High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers

Mayumi Oda1, Jacob L. Glass2, Reid F. Thompson2, Yongkai Mo3, Emmanuel N. Olivier1,4, Maria E. Figueroa5, Rebecca R. Selzer6, Todd A. Richmond6, Xinmin Zhang6, Luke Dannenberg6, Roland D. Green6, Ari Melnick5, Eli Hatchwell7, Eric E. Bouhassira4, Amit Verma3, Masako Suzuki2 and John M. Greally1,2,*

1Department of Medicine (Hematology), 2Department of Genetics (Computational Genetics), 3Department of Developmental and Molecular Biology, 4Department of Cell Biology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, 5Division of Hematology and Medical Oncology, Weill Cornell Medical Center, 1300 York Avenue, New York, NY, 6Roche-NimbleGen Inc., 504 S. Rosa Road, Madison, WI and 7Department of Pathology, State University of New York at Stony Brook, Stony Brook, NY, USA

Received December 10, 2008; Revised April 6, 2009; Accepted April 7, 2009

ABSTRACT

Many genome-wide assays involve the generation of a subset (or representation) of the genome following restriction enzyme digestion. The use of enzymes sensitive to cytosine methylation allows high-throughput analysis of this epigenetic regulatory process. We show that the use of a dualadapter approach allows us to generate genomic representations that includes fragments of <200 bp in size, previously not possible when using the standard approach of using a single adapter. By expanding the representation to smaller fragments using HpaII or MspI, we increase the representation by these isoschizomers to more than 1.32 million loci in the human genome, representing 98.5% of CpG islands and 91.1% of refSeq promoters. This advance allows the development of a new, high-resolution version of our HpaII-tiny fragment Enrichment by Ligation-mediated PCR (HELP) assay to study cytosine methylation. We also show that the MspI representation generates information about copy-number variation, that the assay can be used on as little as 10 ng of DNA and that massively parallel sequencing can be used as an alternative to microarrays to read the output of the assay, making this a powerful discovery platform for studies of genomic and epigenomic abnormalities.

INTRODUCTION

At present, there are several types of assays to test cytosine methylation on a genome-wide scale. Some depend on the relative enrichment of methylated or hypomethylated DNA, mapping back these fractions to the genome by microarray or, more recently, massively parallel sequencing techniques. Enrichment of a certain pattern of methylation can be performed using proteins binding to methylated DNA [antibodies (1,2) or methyl-binding domain proteins (3)], methylation-dependent restriction enzymes [mcrBC (4)], methylation-dependent restriction enzymes [mcrBC (4)] or methylation-sensitive restriction enzymes. The development of massively parallel sequencing promises to allow nucleotide-resolution, quantitative analysis of cytosine methylation throughout mammalian genomes, as demonstrated in Arabidopsis (5,6) although as we have discussed there remain significant practical problems to performing such assays at present (7). These problems include significant cost, due to the need to deeply sequence an entire genome but further complicated by the creation of noncomplementary forward and reverse strands, doubling the effective size of the reference genome, and the difficulty of mapping short, degenerate bisulphite-converted sequences to this reference genome, especially at unmethylated loci where the cytosines are all

*To whom correspondence should be addressed. Tel: 718 678 1234; Fax: 718 678 1020; Email: jgreally@aecom.yu.edu

Present address:
Mayumi Oda, Laboratory of Cellular Biochemistry, Department of Veterinary Medical Sciences/Animal Resource Sciences, University of Tokyo, Bunkyo-ku, Tokyo, Japan

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

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
were differentiated by co-culture on immortalized hESCs (NIH code WA01 from Wicell Research Institute, Madison WI) were cultured on matrigel (BD Biosciences, San Diego). Amplified hESC pluripotency markers was assessed by flow cytometry with SSEA4, CD24 and CD71 and glycophorin A (glycophorin A) markers. Primary acute lymphoblastic leukemia cells were obtained from Our Lady of Mercy Cancer Center, approved by Institutional Review Board of the Albert Einstein College of Medicine. One million fresh CD34 + bone marrow progenitors were purchased from Alcells (Emerville, CA). DNA was extracted from these cells and checked for purity and amount by spectrometry (Nanodrop, Wilmington, DE).

