EM-seq Enables Accurate and Precise Methylome Analysis of Challenging DNA Samples

V. K. Chaithanya Ponnaluri, Louise Williams, Brittany Sexton, Lana Saleh, Katherine Marks, Mala Samaranayake, Bradley Langhorst, Zhiyi Sun, Thomas C. Evans, Jr., Romualdas Vaisvila, Eileen Dimalanta, Theodore B. Davis

New England Biolabs, Ipswich, MA 01938, USA

INTRODUCTION

The cytosine modifications 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC), are key epigenetic factors that play an important role in cellular processes. Misregulation of these cytosine forms results in disease states like cancer. Advances in sample preparation from biological matrices like blood (cell-free DNA (cfDNA)) as well as improvements in sequencing library preparation methods have enabled cancer biomarker identification based on methylation profiling in a minimally invasive manner.

Bisulfite sequencing is the standard method to detect methylation and has been employed for both targeted and whole genome methylation analysis. However, the chemical based bisulfite conversion of cytosines to uracils also results in DNA damage which subsequently results in shorter DNA insert sizes as well as introducing bias into the data. Therefore, analysis of DNA methylation from cfDNA is challenging as the DNA is typically of low quality and quantity. Robust biomarker detection relies primarily on the ability to profile methylation accurately. To overcome the drawbacks of bisulfite sequencing, we developed an enzyme based methylation detection technology, called NEBNext® Enzymatic Methyl-Seq (EM-Seq™), that minimizes damage to DNA requiring less PCR cycles, lower duplication rates and minimal GC bias resulting in more accurate quantification of methylation in the sample DNA.

EM-Seq was used to investigate cfDNA from a healthy human donor. EM-seq libraries demonstrated lower duplication rates, higher percentages of mapped reads and less GC bias compared to WGBS libraries. These libraries also identified a higher number of CpGs and demonstrated higher correlation between the EM-seq libraries compared to WGBS libraries. In conclusion, EM-Seq libraries have superior sequencing metrics resulting in robust methylation profiling for these types of challenging DNA samples.

METHODS

SAMPLE PREPARATION

cfDNA + sheared control DNA → Ultra II End Prep/ Ligation → TET2/ Oxidation Enhancer → APOBEC → PCR Amplification → Illumina Sequencing

- cfDNA was extracted from a healthy individual using single donor human plasma (anticoagulant: Na EDTA, Innovative™ Research). QIAamp Circulating Nucleic Acid Kit was used to extract cfDNA from 5 ml of plasma. No carrier RNA was used during the extraction.
- cfDNA was not sheared but 10ng and 25 ng of cfDNA was combined with two sheared control DNAs: unmethylated lambda (2 ng) and CpG methylated pUC19 (0.1 ng)
- 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 using 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

