Efficient whole-genome DNA methylation analysis of the Human Reference Genome (HuRef)

Gavin D. Meredith, Miroslav Dudas, Elizabeth Levandowsky, Tamara Gilbert, Daniel Krissinger, George Marnellos, Vrunda Sheth, Christopher Adams, Clarence Lee, and Timothy Harkins , Life Technologies / R&D, 5791 Van Allen Way, Carlsbad, CA, USA, 92008

ABSTRACT

Aberrant DNA methylation is characteristic of many cancers and differences in methylation have been observed in a wide variety of genomic contexts; for example, both within "classic" promoter-associated CpG islands and also in distal, non-CpG island regions [1, 2]. Establishing a method to broadly and efficiently survey DNA methylation patterns genome-wide is the objective of the work presented. The method combines the power of methyl-CpG binding domain (MBD) proteins to sensitively and selectively bind methylated DNA sequences with the coverage, precision, and accuracy provided by high-throughput sequencing. Notably, MBD-affinity capture can also be used to sub-fractionate genomic DNA based on its average methyl-CpG content. To illustrate this method, Human Reference Genome (HuRef) DNA [3] was enriched and salt-fractionated with a commercial MBD-based affinity reagent and high-throughput sequencing libraries, both bisulfite converted and unconverted, were prepared from each of the three fractions. The libraries were sequenced using error correcting codes and paired-end technology that yielded 75 bp read-lengths from one end and 30 bp read-lengths from the opposite end on a SOLiD™ 4 System. Pilot analysis of IMR-90 fibroblast cell-line DNA and comparison to its published methylome established the specificity of this approach and the feasibility of obtaining high-quality bisulfite sequencing of human methylation patterns with the SOLiD™ System [4]. Peak analysis of the distribution of mapped unconverted reads can permit the discovery of thousands of locations of putative methylation in different genomic locus classes; the low salt fractions were depleted of CpG islands and enriched for exons while the highest salt fraction was enriched for CpG islands, exons, and promoter regions. Methylation at a large number of these positions can be confirmed by bisulfite-sequencing of the same libraries. We conclude that such enrichment and fractionation, when coupled to high-throughput sequencing with or without bisulfite conversion, can be used to efficiently survey the majority of DNA methylation marks within samples of genomic DNA and to discover genomic loci of differential methylation. This development of this method and reference dataset are intended to provide the community with tools for large-scale methylation studies.

INTRODUCTION

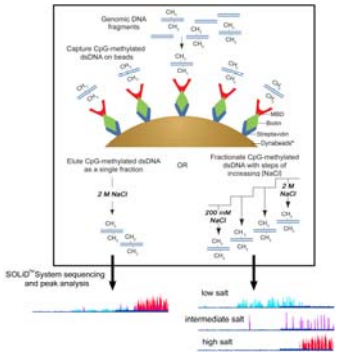
DNA methylation plays a critical role in gene regulation that influences normal organism development and many diseases including cancer. Profiling the DNA methylation patterns of higher organisms is challenging because methylation patterns vary between tissues and with developmental state, hence there are far more methylomes to be analyzed than genomes. Furthermore, in order to map methylation positions with high precision and accuracy, greater depth of sequencing is required than for normal genome sequencing [5]. Affinity-based enrichment of methylated DNA sequences prior to high-throughput sequencing as with the SOLiD™ System provides an avenue to pursue this kind of genome-wide information in a minimally biased and cost-efficient manner. The workflow described here using MethylMiner™ enrichment with stepwise salt gradient elution enables the partitioning of the genome into regions of low, moderate, and high density of methylation. This permits blind discovery of methylated regions and permits detection of differentially methylated regions (DMRs) between samples and across genomic feature subsets that harbor differing degrees of methylation density.