**METHODS**

**Cell preparation and DNA purification**

GM06990 B lymphoblastoid cells were cultured in RPMI 1640 with 15% FBS, 1% glutamine and antibiotics. The cells were harvested and washed twice by PBS and stored in −70°C. The H1 human embryonic stem cells (hESCs, NIH code WA01 from Wicell Research Institute, Madison WI) were cultured on matrigel (BD Biosciences, San Diego). Amplified hESC pluripotency was assessed by flow cytometry with SSEA4, CD24 and Oct4 markers. Erythroid progenitors derived from hESCs were produced as previously described (14). The hESCs were differentiated by co-culture on immortalized human fetal hepatocytes (FH-B-hTERT) for 2 weeks. CD34-positive cells were then magnetically sorted and placed in a serum-free liquid culture for two weeks. The erythroid characteristics of the differentiated cells were assessed by flow cytometry with CD71 and CD235a (glycophorin A) markers. Primary acute lymphoblastic leukemia cells were obtained from Our Lady of Mercy Cancer Center, approved by Institutional Review Board of the Albert Einstein College of Medicine. One million fresh CD34 + bone marrow progenitors were purchased from Alcells (Emerville, CA). DNA was extracted from these cells and checked for purity and amount by spectrometry (Nanodrop, Wilmington, DE).

**LM-PCR with a two adapter set**

The primers used were JHpaII12 (5′-CGGCTTGTTCATG-3′), JHpaII24 (5′-CGACGTGAATCCATCAAGAC-3′), NHpaII12 (5′-CGGCTTCCCTCG-3′) and NHpaII24 (5′-GCAAATGTGATCCGAAGGGAAGC-3′). Our previous protocol (8) was modified with the additional adapter set (NHpaII12/NHpaII24). Five micrograms of genomic DNA were digested by either HpaII or MspI and purified by phenol/chloroform extraction and ethanol precipitation. Each 1 µg of digested genomic DNA was ligated by T4 DNA ligase using four oligos (JHpaII12 and NHpaII12, JHpaII24 and NHpaII24, each 40 mM) in a final volume of 33 µl. The ligated genomic fragments were diluted and used as the template for the LM-PCR. The optimization of PCR conditions for LM-PCR was performed using our prior PCR conditions (8) with/without betaine or dimethyl sulfoxide. PCR products were assessed by gel electrophoresis and purified using a PCR purification kit (Qiagen). The concentrations of PCR products were measured by spectrometry. The intensities of DNA from gel images were processed using ImageJ and Photoshop (Adobe).

To improve the PCR conditions, we tested higher concentrations of magnesium, and explored the use of betaine as a means of improving the amplification of (G + C)-rich templates (15). Dimethyl sulfoxide was also tested with and without betaine, but this failed to enhance short fragment amplification (data not shown).

**Array design, hybridization and data analysis**

We custom-designed two microarrays (Roche-NimbleGen, Inc.) for these experiments. One interrogated the 1% of the genome studied by the ENCODE consortium in their pilot phase using oligonucleotides for each of the 18 529 HpaII/MspI fragments of 50–2000 bp in these ENCODE regions (16). The second tested the >1.32 million loci throughout the human genome that have HpaII sites 50–2000 bp apart and have unique sequences between them that allow oligonucleotide design. The HELP samples were labeled for microarray analysis as described previously (17) using Cy3- or Cy5-conjugated random primers. The HpaII and MspI representations were co-hybridized to the microarray and scanned to quantify the 532- and 635-nm fluorescence at each oligonucleotide on the microarray. The quantile-normalized HpaII/MspI log₂ ratios were calculated as
Correlation of genomic annotations with HELP cytosine methylation data

The start and end positions of MspI fragments were computationally calculated from the hg17 assembly of the human genome sequence at the UCSC Genome Browser (genome.ucsc.edu). Using these DNA sequences, their base compositional characteristics [(C + G) mononucleotide percent, CG dinucleotide frequency per 1 kb] were computationally calculated. The genomic positions of the MspI fragments were used to quantify the overlap with annotations such as CpG islands (epigislandExt table), CG clusters (18), retroelements (rmsk table) and the 2-kb region flanking refSeq transcription start sites.

Use of limited amounts of starting material

Ten nanograms of genomic DNA (representing ~2000 mammalian diploid cell equivalents) from human melanoma samples were digested overnight by HpaII or MspI (New England Biolabs, Inc., Ipswich, MA) in separate 20-μl reactions. The digested DNA fragments were purified by phenol/chloroform/isoamyl alcohol (25:24:1 by volume, pH 5.2, Thermo Fisher Scientific Inc., Waltham, MA) and ethanol precipitated in the presence of glycogen. After washing with 70% ethanol, the DNA pellet was resuspended in 5 μl of Tris–HCl (pH 8.0).

