Enzymatic Methyl-seq: Next Generation Methylomes

L. Williams, V.K.C. Ponnaluri, B.S. Sexton, L. Saleh, K. Marks, M. Samaranayake, L. Ettwiller, S. Guan, H.E. Church, N. Dai, E. Tamanaha, E. Yigit, B. Langhorst, M. Campbell, Z. Sun, T.C. Evans, R. Vaisvila, E.T. Dimalanta, T.B. Davis

New England Biolabs, Ipswich, MA 01938, USA

INTRODUCTION

Cytosine methylation is an important regulatory mechanism in many cellular processes. Bisulfite sequencing is the current gold standard for detecting methylated cytosines. Despite its widespread utilization, bisulfite sequencing can be problematic. DNA is commonly degraded by the harsh, chemical bisulfite reaction and the sequences obtained are frequently GC-biased.

Here we present NEBNext® Enzymatic Methyl-seq (EM-seq™), a novel enzyme-based method, that overcomes the limitations of bisulfite treatment for methylome analysis. Enzymatic Methyl-seq and WGBS (whole genome bisulfite sequencing) libraries were prepared using 10 ng, 50 ng and 200 ng NA12878 human genomic DNA and were used to determine cytosine methylation. BWAMeth was used to align and MethylDackel was used to call methylation levels from 2x100 base paired reads sequenced using the Illumina® NovaSeq® 6000. EM-seq data were in agreement with the expected results from previously characterized genomes as well as the control genomes studied. In addition, genome coverage was more even when compared to bisulfite sequencing. DNA damage is reduced which results in longer sequencing reads. Minimal GC bias is also observed as this method does not preferentially enrich for methylated regions.

EM-seq is an exciting new option for methylation analysis. Compared to WGBS, it is more robust and works over a wide range of DNA inputs, has superior sequencing metrics, and detects more CpGs over a wide range of genomic features.

METHODS

SAMPLE PREPARATION

Sheared Genomic DNA → Ultra II End Prep/ Ligation → TET2/ Oxidation Enhancer → APOBEC → PCR Amplification → Illumina Sequencing

- 10 ng, 50 ng and 200 ng of NA12878 genomic DNA, spiked with unmethylated lambda DNA, was sheared using the Covaris® S2 instrument
- DNA was end repaired and ligated to EM-seq adaptors
- 5mC and 5hmC were protected from APOBEC deamination by TET2/ Oxidation Enhancer
- Cytosines were deaminated to uracils with APOBEC
- Libraries were amplified with NEBNext Q5U™ Master Mix and Unique Dual Index Primer Pairs
- Libraries were sequenced using an Illumina NovaSeq 6000, 2x100 base paired reads
- Bisulfite conversion was performed using Zymo Research EZ DNA Methylation-Gold™ kit

DATA ANALYSIS

Trim → Align (BWAMeth) → Deduplicate → Picard → Methyl Dackel → Methyl Kit

- Reads were aligned to hg38 using BWAMeth
- Methylation numbers were extracted using MethylDackel
- Correlation analysis at 1x and 8x minimum coverage used methylKit 1.4.0
- Picard 2.17.2 was used for determining library insert size and GC bias

RESULTS

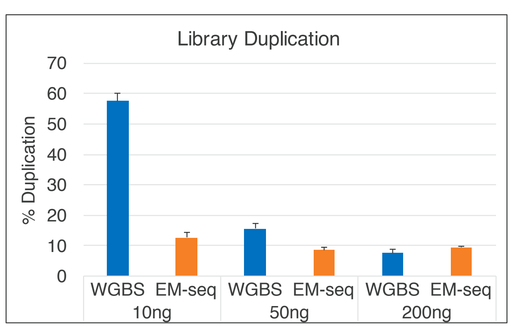
HIGHER QUALITY SEQUENCING DATA WITH EM-seq LIBRARIES

