Library-free methylation sequencing with bisulfite padlock probes

Dinh Diep[1,2,4], Nongluk Plongthongkum[1,4], Athurva Gore[1,4], Ho-Lim Fung[1], Robert Shoemaker[3] & Kun Zhang[1,2]

Targeted quantification of DNA methylation allows for interrogation of the most informative loci across many samples quickly and cost-effectively. Here we report improved bisulfite padlock probes (BSPPs) with a design algorithm to generate efficient padlock probes, a library-free protocol that dramatically reduces sample-preparation cost and time and is compatible with automation, and an efficient bioinformatics pipeline to accurately obtain both methylation levels and genotypes from sequencing of bisulfite-converted DNA.

We have previously developed bisulfite padlock probes for the specific and parallel digital quantification of DNA methylation[1]. Recently, we enhanced BSPPs for improved flexibility and multiplexing capability. These improvements have contributed to recent findings in mouse and human pluripotent stem cells[2–5].

First, target selection and probe design is crucial for BSPPs. To aid in the design of efficient padlock probes for bisulfite analysis, we developed a program called ppDesigner. It accepts as input the genome of any organism, a user's list of arbitrary targets and user-desired probe constraints matching requirements of the experimental protocol. It *in silico* 'bisulfite-converts' the genome (that is, it changes all cytosines to thymines) and outputs padlock probes to cover the chosen targets while avoiding CpGs on the capturing arms that could be methylated and not converted to be recognized as thymine. ppDesigner uses a back-propagation neural network to predict probe efficiency (**Supplementary Fig. 1**). We had previously trained this network using data from probes for exomic targets[6] based on seven properties. Using bisulfite capture data from the first BSPPs[1], we refined the network with two additional factors. ppDesigner can explain ~50% of the variance in capturing efficiency for genomic DNA and ~20% of the variance in capturing efficiency for bisulfite-converted DNA; additional variation could be due to factors such as variability in oligonucleotide synthesis and sample DNA quality. ppDesigner is extremely flexible and has been used to design a variety of genomic and bisulfite probes for *Homo sapiens*[2,3], *Mus musculus*[4] and *Drosophila melanogaster*[7].

Key requirements for methylation analysis of large sample sizes include low cost, simple workflow and automation compatibility. As the cost of DNA sequencing has rapidly decreased, sample processing has become a bottleneck in terms of cost and throughput. A complicated workflow increases variability between samples and reduces power in large-scale studies. To address these issues, we extended a 'library-free' protocol[8] to multiplexed BSPP capture (**Fig. 1**). This method eliminates five steps from Illumina's library-construction protocol such that multiplexed libraries can be generated from DNA in only four steps (**Supplementary Table 1**). Using multiplexed primers with 6–base pair (bp) barcodes, we have routinely generated libraries for 96 samples in 96-well plates and sequenced all at once in a single Illumina HiSeq flowcell. Additionally we designed barcodes to process 384 samples per batch. As sample-specific barcodes were added, barcoded libraries can be pooled for size selection, which is the most time consuming, contamination-prone and error-prone step if performed individually. The protocol is compatible with the use of multichannel pipettes or liquid-handling devices. It dramatically reduced experimental cost and time, and improved reproducibility and read mapping rates (**Supplementary Tables 1** and **2**). For large sample sizes, the library preparation cost (including probes) with our protocol was comparable to that of the restricted-representation bisulfite sequencing and whole-genome bisulfite sequencing protocols, and the sequencing cost was much lower than that of whole-genome bisulfite sequencing owing to targeting of CpG sites of interest. Restricted-representation bisulfite sequencing is more cost-effective than BSPPs, but the former lacks BSPPs' flexibility in selecting specific sites or regions.