Overnight adapter ligation was set up in a 10-μl reaction, scaling the proportional amount of adapter down by a factor of 30 compared with 1 μg HELP, with T4 DNA ligase (Invitrogen Corp., Carlsbad, California). Adapter-ligated DNA fragments were amplified by two rounds of PCR [67 mM Tris–HCl, 4 mM MgCl2, 16 mM (NH4)2SO4, 0.35% β-mercaptoethanol, 0.5 μg/μl BSA, Taq DNA polymerase] using 72°C for 10 min, 15 cycles of 95°C for 30 s and 72°C for 3 min, with a final extension of 72°C for 10 min. For the first round of PCR, 2.5 μl (MspI) or 5 μl (HpaII) of the adapter ligation mixture was used in a 100-μl reaction, of which 1 μl of PCR product was used for a second 100-μl reaction.

Analysis of DNA copy-number variation using the MspI representation

The analysis of HELP data used the published analytical pipeline mentioned earlier (13). The MspI and HpaII signal intensities were summarized for each HpaII fragment as the average of its component probe-level signal intensities. Background noise thresholds were calculated using random probes as 2.5 median absolute deviations (MAD) above the median of MspI and HpaII signal intensities. Those fragments with HpaII above background but MspI below background were removed from consideration. A subset of fragments (those between 250 and 1000 bp) was then selected and centered to align the distribution ranges, with further removal of datapoints that exceeded 2 MADs from the center. Log ratios of MspI signal intensities were calculated for the test sample as a function of a normal cell type. The data were then divided by chromosome and analyzed for copy-number variation using DNAcopy (19), available through BioConductor for the R Statistical Package. Copy-number array (CNA) objects were smoothed and circular binary segmentation applied, with default parameters and change points <1 standard deviation removed.

Massively parallel sequencing of HELP representations

To eliminate adapters from LM-PCR products, we digested the purified LM-PCR product with MspI and concentrated the digested product using isopropanol precipitation. The digested product was loaded onto a 5% acrylamide gel, which was run to resolve DNA fragments in the <1000-bp size range. The DNA in the gel was visualized using visible spectrum transillumination following staining with SYBR Gold (Invitrogen), allowing us to excise the fragments of ~100–600 bp in size, in doing so eliminating the LM-PCR adapters. The DNA was extracted from the gel pieces using Miniprep columns (Qiagen), following which the eluted DNA was precipitated by isopropanol and purified with the PCR purification kit (Qiagen). Illumina libraries were prepared using their standard protocol, and single end sequencing was performed with a single lane per library, 47 bp per read. The first 32 bp were aligned to the reference genome using ELAND (Illumina), and these alignments were analysed using a local copy of the UCSC genome browser database and a downloaded copy of the human genome sequence (hg18 release). All HpaII sites in the genome were identified using a custom set of PERL and BASH shell scripts and loaded into a MySQL database. The HpaII site locations were then compared to the genomic loci corresponding to the Illumina reads using custom PERL scripts and modules based on the DBI PERL MySQL interface. The number of reads per locus (excluding those not starting with CGG) was also calculated using MySQL queries after the locations of the Illumina reads were loaded into the MySQL database. The HpaII representation was normalised in terms of that produced by MspI using the sequencing equivalent of the microarray intensity-dependent ratio (20) of the components as follows: $H \times 2^{(H-M)/(H+M)} + M \times 2^{(H-M)/(H+M)} - 1$, where $H$ = number of HpaII reads and $M$ = number of MspI reads. The results of these HELP and HELP-seq assays were compared against bisulphite MassArray (Sequenom) data from the KCNQ1 promoter region using primers and conditions described in Supplementary Table 1.