MATERIALS AND METHODS

Methylated DNA enrichment and SOLiD™ sequencing.
Genomic DNA from cultured IMR-90 cells was purified with PureLink® columns. Purified HuRef genomic DNA was purchased from the Coriell Institute for Medical Research. Genomic DNA was fragmented to 50-400 bp (mean ~250 bp) with a Covaris™ S2 System (Woburn, MA). Methyl-CpG binding-domain protein affinity capture was with MethylMiner™ Methylated DNA Enrichment kits (Invitrogen, Carlsbad, CA) following the manufacturer's protocol. For salt-gradient elution of IMR-90 DNA, successive fractions were obtained by elution using buffer containing the following NaCl concentrations: 200 mM, 350 mM, 450 mM, 600 mM, and finally 2 M NaCl. For HuRef DNA successive elutions were done with buffer containing 450 mM, 600mM, and 2M NaCl. Each elution step consisted of 2–3 serial incubations of the MethylMiner™ beads at each salt concentration. Following ethanol precipitation the methylated DNA was resuspended in 30-75 uL GIBCO® DNAse- and RNAse-free water (Invitrogen) and quantified by UV absorbance spectroscopy with a NanoDrop® 1000 instrument. The DNA was used for coupled SOLiD™ standard fragment and bisulfite-sequencing (BS-Seq) library construction as shown in Figure 2. SOLiD™-BS-seq on methylation-enriched *A thaliana* eukaryotic DNA is described in the Methods section of Ondov et al and is based on a workflow described by Borman Chung et al [6, 7]. A protocol for this procedure will be available on-line from Life Technologies in November 2010. Each library was sequenced within a "quad" chamber.

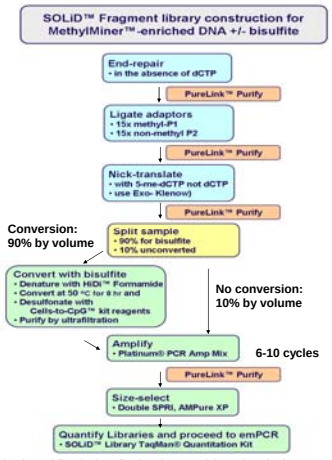
Whole-genome SOLiD™ bisulfite sequencing.
Genomic DNA from cultured IMR-90 cells was purified with PureLink® columns. Genomic DNA was fragmented to 50-400 bp (mean ~250 bp) with a Covaris™ S2 System (Woburn, MA). The workflow is depicted in Figure 2. Five (5) micrograms of genomic DNA fragments was end-repaired using reagents from a SOLiD™ Fragment Library Construction kit with a dNTP-mix lacking dCTP, then PureLink® column purified. A P1 adaptor containing 5-methyl-Cs in its 41 bp strand and a normal P2 adaptor were used in a standard ligation reaction where each adapter was in 15-fold molar excess over end-polished fragments. After re-purification, nick-translation was done with Exo- Klenow enzyme and dNTPs containing 5-methyl-dCTP instead of dCTP. The nick-translation reaction was purified once again and the DNA diluted 2-fold with HiDi™ formamide prior to bisulfite treatment with reagents from a Cells-to-CpG™ Bisulfite Conversion kit at 50 ºC for 8 hours. After desulfonation and purification on an Amicon Ultra-0.5 10 kD device, the converted library DNA was amplified with Platinum® PCR Amp mix for 8 cycles then purified with AMPure XP beads using a double-SPRI protocol.

Figure 1. MethylMiner™ Kit Fractionation of CpG-methylated DNA for SOLiD™ System Sequencing



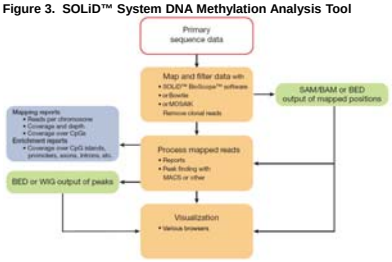
Fragmented double-stranded CpG-methylated genomic DNA is directly and specifically captured on MethylMiner™ MBD-coated magnetic beads then eluted all-at-once with buffer containing 2M NaCl or separated into complementary fractions by step-wise elution with buffers containing progressively increasing concentrations of NaCl up to 2 M. Sequencing after single-step elution using 2M NaCl shows greatest enrichment for densely methylated regions of the sample (lower left). Elution using step-wise salt gradient buffers helps to provide subsets of the methylome with differing degrees of methylation density; sparsely methylated fragments (light blue) elute with low salt, more densely methylated fragments (purple) elute with higher salt, and heavily methylated fragments (red) elute at maximal salt. Selective enrichment prior to sequencing permits clearer identification of these subclasses of methylated genomic DNA fragments.

Figure 2. Workflow for coupled preparation of unconverted and bisulfite-converted SOLiD™ fragment libraries



This basic workflow is described under Materials and Methods. It describes the steps for creating coupled unconverted and bisulfite converted libraries from MethylMiner™-enriched DNA fragments and also applies to whole-genome bisulfite-converted library construction. A protocol for this procedure will be available on-line from Applied Biosystems/Life Technologies in November 2010.