Another bottleneck in sequencing of bisulfite-converted DNA is a lack of computational tools to efficiently analyze sequencing data generated from hundreds of samples. To overcome this issue, we developed an analysis pipeline for read mapping and methylation quantification, called bisReadMapper (**Supplementary Fig. 2**). In previous padlock probe studies, reads had been mapped only against target regions owing to the computational requirements of sequence alignment[1]. In contrast, we designed bisReadMapper to map to the full genome sequence, allowing processing of data from both targeted and whole-genome sequencing of bisulfite-converted DNA. bisReadMapper also determines the origin strand of the read based on base composition and maps reads as if they were fully bisulfite-converted to a fully bisulfite-converted genome sequence, allowing mapping of both bi- and unidirectional bisulfite libraries in an unbiased manner. Another feature is the capability to call single-nucleotide polymorphisms (SNPs) from sequences of bisulfite-converted DNA; this feature not only allows for analysis of allele-specific methylation[9]

[1]Department of Bioengineering, University of California at San Diego, La Jolla, California, USA. [2]Bioinformatics and System Biology Graduate Program, University of California at San Diego, La Jolla, California, USA. [3]Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, California, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to K.Z. (kzhang@bioeng.ucsd.edu).

RECEIVED 1 SEPTEMBER 2011; ACCEPTED 21 DECEMBER 2011; PUBLISHED ONLINE 5 FEBRUARY 2012; DOI:10.1038/NMETH.1871

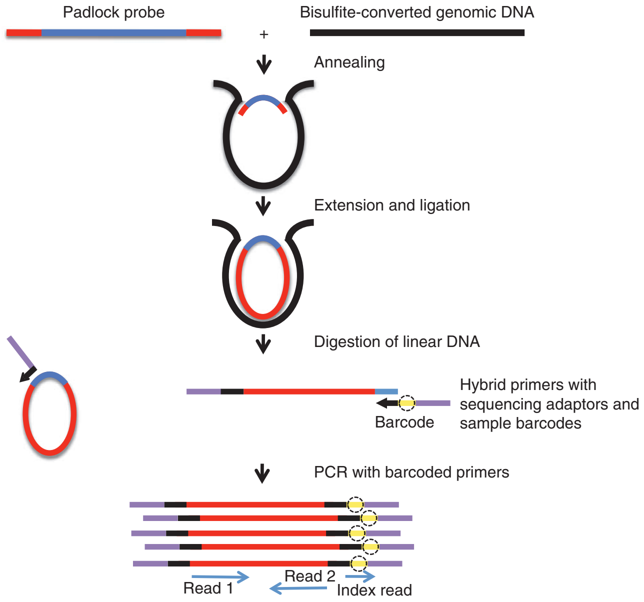
Padlock probe
+
Bisulfite-converted genomic DNA
Annealing
Extension and ligation
Digestion of linear DNA
Hybrid primers with sequencing adaptors and sample barcodes
Barcode
PCR with barcoded primers
Read 1
Read 2
Index read

Figure 1 | Library-free BSPP protocol. Each padlock probe has a common linker sequence flanked by two target-specific capturing arms (red) that anneal to bisulfite converted genomic DNA. The 3′ end is extended and ligated with the 5′ end to form circularized DNA. After removal of linear DNA, all circularized captured targets are PCR-amplified with barcoded primers and directly sequenced with an Illumina sequencing platform (GA II(x) or HiSeq). Amplicon size is 363 bp, which includes captured target (180 bp), capturing arms (55 bp), and amplification primers and adapters (128 bp). The inserts can be read through with paired-end 120-bp sequencing reads.

but also allows accurate sample tracking in large-scale experiments. Finally, bisReadMapper can call methylation levels at both CpG and non-CpG sites.

To test our assay, we generated a genome-scale probe set based on our previous results and new information about differential methylation[1,10–12]. We targeted our new design for evaluation of methylation at genomic locations known to contain differentially methylated regions or differentially methylated sites (DMSs)[10–13], transcriptional repressor CTCF binding sites and DNase I–hypersensitive regions. We also targeted all microRNA genes and all promoters for human US National Center for Biotechnology Information reference sequence (RefSeq) genes. Using ppDesigner, we designed ~330,000 padlock probes that covered 140,749 non-overlapping regions with a total size of 34 megabases. We performed capturing experiments and end-sequencing, and found that these probes were slightly more specific (~96% on-target) and uniform than previous probes[1] (**Supplementary Fig. 3**). To improve uniformity, we normalized the experimental capturing performance of these probes using subsetting and suppressor oligonucleotides as described previously[1]. We could characterize roughly 500,000 CpG sites with ~4 gigabases of sequencing reads, and additional sites became callable with deeper sequencing (**Supplementary Fig. 4** and **5**).

