H., Miura,F., Soejima,H., Uchiyama,M., Iwasaka,T., Mukai,T., Sakaki,Y. and Ito,T. (2004) A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. *Genome Res.*, **14**, 247–266.
23. Bedell,J.A., Budiman,M.A., Nunberg,A., Citek,R.W., Robbins,D., Jones,J., Flick,E., Rholfing,T., Fries,J., Bradford,K. *et al.* (2005) Sorghum genome sequencing by methylation filtration. *PLoS Biol.*, **3**, e13.
24. Strichman-Almashanu,L.Z., Lee,R.S., Onyango,P.O., Perlman,E., Flam,F., Frieman,M.B. and Feinberg,A.P. (2002) A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. *Genome Res.*, **12**, 543–554.
25. Rabinowicz,P.D., Schutz,K., Dedhia,N., Yordan,C., Parnell,L.D., Stein,L., McCombie,W.R. and Martienssen,R.A. (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nature Genet.*, **23**, 305–308.
26. Weber,M., Davies,J.J., Wittig,D., Oakeley,E.J., Haase,M., Lam,W.L. and Schubeler,D. (2005) Chromosome-wide and promotor-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genet.*, **37**, 853–862.
27. Frommer,M., McDonald,L.E., Millar,D.S., Collis,C.M., Watt,F., Grigg,G.W., Molloy,P.L. and Paul,C.L. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl Acad. Sci. USA*, **89**, 1827–1831.
28. Novik,K.L., Nimmrich,I., Genc,B., Maier,S., Piepenbrock,C., Olek,A. and Beck,S. (2002) Epigenomics: genome-wide study of methylation phenomena. *Curr. Issues Mol. Biol.*, **4**, 111–128.
29. Rakyan,V.K., Hildmann,T., Novik,K.L., Lewin,J., Tost,J., Cox,A.V., Andrews,T.D., Howe,K.L., Otto,T., Olek,A. *et al.* (2004) DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol.*, **2**, e405.
30. Garnes,J., Ciancio,M. and Gnirke,A. (2002) Construction Of Large-Insert Bacterial Clone Libraries. In Genomic Mapping and Sequencing. Horizon Scientific Press, UK.
31. Paulin,R., Grigg,G.W., Davey,M.W. and Piper,A.A. (1998) Urea improves efficiency of bisulphite-mediated sequencing of 5′-methylcytosine in genomic DNA. *Nucleic Acids Res.*, **26**, 5009–5010.
32. Don,R.H., Cox,P.T., Wainwright,B.J., Baker,K. and Mattick,J.S. (1991) 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res.*, **19**, 4008.
33. Altshuler,D., Pollara,V.J., Cowles,C.R., Van Etten,W.J., Baldwin,J., Linton,L. and Lander,E.S. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.
34. Ventura,A., Meissner,A., Dillon,C.P., McManus,M., Sharp,P.A., Van Parijs,L., Jaenisch,R. and Jacks,T. (2004) Cre-lox-regulated conditional RNA interference from transgenes. *Proc. Natl Acad. Sci. USA*, **101**, 10380–10385.
35. Chapman,V., Forrester,L., Sanford,J., Hastie,N. and Rossant,J. (1984) Cell lineage-specific undermethylation of mouse repetitive DNA. *Nature*, **307**, 284–286.

36. Walsh,C.P., Chaillet,J.R. and Bestor,T.H. (1998) Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genet.*, **20**, 116–117.
37. Lucifero,D., Mertineit,C., Clarke,H.J., Bestor,T.H. and Trasler,J.M. (2002) Methylation dynamics of imprinted genes in mouse germ cells. *Genomics*, **79**, 530–538.
38. Grunau,C., Clark,S.J. and Rosenthal,A. (2001) Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res.*, **29**, E65–65.
39. Fogg,M.J., Pearl,L.H. and Connolly,B.A. (2002) Structural basis for uracil recognition by archaeal family B DNA polymerases. *Nature Struct. Biol.*, **9**, 922–927.
40. Eads,C.A. and Laird,P.W. (2002) Combined bisulfite restriction analysis (COBRA). *Methods Mol. Biol.*, **200**, 71–85.
41. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
42. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
43. Chen,T., Ueda,Y., Dodge,J.E., Wang,Z. and Li,E. (2003) Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Mol. Cell. Biol.*, **23**, 5594–5605.
44. Jackson,M., Krassowska,A., Gilbert,N., Chevassut,T., Forrester,L., Ansell,J. and Ramsahoye,B. (2004) Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells. *Mol. Cell. Biol.*, **24**, 8862–8871.
45. Dodge,J.E., Ramsahoye,B.H., Wo,Z.G., Okano,M. and Li,E. (2002) *De novo* methylation of MMLV provirus in embryonic stem cells: CpG versus non-CpG methylation. *Gene*, **289**, 41–48.
46. Ramsahoye,B.H., Biniszkiewicz,D., Lyko,F., Clark,V., Bird,A.P. and Jaenisch,R. (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl Acad. Sci. USA*, **97**, 5237–5242.
47. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057), 376–380.