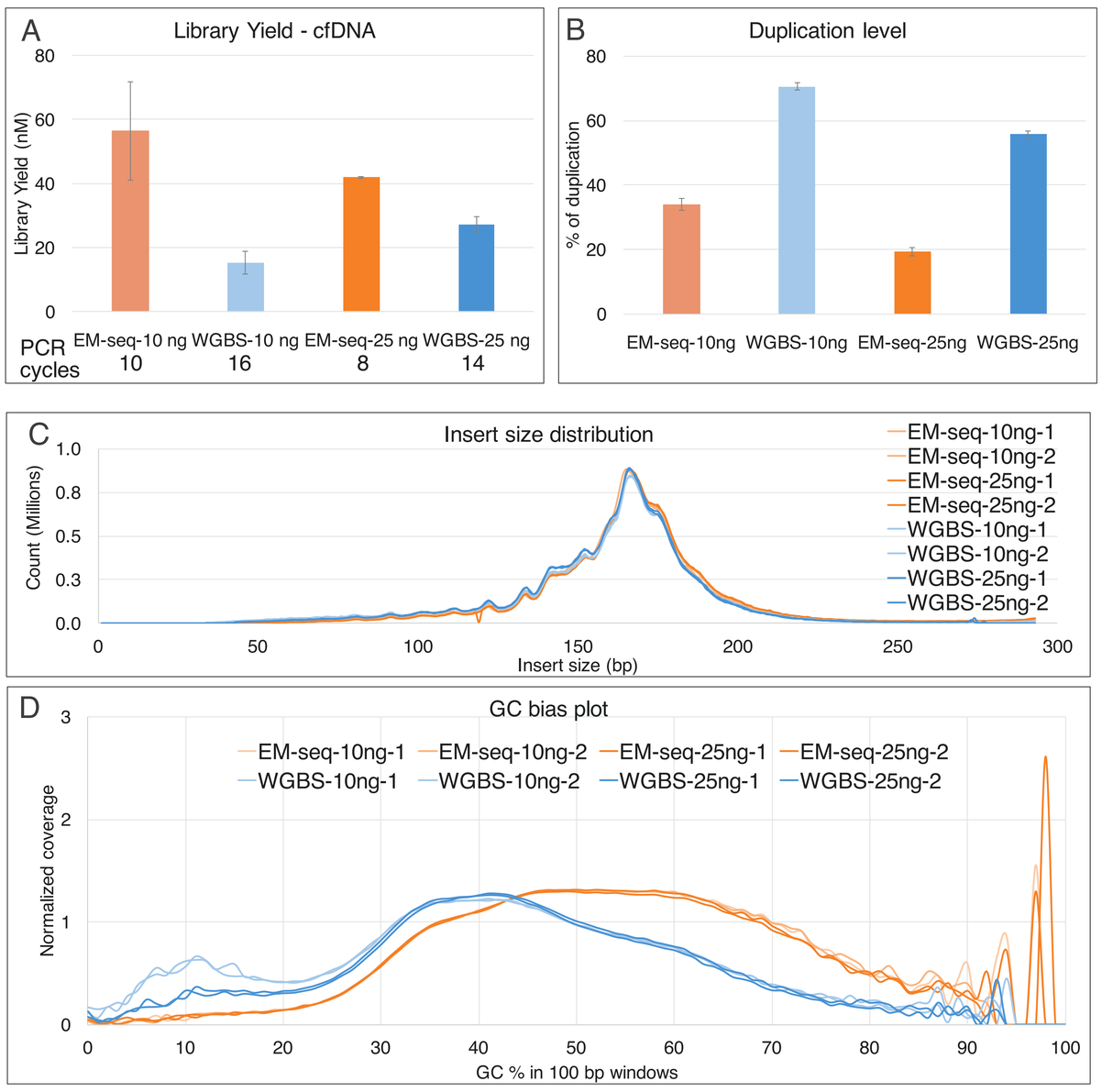
DATA ANALYSIS

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

- Reads were aligned to hg38 using BWA-Meth
- Methylation levels were extracted using MethylDackel
- Correlation analysis at 1x minimum coverage was performed 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



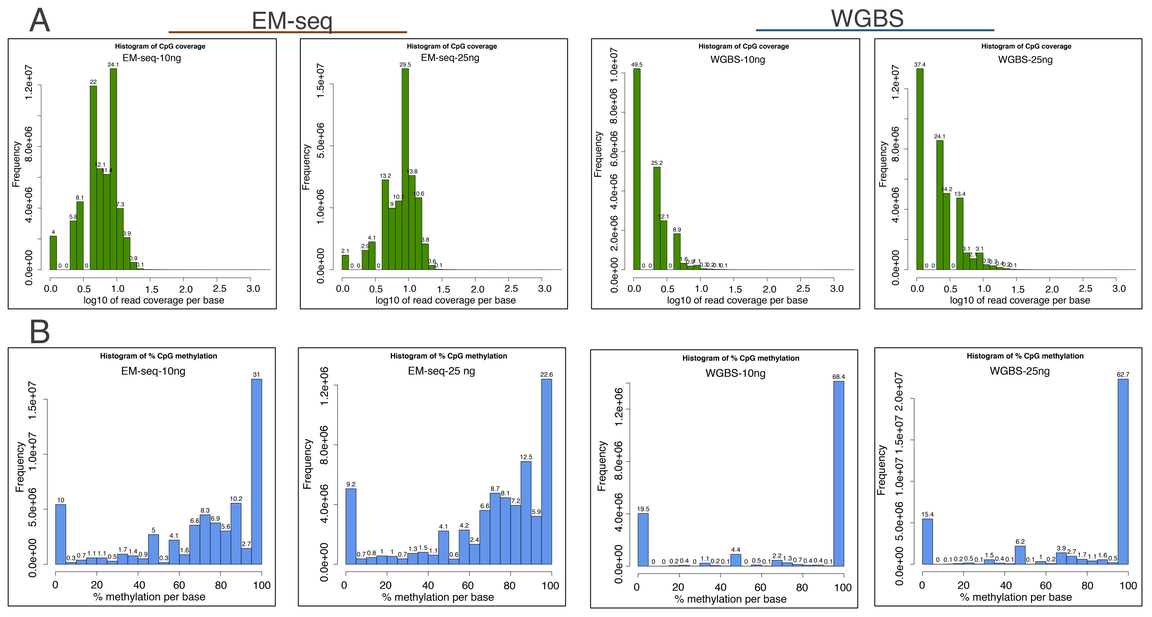
EM-seq and WGBS metrics from 10 ng and 25 ng cfDNA. Each library was sequenced using the Illumina NovaSeq 6000. (A) EM-seq libraries have higher yield using fewer PCR cycles compared to WGBS. (B) Library duplication percentages are lower for EM-seq. (C) Insert size distribution is similar between EM-seq and WGBS libraries 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.