We used these probes to analyze H1 embryonic stem cells (H1 ESCs), PGP1 fibroblasts and two technical replicates of PGP1 fibroblast–derived induced pluripotent stem cells (PGP1-iPSCs). For each sample, we sequenced on average ~3.66 gigabases and measured methylation for an average of 480,904 CpG sites. To assess whether these data could be used to identify potential epigenetic regulation of transcription, we used the genomic regions enrichment of annotations tool[14] to predict the *cis*-regulatory potential of regions around captured CpG sites. In total, the padlock probes captured CpG sites in regions predicted to regulate 98% of RefSeq genes (**Supplementary Fig. 6**).

The data generated with BSPPs accurately represented the methylation status of the target regions. Methylation levels for the two technical replicates of PGP1-iPSCs were consistent both within a single batch and between separate batches (Pearson's correlation coefficient $R = 0.97–0.98$, **Supplementary Fig. 7a,b**). Additionally, when we compared methylation levels between technical replicates, no CpG site was different by a Fisher Exact Test with Benjamini-Hochberg multiple testing correction (false discovery rate = 0.01, $n = 439,090$). In comparison, large fractions of sites were differentially methylated owing to either the process of nuclear reprogramming (27.9% DMSs between PGP1-iPSCs and PGP1 fibroblasts) or the difference in cell type (31.3% DMSs between PGP1 fibroblasts and H1 ESCs) with the same criteria (false discovery rate = 0.01, $n = 444,111$ and 359,290, respectively). Our BSPP results with H1 ESCs were consistent with the published whole-genome sequencing of bisulfite-converted DNA[12] (Pearson's correlation coefficient $R = 0.95$, **Supplementary Fig. 8**).

Our assay has very low technical variability. We performed the assay on over 150 samples in 96-well plates; the yield for each was similar (**Supplementary Fig. 9**). Approximately 10% of CpG sites were targeted separately on each strand, allowing low-quality datasets with poor correlation between these built-in technical replicates to be identified (**Supplementary Fig. 7c–e**). As our BSPP assay measures absolute methylation, no normalization is necessary as long as the internal replicates are consistent. Therefore, many datasets, even those generated in different laboratories, can be directly compared without batch effects, which is important for case-control studies on large samples or for meta-analyses. Additionally, the SNP-calling feature of bisReadMapper allowed us to characterize roughly 20,000 SNPs for each sample with an accuracy of 96% or better. This allowed us to unambiguously track samples, which is crucial for projects involving large sample sizes.

Our library-free BSPP method is flexible for different study designs. Whereas our genome-scale probe set allows global profiling on thousands of samples, a focused assay is often necessary to follow up on tens to hundreds of candidate regions identified in genome-scale scanning. Such an assay needs to be customizable to different genomic targets, scalable to a very large sample size (1,000–100,000 samples), and inexpensive. To additionally test the flexibility, we designed a second set of 3,918 probes to evaluate the methylation state 1 kbp upstream and downstream of 120 genomic regions previously known and confirmed by BSPP to carry aberrant methylation in induced pluripotent stem cells[15]. We acquired the oligonucleotides from a second vendor (LC Sciences). Even with shorter capturing sequences (40 bp total for capturing arms rather than 50 bp on average, **Supplementary Fig. 10**) and a 100-fold smaller target size, an average of 56% of mappable bases were on-target, equivalent to an enrichment factor of ~6,500. With the data from three cell lines (H1 ESCs, PGP1 fibroblasts and PGP1-iPSCs) we identified regions of aberrant methylation in induced pluripotent stem cells (**Supplementary Fig. 11**) and demonstrated that aberrant methylation continues further upstream and downstream than observed previously. This analysis demonstrated that a focused probe set can be used to validate specific regions of interest identified in global scanning using either our genome-wide probe set or other methods.