RESULTS

Our original assay failed to represent HpaIIMspI fragments <200 bp despite their relative abundance in the genome, as shown by our in silico digestion of the human genome at HpaII/MspI digestion sites (Figure 1). Of all the HpaII/MspI fragments in human DNA,
44.5% of fragments in the 200–2000-bp range are represented in the original HELP assay, while the 22.5% in the 50–200-bp range were not generated using our original protocol (Table 1). We considered two possible reasons for this. First, the regions with HpaII/MspI CCGG sites at higher frequency are enriched in (C + G) mononucleotide content, making them relatively difficult to amplify by conventional PCR techniques. The (C + G) content of the HpaII/MspI fragments of 50–2000-bp averages 60.0%, in contrast to the 47.7% for 200–2000-bp fragments. Second, the presence of a ligated adapter sequence may cause intramolecular self-annealing for shorter fragments preferentially, preventing PCR amplification of these hairpin (panhandle) structures (21). To test the second hypothesis, we added a second adaptor for the ligation step. This is expected to provide the ligated product with heterologous ends at 50% frequency (Figure 2a), preventing intramolecular annealing in this subpopulation of molecules. We find that the addition of the second adapter shifts the size range of PCR products compared with the use of a single adapter (Figure 2b). We conclude that the use of a single adapter for genomic representations fails to amplify fragments <200 bp because of a strong panhandle effect, a problem alleviated by the use of heterologous adapters.

These improvements in size range representation came at the expense of yield, which may be due to the increased (C + G) mononucleotide content of the template. By increasing the concentration of magnesium and introducing betaine to the PCR reaction (Figure 2c), we could generate amounts of PCR product comparable to the original protocol. The amplification products shown in Figure 2d include the adapter sequences (~50 bp), indicating that we are also amplifying products ≤50 bp in size while preserving our ability to amplify up to 2000 bp.

Our next concern was whether the shorter fragments generated could be labeled and hybridized to a microarray as successfully as larger fragments. We custom-designed a NimbleGen DNA methylation 385K microarray (Roche-NimbleGen Inc.) using oligonucleotides for each of the 18 529 HpaII/MspI fragments of 50–2000 bp in the ENCODE regions of the genome (16). We prepared HpaII and MspI representations from the GM06990 cell line using the optimized conditions described above (Figure 2d), labeling the representations by random priming and co-hybridizing the labeled representations to the microarray. This customized ENCODE array also included a set of 10 100 oligonucleotides representing random sequences. To assess the microarray data, we defined the background fluorescence level using a threshold of 2.5 median absolute deviations (2.5 MAD) above the median intensity of these control oligonucleotides. We have found this value to define consistently the distinctive population of loci in HpaII representations that does not amplify due to local methylation, and applied the same parameter to define signal intensities in the MspI channel, indicating loci that have failed to amplify adequately (Supplementary Figure 1). In the current experiment, we observed a failure rate of 5–9% in total, but only 2–4% were in the size range of 50–199 bp (Supplementary Figure 1). When the HpaII representation was studied, we observed a bimodal distribution, demonstrating that the two-adapter HELP method discriminates between methylated and unmethylated loci (Supplementary Figure 1).

Table 1. Annotation of human HpaII/MspI fragments in high-resolution HELP

<table>
<thead>
<tr>
<th></th>
<th>Original representation</th>
<th>Added representation</th>
<th>Combined representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>Size range/bp</td>
<td>200–2000</td>
<td>50–199 50–2000</td>
</tr>
<tr>
<td></td>
<td>Number</td>
<td>1 016 980</td>
<td>514 387 1 531 367</td>
</tr>
<tr>
<td>CpG islands</td>
<td>Number overlapped</td>
<td>23 881</td>
<td>23 716 27 035</td>
</tr>
<tr>
<td></td>
<td>Proportion represented</td>
<td>87.00%</td>
<td>86.40% 98.50%</td>
</tr>
<tr>
<td></td>
<td>Mean number of fragments/locus</td>
<td>1.6</td>
<td>3.9 5.5 9.8</td>
</tr>
<tr>
<td>CG clusters</td>
<td>Number overlapped</td>
<td>42 226</td>
<td>39 206 43 740</td>
</tr>
<tr>
<td></td>
<td>Proportion overlapped</td>
<td>95.10%</td>
<td>88.30% 98.60%</td>
</tr>
<tr>
<td></td>
<td>Mean number of fragments/locus</td>
<td>2.3</td>
<td>3.9 5.5 6.3</td>
</tr>
<tr>
<td>refSeq</td>
<td>Number overlapped</td>
<td>16 848</td>
<td>14 923 17 223</td>
</tr>
<tr>
<td>promoters*</td>
<td>Proportion overlapped</td>
<td>89.10%</td>
<td>78.90% 91.10%</td>
</tr>
<tr>
<td></td>
<td>Mean number of fragments/locus</td>
<td>2.6</td>
<td>4.2 6.8 8.8</td>
</tr>
</tbody>
</table>

*2 kb flanking transcription start site.
In addition, the scatterplots and Pearson’s correlation coefficients for log₂ MspI intensities and HpaII/MspI intensity ratios illustrated that both technical and biological replicates show high correlation in MspI intensities (98–99%) and HpaII/MspI ratios (95–99%) (Supplementary Figure 1). We conclude that the labeling and hybridization of the additional representation of shorter fragments is as efficient and reproducible as the original method.