RESULTS

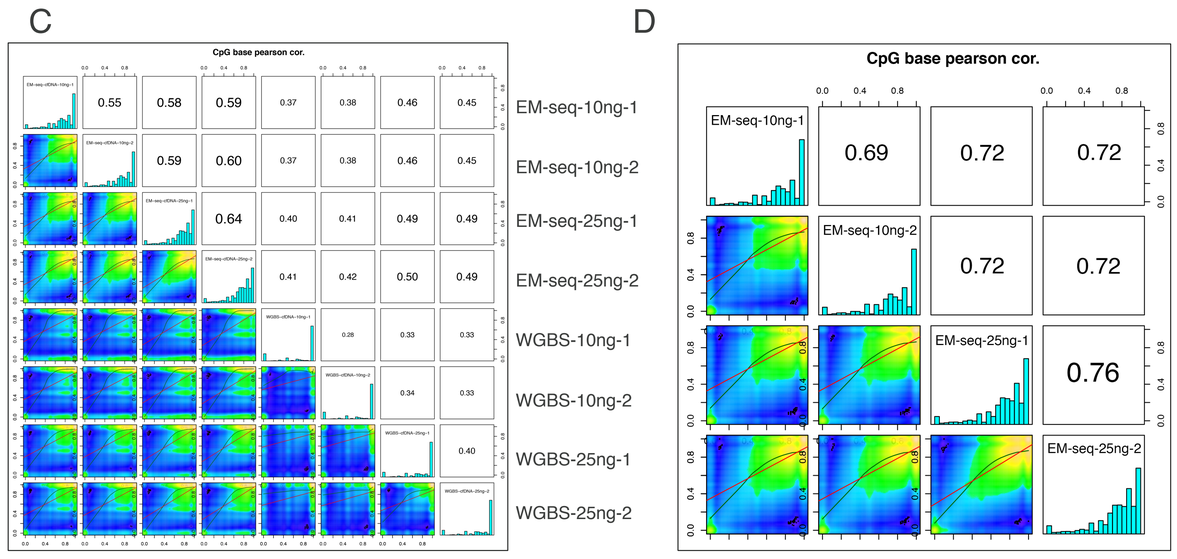
SIMILAR GLOBAL METHYLATION LEVELS BETWEEN EM-seq AND WGBS

	% methylation (10 ng)			% methylation (25 ng)		
cfDNA	CpG	CHG	CHH	CpG	CHG	CHH
EM-seq	76.48 ± 0.42	0.95 ± 0.07	0.90 ± 0.14	76.45 ± 0.07	0.75 ± 0.07	0.75 ± 0.07
WGBS	77.80 ± 0.14	0.35 ± 0.07	0.35 ± 0.07	78.7 ± 0.14	0.60 ± 0.14	0.60 ± 0.14

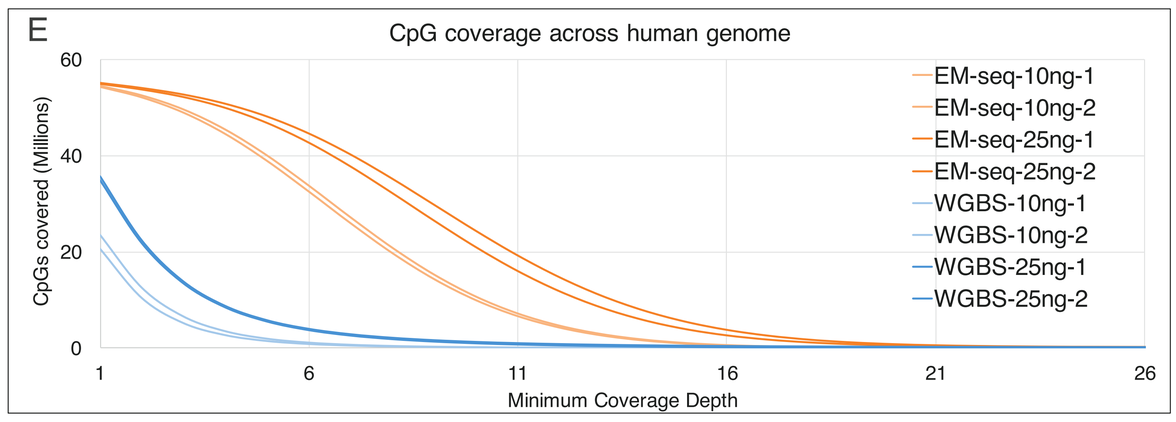
The percentage methylation for 10 ng and 25 ng cfDNA and unmethylated lambda DNA in CpG/CHG/CHH contexts. cfDNA: CpG methylation levels are similar for all libraries. Unmethylated Lambda: <1% methylated Cs in CpG, CHG and CHH were detected for all libraries.