This method can be implemented to aid in identifying the effects of DNA methylation in any organism by using the computational tools at http://genome-tech.ucsd.edu/public/Gen2_BSPP/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

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

ACKNOWLEDGMENTS

We thank A. Feinberg and J. Stamatoyannopoulos for providing informative genomic targets, and E. LeProust (Agilent Technology) for long-oligonucleotide synthesis. This work was funded by grants from US National Institutes of Health (R01 DA025779 and R01 GM097253) to K.Z.; A.G. is sponsored by a California Institute of Regenerative Medicine predoctoral fellowship. N.P. is supported by the Royal Thai Government Scholarship.

AUTHOR CONTRIBUTIONS

K.Z. oversaw the project. D.D. and N.P. performed experiments and bioinformatic analyses. A.G. developed a probe design algorithm. R.S. and A.G. designed probes. H.-L.F. performed sequencing. D.D., N.P., A.G. and K.Z. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

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

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Deng, J. *et al. Nat. Biotechnol.* **27**, 353–360 (2009).
2. Liu, G.H. *et al. Nature* **472**, 221–225 (2011).
3. Liu, G.H. *et al. Cell Stem Cell* **8**, 688–694 (2011).
4. Xu, Y. *et al. Mol. Cell* **42**, 451–464 (2011).
5. Hansen, K.D. *et al. Nat. Genet.* **43**, 768–775 (2011).
6. Gore, A. *et al. Nature* **471**, 63–67 (2011).
7. Wang, H. *et al. Genome Res.* **20**, 981–988 (2010).
8. Turner, E.H., Lee, C., Ng, S.B., Nickerson, D.A. & Shendure, J. *Nat. Methods* **6**, 315–316 (2009).
9. Shoemaker, R., Deng, J., Wang, W. & Zhang, K. *Genome Res.* **20**, 883–889 (2010).
10. Irizarry, R.A. *et al. Nat. Genet.* **41**, 178–186 (2009).
11. Doi, A. *et al. Nat. Genet.* **41**, 1350–1353 (2009).
12. Lister, R. *et al. Nature* **462**, 315–322 (2009).
13. Figueroa, M.E. *et al. Cancer Cell* **17**, 13–27 (2010).
14. McLean, C.Y. *et al. Nat. Biotechnol.* **28**, 495–501 (2010).
15. Lister, R. *et al. Nature* **471**, 68–73 (2011).

ONLINE METHODS

Probe design. Probe design and read mapping algorithms, probe sequences and additional information are available at http://genome-tech.ucsd.edu/public/Gen2_BSPP/. A schematic for the padlock probes is illustrated in **Supplementary Figure 10**.

Bisulfite padlock probe production (oligonucleotides from Agilent). Libraries of oligonucleotides (~150 nucleotides (nt)) were synthesized by ink-jet printing on programmable microarrays (Agilent Technologies) and released to form a combined library of 330,000 oligonucleotides. The oligonucleotides were amplified by PCR in 96 reactions (100 µl each) with 0.02 nM template oligonucleotide, 400 nM each of pAP1V61U primer and AP2V6 primer (**Supplementary Table 3**), and 50 µl of KAPA SYBG fast Universal 2× qPCR Master Mix (Kapabiosystems) at 95 °C for 30 s, 15–16 cycles of 95 °C for 3 s; 55 °C for 30 s; and 60 °C for 20 s and 60 °C for 2 min. The amplified amplicons were purified by ethanol precipitation and repurified with Qiaquick PCR purification columns (Qiagen). Approximately 20 µg of the purified amplicons were digested with 50 units of lambda exonuclease (5 U μl^{-1}; New England Biolabs (NEB)) at 37 °C for 1 h in lambda exonuclease reaction buffer. The resulting single-strand amplicons were purified with Qiaquick PCR purification column. Approximately 5–8 µg of single strand amplicons were subsequently digested with 5 units USER (1 U μl^{-1}, NEB) at 37 °C for 1 h. The digested DNAs were annealed to 5.88 µM RE-DpnII-V6 guide oligo (**Supplementary Table 3**) and denatured at 94 °C for 2 min decreased the temperature to 37 °C and incubated at 37 °C for 3 min. The mixture was digested with 50 U of DpnII (10 U μl^{-1}, NEB) in NEBuffer DpnII at 37 °C for 2 h. Then the mixture was further digested with 5 U of USER at 37 °C for 2 h followed by enzyme inactivation at 75 °C for 20 min. The USER and DpnII–digested DNAs were purified with Qiaquick PCR purification column. The single-strand 102-nt probes were purified with 6% denaturing PAGE (6% TB-urea two dimensional (2D) gel; Invitrogen).