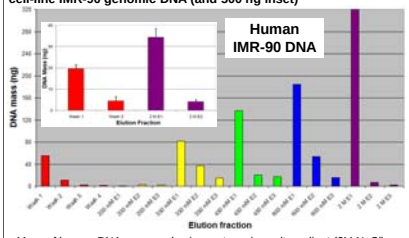
Figure 3. SOLiD™ System DNA Methylation Analysis Tool



Data Analysis Pipeline.
A data analysis pipeline can be assembled using SOLiD™ System software (e.g. SOLiD™ BioScope™ Software), free public tools (e.g., Bowtie [8], SAMtools, MACS), and auxiliary scripts and programs (this is described more extensively in a recent Life Technologies Application Note entitled "Genome-wide methylation analysis at 150 bp resolution using the MethylMiner™ kit and SOLiD™ System sequencing). The pipeline can run on a Linux® desktop with multiple processors and sufficient RAM and disk space; minimum recommended requirements include 4–6 processors, 16GB RAM, and 2 TB storage. With Bowtie, SOLiD™ reads were mapped against the hg19 reference genome build. When using BioScope™, HuRef reads were mapped against the HuRef genome. The MACS [9] algorithm is called for detection of peaks of methylated sequence enrichment. This pipeline will be made available for download from Applied Biosystems/Life Technologies in November 2010.

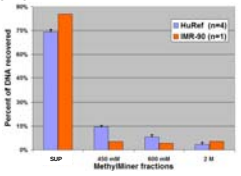
RESULTS

Figure 4. MethylMiner™ Elution profile from 25 µg of human cell-line IMR-90 genomic DNA (and 500 ng inset)



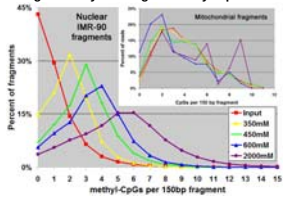
Mass of human DNA recovered using a step-wise salt gradient (2M NaCl) elution. The majority of input DNA is not captured on the MethylMiner™ beads because only about 0.5-1% of all bases are 5-methyl-C in CpG dinucleotides. This fact, and the observation that diminishing amounts of DNA are recovered in successive wash and step-wise elutions (e.g., fractions 350a, 350b, and 350c in Fig. 2A-B), indicate that the methylated DNA–enrichment protocol worked properly. Importantly, as shown in the inset of Figure 2, MethylMiner™ enrichment from as little as 500 ng of fragmented genomic DNA can yield enough methylated DNA for SOLiD™ library construction.

Figure 5. MethylMiner™ Elution profiles from 25 µg of human cell-line genomic DNA



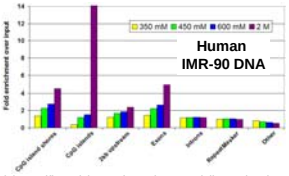
Relative mass of human DNA recovered using a step-wise salt gradient (2M NaCl) elution. The majority of input DNA is not captured on the MethylMiner™ beads because only about 0.5-1% of all bases are 5-methylC in CpG motifs. Typically 5-15% of input mass is recovered as the methylated subset.

Figure 6. Higher salt yields higher methyl-CpG content



Counts of the number of methyl-CpG dinucleotides within 150 bp downstream of uniquely-mapped sequencing start-sites show that differing subsets of the genome are obtained with salt fractionation. Importantly, the number of sequenced fragments showing 0 CpG content drops with enrichment and the hypo-methylated human mitochondrial reads (inset) show no significant changes in their distributions; they are suppressed >50-fold with enrichment (not shown) indicating strong specificity of capture for CpG-methylated DNA.

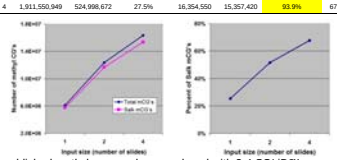
Figure 7. Enrichment trends for human genome features



Differential, specific enrichment for various, partially overlapping, annotated genomic features is obtained with salt-fractionation. Notably, CpG islands, shores, and exonic sequences increase in relative representation with increasing ionic strength. Single-fraction elution with high-salt (2M NaCl) yields an average behavior in this respect (not shown).