We then used a genome-wide microarray consisting of >1.32 million loci for the human genome to test the cytosine methylation pattern in human embryonic stem cells. We show the results of the methylation patterns observed in Figure 3. We see that CG-dense loci [CpG islands or CG clusters (18)] and promoters are less methylated than the CG-depleted sequences within gene bodies and intergenic sequences. These are results concordant with prior expectations (22) and demonstrate the assay to be able to report expected patterns of methylation in a variety of sequence contexts.

Adaptation of protocol to limited starting amounts of DNA: nanoHELP

A limitation in many epigenomic assays is their requirement for large amounts of sample. As the HELP assay involves a PCR amplification step, we tested whether we could use substantially less starting amounts of DNA and still get reproducible results. We performed the assay using 1 μg of DNA as a reference sample and 10 ng of the same DNA for LM-PCR, using two rounds of amplification with 15 cycles per round to stay within the intensities (98–99%) and HpaII/MspI ratios (95–99%) (Supplementary Figure 1). We conclude that the labeling and hybridization of the additional representation of shorter fragments is as efficient and reproducible as the original method.

We then used a genome-wide microarray consisting of >1.32 million loci for the human genome to test the cytosine methylation pattern in human embryonic stem cells. We show the results of the methylation patterns observed in Figure 3. We see that CG-dense loci [CpG islands or CG clusters (18)] and promoters are less methylated than the CG-depleted sequences within gene bodies and intergenic sequences. These are results concordant with prior expectations (22) and demonstrate the assay to be able to report expected patterns of methylation in a variety of sequence contexts.

Figure 3. The distribution of cytosine methylation by genomic compartment. The distribution of cytosine methylation, using the new high-resolution genome-wide HELP microarray, is categorized as hypomethylated (white), methylated (black) and intermediate (grey) based on thresholds categorising the bimodal distribution of HpaII/MspI ratios. The results show relative hypomethylation of Cpg islands or CG clusters (18) whether at promoters, within gene bodies or intergenically compared with other sequences. This is consistent with expectations for the distribution of methylation and indicates that the assay is capable of defining methylation states in a range of genomic contexts.

Adaptation of protocol to limited starting amounts of DNA: nanoHELP

A limitation in many epigenomic assays is their requirement for large amounts of sample. As the HELP assay involves a PCR amplification step, we tested whether we could use substantially less starting amounts of DNA and still get reproducible results. We performed the assay using 1 μg of DNA as a reference sample and 10 ng of the same DNA for LM-PCR, using two rounds of amplification with 15 cycles per round to stay within the intensities (98–99%) and HpaII/MspI ratios (95–99%) (Supplementary Figure 1). We conclude that the labeling and hybridization of the additional representation of shorter fragments is as efficient and reproducible as the original method.

We then used a genome-wide microarray consisting of >1.32 million loci for the human genome to test the cytosine methylation pattern in human embryonic stem cells. We show the results of the methylation patterns observed in Figure 3. We see that CG-dense loci [CpG islands or CG clusters (18)] and promoters are less methylated than the CG-depleted sequences within gene bodies and intergenic sequences. These are results concordant with prior expectations (22) and demonstrate the assay to be able to report expected patterns of methylation in a variety of sequence contexts.

Adaptation of protocol to limited starting amounts of DNA: nanoHELP

A limitation in many epigenomic assays is their requirement for large amounts of sample. As the HELP assay involves a PCR amplification step, we tested whether we could use substantially less starting amounts of DNA and still get reproducible results. We performed the assay using 1 μg of DNA as a reference sample and 10 ng of the same DNA for LM-PCR, using two rounds of amplification with 15 cycles per round to stay within the intensities (98–99%) and HpaII/MspI ratios (95–99%) (Supplementary Figure 1). We conclude that the labeling and hybridization of the additional representation of shorter fragments is as efficient and reproducible as the original method.
linear stage of amplification. We show in Figure 4 that we can amplify the DNA robustly from the 10-ng samples, and that microarray hybridizations generate correlations of HpaII/MspI ratios measured by $R$-values in excess of 0.96. We conclude that HELP can be used on amounts of DNA that represent approximately the amount in 2000 cells, and is likely to be amenable to further scaling to even more limited starting amounts.