Bisulfite padlock probe production (oligonucleotides from LC Sciences). The oligonucleotides (100 nt) were synthesized using a programmable microfluidic microarray platform (LC Sciences) and released to form a mix of 3,918 oligonucleotides. The oligonucleotides were amplified by two-step PCR in a 200 µl reaction with 1 nM template oligonucleotides, 400 nM each of eMIP_CA1_F primer and eMIP_CA1_R primer (**Supplementary Table 3**), and 100 µl of KAPA SYBR fast Universal qPCR Master Mix at 95 °C for 30 s, 5 cycles of 95 °C for 5 s; 52 °C for 1 min; and 72 °C for 30 s, 10–12 cycles of 95 °C for 5 s; 60 °C for 30 s; and 72 °C for 30 s, and 72 °C for 2 min. The resultant amplicons were purified with Qiaquick PCR purification columns and re-amplified in 32 PCRs (100 µl each) with 0.02 nM first round amplicons, 400 nM each of eMIP_CA1_F primer and eMIP_CA1_R primer and 50 µl of KAPA SYBR fast Universal qPCR master mix at 95 °C for 30 s, 13–15 cycles of 95 °C for 5 s; 60 °C for 30 s; and 72 °C for 30 s and 72 °C for 2 min. The resultant amplicons were purified by ethanol precipitation and repurified with Qiaquick PCR purification columns as described above. Approximately 4 µg of the purified amplicons were digested with 100 U of Nt.AlwI (100 U μl^{-1}, NEB) at 37 °C for 1 h in NEBuffer 2. The enzyme was heat-inactivated at 80 °C for 20 min. The digested amplicons were then incubated with 100 U of Nb.BrsDI (10 U μl^{-1}, NEB) at 65 °C for 1 h. The nicked DNA was purified by Qiaquick PCR purification column. The probe molecules (~70 bases) were purified by 6% denaturing PAGE (6% TB-urea 2D gel).

Sample preparation and capture. Genomic DNA was extracted using the AllPrep DNA/RNA Mini kit (Qiagen) and bisulfite-converted with the EZ-96 DNA methylation Gold kit (Zymoresearch) in a 96-well plate. Normalized amount of padlock probes, 200 ng of bisulfite converted gDNA and 4.2 nM oligo suppressor were mixed in 25 µl 1× Ampligase buffer (Epicentre) in 96-well plate, denatured at 95 °C for 10 min, gradually lowered the temperature at 0.02 °C s^{-1} to 55 °C in a thermocycler and hybridized at 55 °C for 20 h. 2.5 µl of SLN (Stoffel fragment, ligase, nucleotides) mix (100 µM dNTP, 2 U μl^{-1}AmpliTaq Stoffel Fragment (ABI) and 0.5 U μl^{-1} Ampligase (Epicentre) in 1× Ampligase buffer) was added to the reaction for gap-filling. For circularization, the reactions were incubated at 55 °C for 20 h, followed by enzyme inactivation at 94 °C for 2 min. To digest linear DNA after circularization, 2 µl of exonuclease mix (10 U μl^{-1} exonuclease I and 100 U μl^{-1}exonuclease III, USB) was added to the reactions, and the reactions were incubated at 37 °C for 2 h and then inactivated at 94 °C for 2 min.