Table 1 and Figure 8. SOLiD™ Bisulfite sequencing of IMR-90 fragment library yields high concordance with published methylome

# slides	Total Reads	# Uniquely Mapped	Uniquely Mapped (%)	Total mCG's (this work)	Salk mCG's	Concordance with Salk mCG's	% Salk mCG's
1	495,265,636	151,321,772	30.6%	5,632,473	5,339,933	94.8%	23.6%
1	515,699,253	160,975,121	31.2%	6,096,313	5,770,102	94.6%	25.5%
2	1,010,964,889	312,296,893	30.9%	12,359,341	11,692,335	94.6%	51.6%
4	1,911,550,949	524,998,672	27.5%	16,354,550	15,357,420	93.9%	67.8%



The published methylome can be reproduced with 3-4 SOLiD™ runs or about 4 billion raw 50 bp reads.

Table 2. Statistics for 50-color F3 reads from "quad" chambers mapped with Bowtie on a basic Linux workstation.

Dataset	# reads processed	# of uniquely mapped reads	% reads uniquely mapped	unique starts	alignments per start	starts*read-length (Gb)
IMR-90 unconverted SOLiD 3Plus Opti-reads trimmed to 50bp						
Input	86,033,985	52,804,281	61.4%	48,126,872	1.10	2.41
350 mM	86,942,318	44,292,234	50.9%	37,632,630	1.18	1.88
450 mM	83,790,547	40,991,116	48.9%	34,193,582	1.20	1.71
600 mM	84,897,767	40,368,266	47.5%	33,372,789	1.21	1.67
2 M	84,851,179	27,002,327	31.8%	21,223,792	1.27	1.06
HuRef unconverted SOLiD 4 ToP-reads trimmed to 50bp						
HuRef SUP	106,973,202	65,567,763	61.3%	56,607,178	1.16	2.83
HuRef 450 mM	110,518,004	70,647,916	63.9%	61,878,446	1.14	3.09
HuRef 600 mM	109,445,214	56,638,857	51.8%	48,225,810	1.17	2.41
HuRef 2M	102,626,098	45,932,465	44.8%	36,399,287	1.26	1.82
HuRef BS-converted SOLiD 4 ToP-reads trimmed to 50bp (non-CpG CZT converted reference)						
HuRef SUP	106,495,397	33,434,814	31.4%	29,037,672	1.15	1.45
HuRef 450 mM	110,953,701	35,321,299	31.8%	30,614,648	1.15	1.53
HuRef 600 mM	103,354,357	28,971,297	28.0%	23,771,167	1.22	1.19
HuRef 2M	104,090,821	24,447,215	23.5%	18,818,924	1.30	0.94

Mapping rates and library complexity decrease with increasing ionic strength of elution. Bisulfite-converted F5 reads map nearly as efficiently as unconverted reads to an *in silico* reference that assumes all non-CpG Cs get converted to Ts. ToP probes and SOLiD™ 4 slide-densities improve yields too.

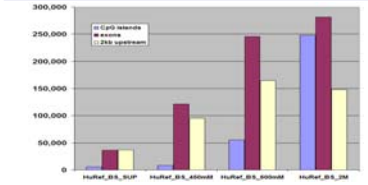
Table 3. Paired-end mapping stats for HuRef libraries with and without ECC (Exact Call Chemistry)

sample	F3	F5	non redundant pairings	reads per start point -F3	reads per start point -F5
HuRef_BS_SUP	53,577,466 (50.76%)	49,469,064 (46.87%)	27,397,712	1.15	1.12
HuRef_BS_SUP_ECC	56,305,625 (52.87%)	49,469,064 (46.87%)	28,946,054		
HuRef_BS_450mM	53,854,619 (48.95%)	46,488,311 (42.26%)	26,415,015	1.15	1.1
HuRef_BS_450mM_ECC	55,628,891 (50.14%)	46,488,311 (42.26%)	21,528,619		
HuRef_BS_600mM	48,413,158 (47.45%)	34,560,613 (33.87%)	13,143,685	1.15	1.11
HuRef_BS_600mM_ECC	51,790,121 (50.11%)	34,560,613 (33.87%)	13,922,181		
HuRef_BS_2M	36,350,685 (35.18%)	22,026,902 (21.32%)	5,272,821	1.14	1.19
HuRef_BS_2M_ECC	40,118,031 (38.54%)	22,026,902 (21.32%)	5,712,030		
sample	**F3**	**F5**	**non redundant pairings**	**reads per start point -F3**	**reads per start point -F5**
HuRef_MM_SUP	81,428,617 (76.65%)	45,591,862 (42.97%)	45,682,214	1.23	1.17
HuRef_MM_SUP_ECC	85,086,709 (79.54%)	45,591,862 (42.97%)	47,268,892		
HuRef_MM_450mM	91,564,604 (83.99%)	48,130,909 (43.94%)	49,852,817	1.2	1.14
HuRef_MM_450mM_ECC	92,977,698 (84.13%)	48,130,909 (43.94%)	50,783,640		
HuRef_MM_600mM	85,394,256 (79.47%)	44,713,705 (41.61%)	40,183,342	1.29	1.2
HuRef_MM_600mM_ECC	88,131,539 (80.53%)	44,713,705 (41.61%)	41,132,928		
HuRef_MM_2M	81,505,487 (81.17%)	42,030,634 (41.86%)	34,437,218	1.19	1.13
HuRef_MM_2M_ECC	83,786,530 (81.65%)	42,030,634 (41.86%)	35,274,270		