Simultaneous copy-number analysis: HELP-CNV

We investigated whether the HELP data could simultaneously be used to provide copy-number information, given that the MspI representation should be unaffected by cytosine methylation and instead dependent on the amount of DNA present. In a manner analogous to the ROMA technique (23), we compared the MspI (methylation-insensitive) representations of two different samples. One sample was a primary leukemia sample from a human subject that we have previously studied (24) and found to have copy-number changes by array comparative genomic hybridization (aCGH), the other was a normal CD34 + hematopoietic stem cell sample used as a control with presumed minimal copy-number variability. The copy-number variation analysis of HELP data (HELP-CNV) identified the same patterns of amplifications and deletions on chromosome 20 found using aCGH (Figure 5). The HELP-CNV approach allows the data generated as part of a HELP assay to harvest a second source of information about DNA copy number while testing cytosine methylation genome wide, a useful combination for diseases such as cancer.

HELP studies using massively parallel sequencing: HELP-seq

The development of massively parallel sequencing has provided an alternative to microarrays for measuring the enrichment of genomic representations following chromatin immunoprecipitation (ChIP-seq). We sought to test whether the HELP representations could allow similar mapping of hypomethylated loci. Contaminating fragments in the assay can be recognized because they lack the HpaII/MspI site remnant (CGG) that starts each sequence read following the digestion and end-polishing when creating the libraries. In Table 2 we show a summary of the results for each library, including a proportion (~3%) in each library that starts with a CGG but is not located at a canonical HpaII site (CCGG) in the reference genome sequence, indicating that we are identifying a substantial number of polymorphic HpaII sites in this ES cell DNA sample.

When we map these sites to the genome we observe ‘piling up’ of reads at HpaII sites, indicating their enrichment by the representation. We note that the control MspI representation is not enriched equally at all sites, a finding...
similar to those we have observed previously using microarray-based approaches (13). The HpaII representations are also heterogeneous but this is subject to the additional influence of cytosine methylation. We normalized the HpaII representation in terms of the MspI data and plotted wiggle tracks for the resulting data. In general, we found the microarray-based HELP and the HELP-seq data to be strongly concordant, as represented in Figure 6. We noted that the discordance of the assays appeared to be mostly in terms of HELP-seq identifying additional hypomethylated loci, such as the alternative, non-CpG island, non-CG cluster (18) promoter of KCNQ1, which appears to become hypomethylated in erythroid progenitor cells (Figure 6). We tested this locus using bisulphite MassArray and confirmed its hypomethylation (Figure 6), indicating that HELP-seq is more sensitive than microarray-based HELP in identifying hypomethylated loci. We note that HELP-seq allows mapping of sequences shorter than the 50-bp minimum that we can represent on a microarray, and can thus give more detailed information about CG-dense regions of the genome than when using microarrays.

DISCUSSION

Genomic representations by LM-PCR are used by a number of different applications, including representational oligonucleotide microarray analysis (ROMA) (23), high-density SNP microarrays (25) and other epigenomic assays testing cytosine methylation (26–28). The HELP assay uses representations from HpaII to distinguish methylated from unmethylated loci in the genome, with a concurrent MspI representation defining the full range of potential HpaII-amplifiable loci. The more fragments that can be represented, the greater the level of detail we can achieve in ROMA, SNP or epigenomic analyses. By increasing the representation of shorter fragments using dual adapters and modified PCR conditions, we achieve greater coverage of CG-dense regions in particular. By computational analysis, we show that the proportional coverage for CpG islands or our updated definition of CG clusters (18) approaches the maximum possible (98.5% and 98.6%, respectively, Table 1). In addition, the number of fragments at each CG-dense locus and refSeq promoter is increased 2- to 3-fold compared with the previous representation, which enables a more detailed analysis of DNA methylation in these promoter regions. We demonstrate that these shorter fragments (50–200 bp) can be labeled with the random priming technique and hybridized to microarrays with signal ranges comparable to those of larger fragments. Our practical lower limit for microarray analysis is constrained by the size of the oligonucleotides we use (≥50 nt), making it difficult to represent fragments smaller than the oligonucleotides themselves. For massively parallel sequencing applications, the sequences generated are
generally shorter (≈35 bp) and need only be long enough
to allow accurate mapping to the reference genome,
increasing the number of testable loci compared with microarrays.
This new representation allowed us to design high-
resolution HELP microarrays for the human, mouse, rat
and cow genomes, each representing over 1 million loci
throughout these genomes. By using a high-density plat-
form, we can represent these loci on a single microarray,
allowing analysis of the entire genome in a single hybrid-
ization experiment. We performed a high-resolution
HELP assay on human ES cells to test how the cytosine
methylation patterns observed correlate with genomic
annotations. We found relative hypomethylation of CpG
islands and CG clusters compared with other sequences,
and less methylation at promoters compared with gene
bodies or intergenic regions, patterns consistent with
prior observations about the distribution of cytosine
methylation (22).