Capture-circles amplification (library-free BSPP protocol, Agilent oligonucleotides). Ten microliters of circularized DNA was amplified and barcoded in 100-µl reactions with 400 nM each of AmpF6.3Sol primer (**Supplementary Table 3**) and AmpR6.3 indexing primer (**Supplementary Table 3**), 0.4× SYBR Green I (Invitrogen) and 50 µl Phusion High-Fidelity 2× master mix (NEB) at 98 °C for 30 s, 5 cycles of 98 °C for 10 s; 58 °C for 20 s; and 72 °C for 20 s, 9-12 cycles of 98 °C for 10 s; and 72 °C for 20 s and 72 °C for 3 min.

Capture-circles amplification (library-free BSPP protocol, LC Sciences oligonucleotides). Ten microliters of circularized DNA was amplified in a 100-µl reaction with 200 nM each of CP-2-FA primer and CP-2-RA primer (**Supplementary Table 3**) and 50 µl KAPA SYBR fast Universal qPCR Master Mix at 98 °C for 30 s, 5 cycles of 98 °C for 10 s; 52 °C for 30 s; and 72 °C for 30 s, 15 cycles of 98 °C for 10 s; 60 °C for 30 s; and 72 °C for 30 s and 72 °C for 3 min. The resultant amplicons with the corresponding expected size of ~260 bp were purified by 6% PAGE (6% 5-well gel, Invitrogen) and resuspended in 12 µl of TE buffer. Thirty percent of the gel-purified amplicons were reamplified and barcoded in a 100-µl reaction with 200 nM each of two different sets of primers to enable single-end sequencing of both ends of the amplicons (CP-2-FA.IndSol primer and CP-2-RA.Sol primer or Switch.CP-2-FA and Switch.CP-2-RA.IndSol) and 50 µl KAPA SYBR fast Universal qPCR Master Mix at 98 °C for 30 s, 4 cycles of 98 °C for 10 s; 54 °C for 30 s; and 72 °C for 30 s and 72 °C for 3 min.

Primer barcode design for multiplexing. An in-house Perl script was written to randomly generate 6-nt-long sequences. A sequence was kept if it did not have more than two matching positions with another accepted barcode and if it had 2–4 guanines or cytosines. The script reiterated until the desired number of barcodes have been obtained. A total of 384 primers were designed (**Supplementary Table 4).**

doi:10.1038/nmeth.1871

Bisulfite read mapping and data analysis. Bisulfite-converted data were processed as previously described. Reference genome was computationally converted by changing all cytosines to thymines on the two strands separately. Sequencing reads were encoded by (i) predicting the mapping orientation, (ii) converting all predicted forward mapping reads by changing all cytosines to thymines and converting all predicted reverse complementary mapping reads by changing all guanines to adenines, the original reads are maintained. The bisulfite reads were then mapped to the converted reference separately using SOAP2Align (http://soap.genomics.org.cn/soapaligner.html) with the parameters $r = 0$ (report uniquely mapped reads only), $v = 2$ (number of allowable mismatch), paired-end: $m = 0$ (minimal insert size), $x = 400$ (maximum insert size). Alignment files were then combined, and one alignment per read was selected. Original C calls were placed back into the alignment information. Alignments were then converted to pileup format using SamTools (http://samtools.sourceforge.net/). Raw SNPs and methylation frequency files were computed from pileup counts. Methylation frequencies were called using a method described previously[1].

Correlation of methylation between two samples. To check whether methylation levels were similar between two samples, the Pearson's correlation was calculated on all CpG sites characterizable in both samples. First, a list of CpG sites with read depth of at least ten in both samples was generated. The methylation frequencies at these sites were obtained from bisReadMapper output and input into the statistical package R. Finally, Pearson's correlation for the two samples was computed using the cor() function.

Analysis of methylation. From the bisReadMapper output, the raw read counts showing methylation and lack of methylation were assembled for each line. Using these counts, a Fisher Exact Test with Benjamini-Hochberg multiple testing correction (false discovery rate = 0.01) was carried out on each CpG site with minimum 10× depth coverage. This resulted in a set of DMSs between the two lines; at each of these sites, the methylation had at least 0.1 methylation level difference. Technical replicates did not show any differential methylation, and different cell types showed a large extent of differential methylation (~33%).

doi:10.1038/nmeth.1871