Paired-end sequencing (75x35) shows very high rates of non-redundant pairings for unconverted MethylMiner™-enriched libraries. ECC consistently permits higher yields of unique mapping too. For bisulfite-converted libraries, there is a reduction in pairing efficiency at higher ionic strengths, presumably due to the reduced sequence complexity both of the selected fragments and caused by bisulfite treatment.

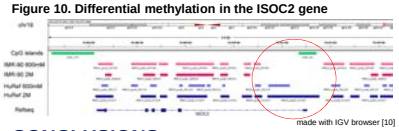
Table 3 and Figure 9. Methyl-cytosines in HuRef samples detected using BioScope™ SNP-calling

sample	methylCs	me-CpGs	me-CpHs	overlap with Salk	CpG islands	exons	2kb upstream
HuRef_BS_SUP	1,640,358	1,635,245	5,113	1,477,719	5,797	36,377	36,832
HuRef_BS_450mM	5,232,402	5,230,498	1,904	4,840,232	8,211	121,714	95,167
HuRef_BS_600mM	6,483,756	6,481,321	2,435	5,928,973	55,222	245,710	164,585
HuRef_BS_2M	4,205,183	4,202,955	2,228	3,704,143	247,953	281,384	147,641



Since bisulfite-conversion can be thought of as the genome-wide chemical induction of non-methyl-C>T SNPs, the tools within BioScope™ can be used to identify methylated Cs when an *in silico* fully-converted reference is used. An accounting of all methyl-Cs detected in the HuRef samples is given in the table and the distribution of methyl-Cs in CpG islands, exons and regions within 2 kb upstream of TSSs is graphed. Approximately 16 million methyl-Cs are detected, ~90% of which are found in the MethylMiner™-enriched libraries each sequenced within a "quad" chamber. As seen for IMR-90 DNA, there is specific enrichment in regions corresponding to these genomic elements; however, a large number of methyl-Cs clearly reside elsewhere throughout the genome.

Figure 10. Differential methylation in the ISOC2 gene



made with IGV browser [10]

CONCLUSIONS

1. Bisulfite-sequencing of human methylomes is tractable on the SOLiD™ 4 System; it requires 3-4 full runs per methylome. The results are highly concordant with published data.
2. MBD-based enrichment of methylated sequences with the MethylMiner™ kit is an efficient means to focus of SOLiD™ System sequencing on genomic feature subsets.
3. MBD-based enrichment permits sub-fractionation of the genome based on the density of methylated CpGs. The methylation density is directly correlated to the ionic strength required for elution.
4. Methylated DNA enrichment can be coupled with bisulfite-sequencing. This permits single-nucleotide resolution validation of specific positions of methylation. This permits at least 4-fold reduction in sequencing cost.

REFERENCES

1. Irizarry RA, et al *Nat Genet.* 2009; **41**:178-86
2. Schmidl C, et al *Genome Res.* 2009; **19**:1165-74
3. Levy S, et al PLoS Biology 2007; **5**:2113-44
4. Lister R, et al *Nature* 2009; **462**:315-22
5. Zhang Y and Jeltsch A *Genes* 2010; **1**:85-101
6. Borman Chung CA, et al *PLoS One* 2010; **5**:e9320
7. Ondov BD, et al *Bioinformatics* 2010; **26**:1901-2
8. Langmead B, et al *Genome Biol.* 2009; **10**:R25
9. Zhang et al. Genome Biol. 2008; **9**:R137
10. http://www.broadinstitute.org/igv

ACKNOWLEDGEMENTS

© 2010 Life Technologies Corporation. All rights reserved.