A

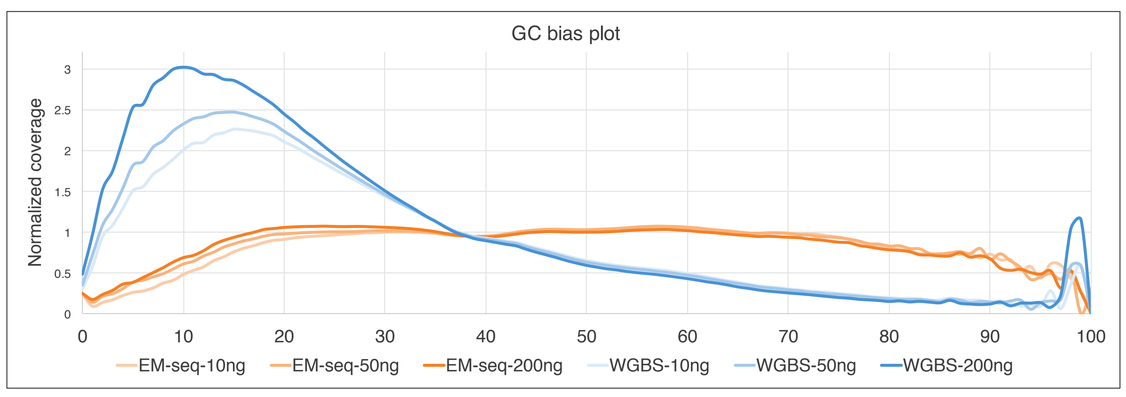
	PCR CYCLES		YIELD (nM)	
DNA inputs (ng)	EM-seq	WGBS	EM-seq	WGBS
10	8	10	15	3
50	6	8	22	6
200	4	6	63	42

B

C



D



EM-seq and WGBS metrics from 10 ng, 50 ng and 200 ng NA12878 genomic DNA. Each library was sequenced using the Illumina NovaSeq 6000. 324 million, 2 x100 base reads were used for methylation analysis. (A) EM-seq libraries have higher yield but require fewer PCR cycles. (B) EM-seq library insert sizes are larger than bisulfite libraries. EM-seq libraries peak at 170 bp compared to 100 bp for bisulfite libraries. (C) Library duplication percentages are lower for EM-seq and (D) the GC distribution of EM-seq and bisulfite libraries indicate that EM-seq libraries show more even coverage than bisulfite libraries. The bisulfite libraries are AT rich and have lower GC coverage.

SIMILAR GLOBAL METHYLATION LEVELS BETWEEN EM-seq AND WGBS

		% Methylation (10 ng)			% Methylation (50 ng)			% Methylation (200 ng)		
		CpG	CHG	CHH	CpG	CHG	CHH	CpG	CHG	CHH
NA12878	EM-seq	52.75 ± 0.07	0.55 ± 0.07	0.55 ± 0.07	52.15 ± 0.07	0.1 ± 0.00	0.1 ± 0.00	52.2 ± 0.00	0.1 ± 0.00	0.1 ± 0.00
	WGBS	53.85 ± 0.07	0.2 ± 0.00	0.2 ± 0.00	53.9 ± 0.14	0.2 ± 0.00	0.2 ± 0.00	54.15 ± 0.07	0.2 ± 0.00	0.2 ± 0.00
Lambda	EM-seq	0.55 ± 0.07	0.65 ± 0.07	0.6 ± 0.00	0.2 ± 0.00	0.15 ± 0.07	0.1 ± 0.00	0.1 ± 0.00	0.1 ± 0.00	0.1 ± 0.00
	WGBS	0.2 ± 0.00	0.2 ± 0.00	0.2 ± 0.00	0.2 ± 0.00	0.2 ± 0.00	0.25 ± 0.07	0.25 ± 0.07	0.25 ± 0.07	0.2 ± 0.00

The percentage methylation for 10 ng, 50 ng and 200 ng NA12878 and unmethylated lambda DNA in CpG/CHG/CHH contexts. NA12878: CpG methylation levels are similar for all libraries. Unmethylated Lambda: <1% methylated Cs in CpG, CHG and CHH were detected for all libraries.

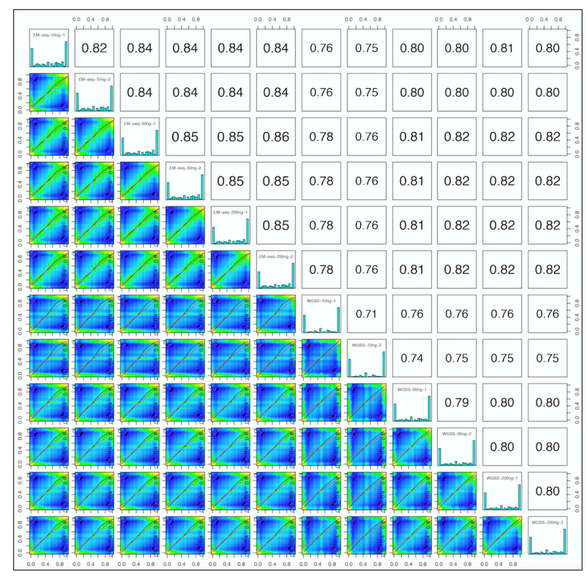
EM-seq LIBRARIES SHOW HIGHER CORRELATIONS THAN WGBS