When moving epigenomic assays to the clinical setting,
a critical issue to address is the potential to use the limited
sample amounts that can be acquired from biopsies.
As the generation of genomic representations in the
HELP assay involves PCR, we tested whether the ampli-
fication step could allow us to generate adequate amounts
of material from limited starting amounts of DNA.
We used 10 ng of DNA in our studies as an amount repre-
senting approximately 2 \times 10^3 cells, and found the profiles
generated to be highly comparable with those generated
using our usual microgram quantities of starting material.
We conclude that the HELP assay can be used on rela-
tively limited numbers of cells. Our ongoing studies are
testing whether we can decrease the starting amounts of
DNA still further.

We also found that the MspI representation on its own
allows an accurate and detailed analysis of copy number.
This approach is similar to the published ROMA tech-
nique (23) in that it uses genomic representations by
methylation-insensitive restriction enzymes, but is sub-
stantially higher resolution with \(|>1.32 million loci repre-
sented. With MspI/HpaII sites more frequent in (C + G)
mononucleotide-rich regions which tend also to be
more gene-rich (9), the MspI representation offers
selectively higher resolution in gene-rich regions,
which is potentially advantageous. The HELP-CNV appli-
cation allows DNA to be tested for cytosine methylation
and copy-number variation simultaneously, a valuable
tool especially in cancer research, in which epigenomic
alterations (29) and copy-number changes (30) are
frequent.

A major advantage of HELP-seq compared with micro-
array-based HELP appears to be one of sensitivity of
detection of hypomethylated loci. We believe that this is
due to the lower background noise for sequencing com-
pared with the background noise for microarray-based HELP.
There are other potential advantages of HELP-seq, including
the capacity to detect allelic differences, especially in cancer
samples, and the ability to detect events in repetitive
regions smaller than the 50-bp lower limit for microarray
microarrays and allowing the number of testable loci compared with
increasing the number of testable loci compared with
increasing the number of testable loci compared with
microarrays.

### Table 2. HELP-seq data

<table>
<thead>
<tr>
<th></th>
<th>Total reads</th>
<th>Low quality</th>
<th>Not aligned</th>
<th>Unique aligned read</th>
<th>Number of unique aligned reads starting with CGG</th>
<th>Number of unique aligned reads starting with CGG at annotated HpaII sites</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MspI ES</strong></td>
<td>3,510,528</td>
<td>40,179</td>
<td>754,146</td>
<td>473,994</td>
<td>2,342,209</td>
<td>2,205,347</td>
</tr>
<tr>
<td><strong>HpaII ES</strong></td>
<td>5,872,425</td>
<td>65,788</td>
<td>1,358,445</td>
<td>88,096</td>
<td>4,360,096</td>
<td>4,232,845</td>
</tr>
<tr>
<td><strong>HpaII EP</strong></td>
<td>7,373,981</td>
<td>78,564</td>
<td>1,789,257</td>
<td>254,132</td>
<td>5,616,028</td>
<td>5,562,184</td>
</tr>
</tbody>
</table>

In the upper panel, we show the results of analysis of all of the reads obtained from the HELP-seq experiment. Most sequences mapped uniquely to the reference human genome, of which the vast majority had the CGG remnant of the original HpaII/MspI digestion. Multiple coincident reads defined the distinct loci described in the lower panel. Again, most loci had the CGG sequence and mapped to HpaII sites in the reference genome, but ~3% of the distinct loci had evidence of an original HpaII site where none exists in the reference genome, defining polymorphic HpaII sites in our cell sample.
Figure 6. Adaptation of the HELP assay to use massively parallel sequencing: HELP-seq. Instead of using microarrays to map the representations to their genomic source, we sequenced the ends of the fragments using massively parallel sequencing (Illumina). We show the promoter region of the KCNQ1 gene as an example of the concordant data obtained, with hypomethylated sites represented as an upward/positive peak and methylated as downward/negative, in both the microarray-based (HELP) and Illumina-based (HELP-seq) assays. The HELP data are represented by quantile-normalized HpaII/MspI log2 ratios (13) while the HELP-seq data are represented by the number of HpaII reads per locus normalised to MspI reads for the same locus as described in the main text. Both embryonic stem (ES, red) and derived in vitro-differentiated erythroid progenitor (blue) cells are shown. The upstream (left) promoter of KCNQ1 is hypomethylated in both cell types, with a broader region of hypomethylation in erythroid compared with ES cells in both the HELP and HELP-seq studies. The downstream promoter appears to be methylated in ES cells in both assays, but microarray-based HELP weakly indicates hypomethylation in erythroid cells, whereas HELP-seq shows a strong hypomethylated signal in these cells. To determine which technique was more accurate, we performed validation using bisulphite MassArray (bottom), confirming the underlying hypomethylation of this locus in the erythroid and not ES cells, and demonstrating that HELP-seq offers improved capacity to identify sites of hypomethylated DNA in the genome.
studies. We also note that we identified a substantial number of HpaII sites (~3%) that are not present in the reference human genomic sequence, variability that would not be captured by or that would cause errors in microarray-based approaches.

The ideal assay to test cytosine methylation would test every CG dinucleotide individually and quantitatively throughout the genome, preserving information about cis-relationships of methylation states between CGs, and allowing high sample throughput. No such assay exists at present. While massively parallel sequencing-based approaches promise to make nucleotide-resolution studies possible (2,5,6,31), at present their sample quantity demands and costs remain daunting. The HELP assay falls into a category of assays that act to screen the genome at lower resolution, the ‘discovery’ step that defines loci for more detailed, nucleotide-resolution studies. Other assays in this category include methylation-dependent restriction enzyme assays (32) and affinity-based assays using antibodies (1) or other natural methyl-binding proteins (3). The comparative advantages of HELP include its capability of using a single array and the easy technical validation of results, as the focus is solely on methylation at the HpaII/MspI sites (CCGG) generating the representations. While HpaII/MspI sites constitute only ~8% of the CG dinucleotides in the genome (9), the presence of methylation ‘states’ in cis that may extend over as much as 1 kb (33) allow discovery assays to flag interesting regions by testing a subset of CGs. When we measure the proportion of CGs in the human genome residing in proximity to the HpaII sites on the HELP microarray we used, we find that approximately two-thirds are within 1 kb of these sites (Supplementary Figure 2). While epigenomic discovery approaches directly test only a minority of CGs, they have the potential of ‘flagging’ the majority of CGs in the genome. We conclude that the high-resolution HELP assay is technically simple and robust and offers the capacity to test both the epigenome and copy-number variability, and has the potential for use with limited numbers of cells and adaptation to massively parallel sequencing platforms.

**ADDENDUM**

While this manuscript was in review, an assay fundamentally similar to HELP-seq, described as Methyl-seq, was published by Brunner et al. (34).

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online.

**ACKNOWLEDGEMENTS**

The authors acknowledge the support of the Einstein Center for Epigenomics, the Bioinformatics Shared Resource and the Genomics Core Facility.

**FUNDING**

The National Institutes of Health (NIH) (NHGRI) R01 HG004401 to J.M.G, (NHLBI) R01 HL082946 to A.V., (NIA) P01 AG027734 and R01 AG18728 to R.F.T., and (NIGMS) T32 GM007288 to J.L.G.; the Japan Society for the Promotion of Science (to M.O.). A Leukemia and Lymphoma Society Translational Research Award and funding from the Hershert Foundation (to A.V.) and the Leukemia and Lymphoma Society (to A.M.). Funding for open access charge: NHGRI R01 HG004401.

Conflict of interest statement. None declared.

**REFERENCES**