A Total unique CpGs identified

		1x coverage			8x coverage		
	Input (ng)	10	50	200	10	50	200
CpGs covered (millions)	EM-seq	53.7	54.2	54.1	11	16.6	15.3
	WGBS	36	44.6	44.8	1.6	6.6	7.1
	Common	35.7	44.5	44.6	0.75	4.2	4.3

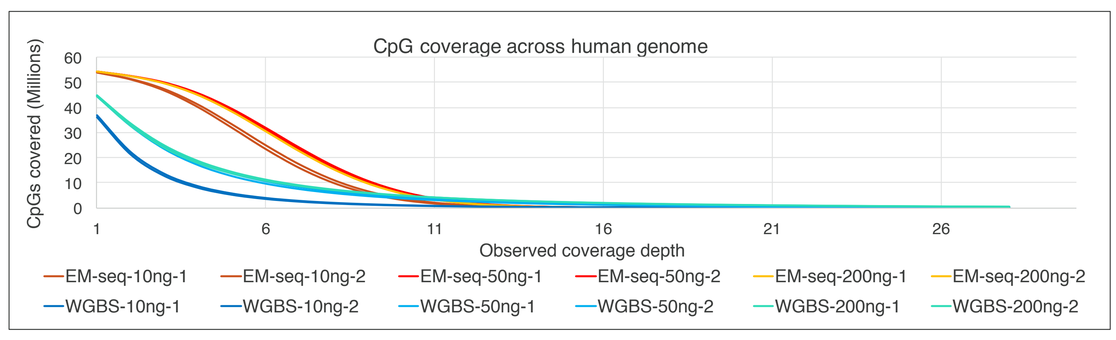
Top and bottom strand CpGs were counted independently, yielding a maximum of 56 M possible CpG sites. (A) Unique CpGs identified at 1x and higher or 8x and higher coverage for EM-seq, WGBS libraries and CpGs common to both. EM-seq detects more CpGs than WGBS. (B) Correlations were plotted using methylKit for 10 ng, 50 ng and 200 ng EM-seq and bisulfite libraries at 1x minimum coverage (21 million CpGs common to all libraries).

B

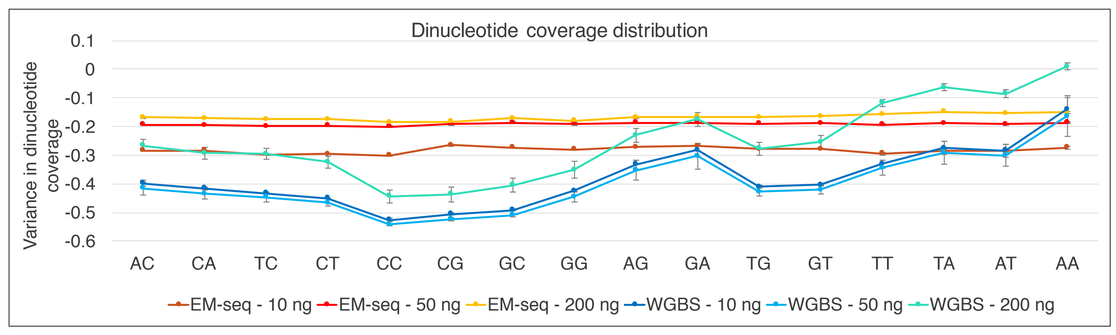


INCREASED CpG COVERAGE FOR EM-seq LIBRARIES

A

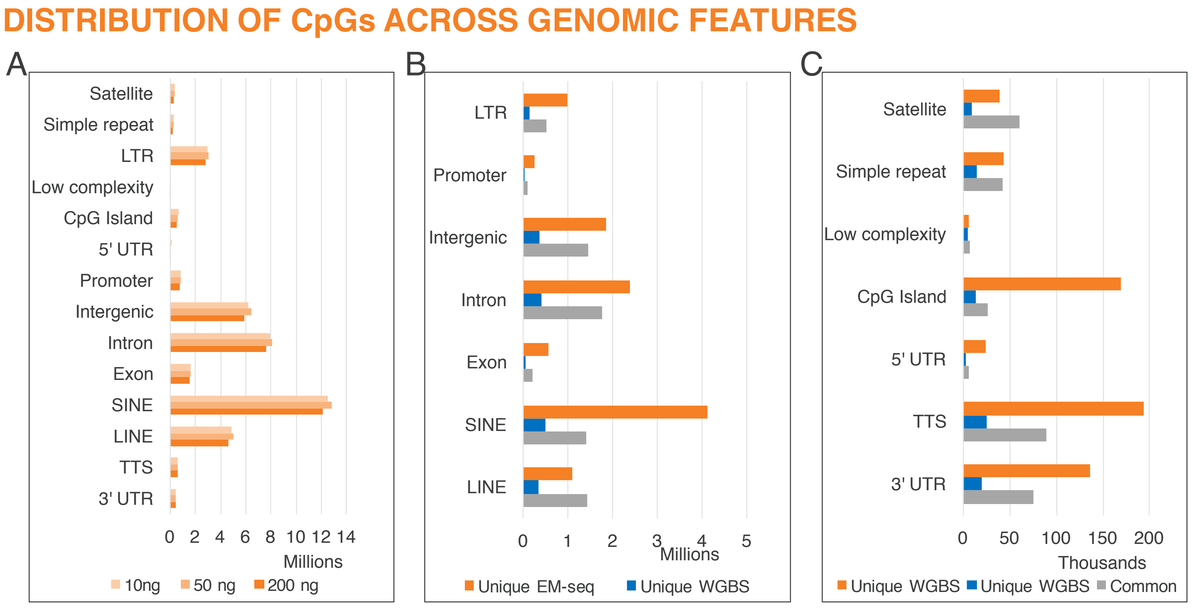


B



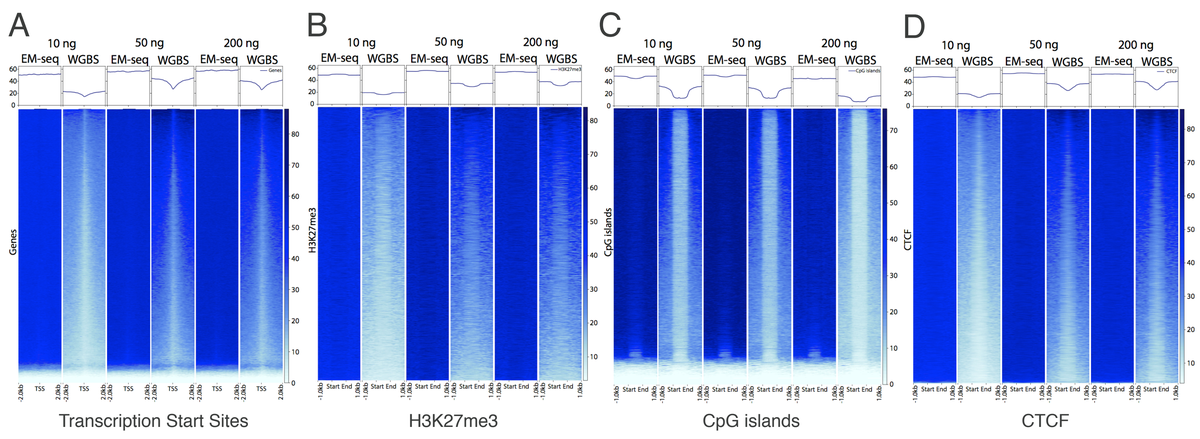
Top and bottom strand CpGs were counted independently, yielding a maximum of 56 M possible CpG sites. (A) EM-seq libraries identified more unique CpGs than bisulfite libraries using the same number of reads (324 M) for 10 ng, 50 ng and 200 ng inputs. EM-seq libraries show a higher percentage of CpGs covered at lower minimum coverage depths. (B) Dinucleotide coverage plot demonstrates that EM-seq library sequences are distributed more evenly over all dinucleotides. Data normalized to coverage observed in an unconverted Ultra II DNA library using the same input DNA as EM-seq and WGBS libraries.

DISTRIBUTION OF CpGs ACROSS GENOMIC FEATURES



Unique CpGs were compared between EM-seq and WGBS. (A) CpGs identified at 1x coverage are similar using 10ng, 50ng and 200 ng NA12878 DNA input amounts for EM-seq. (B) and (C) More CpGs can be identified in the EM-seq libraries compared to WGBS at 8x and higher coverage.

CpGs DETECTED WITHIN GENOMIC FEATURES



CpG coverage across genomic features are represented as heatmaps. (A) Transcription start sites (TSS), (B) H3K27me3, (C) CpG islands and (D) CTCF transcription factor binding sites are represented. Regions around the TSS are covered +/-2 kb and the H3K27, CpG islands and CTCF sites are represented +/- 1kb from the start and end sites. Dark blue indicates high coverage and light blue/white indicate little or no coverage. The heatmaps show that EM-seq has higher coverage at all DNA inputs across these genomic features.

CONCLUSIONS

Identification of CpGs using the EM-seq method is superior to whole genome bisulfite sequencing.

EM-seq compared to WGBS:

- Higher library yields with less PCR cycles
- Larger library insert sizes
- Lower percent duplication
- More even base coverage
- Detects more CpGs with fewer reads
- Less GC bias

We thank Laurie Mazzola, Danielle Fuchs, Aine Quimby from the NEB Sequencing Core for their assistance.