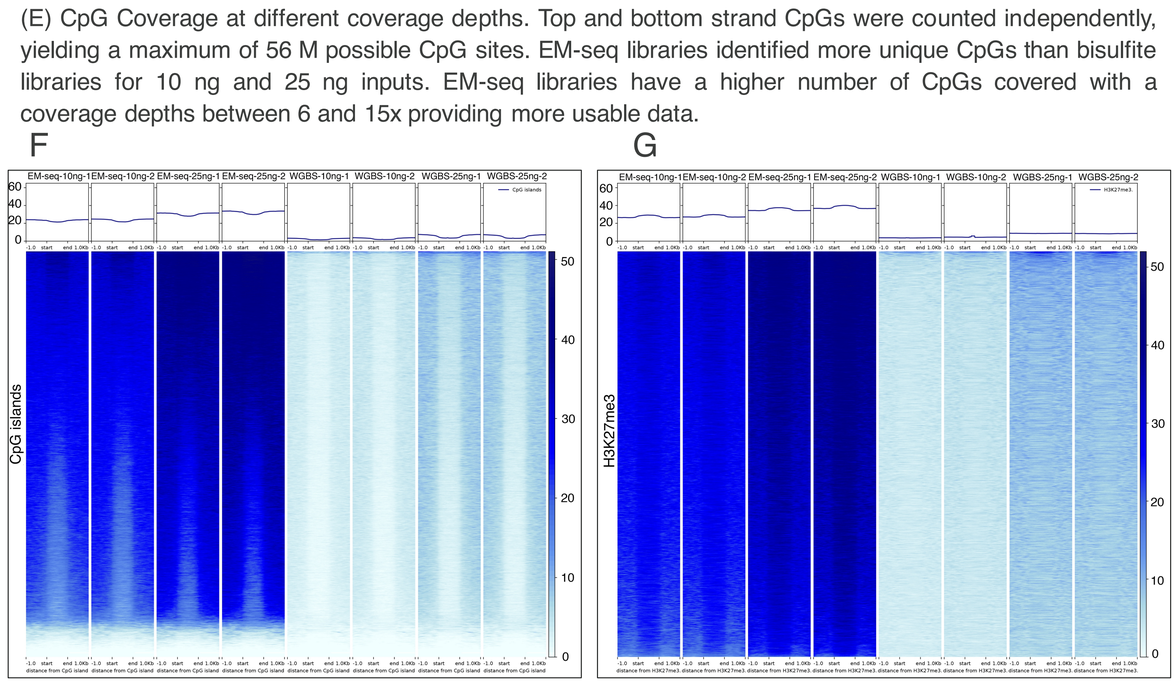
(A) CpG coverage of 10 ng and 25 ng EM-seq and WGBS libraries were plotted using MethylKit. The histogram is shifted right for EM-seq compared to WGBS indicating that more CpGs are detected at higher coverage for EM-seq libraries compared to WGBS for the 10 and 25 ng inputs. (B) The number of methylated cytosine for each possible percentage of methylated CpG bases are shown for EM-seq and WGBS.



(C) Pearson's correlations were plotted using MethylKit for 10 ng and 25 ng EM-seq and bisulfite libraries at 1x minimum coverage (8 million CpGs common to all libraries). (D) Pearson's correlation of 10 and 25 ng EM-seq libraries using 1x minimum coverage (53 million CpGs common to all libraries). Higher correlations are observed for EM-seq libraries demonstrating their robustness.



(E) CpG Coverage at different coverage depths. Top and bottom strand CpGs were counted independently, yielding a maximum of 56 M possible CpG sites. EM-seq libraries identified more unique CpGs than bisulfite libraries for 10 ng and 25 ng inputs. EM-seq libraries have a higher number of CpGs covered with a coverage depths between 6 and 15x providing more usable data.



CpG coverage across genomic features are represented as heatmaps. (F) CpG islands and (G) H3K27me3 are represented. Coverage of CpGs in a region around the CpG islands and H3K27me3 are represented in a +/- 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 robust compared to whole genome bisulfite sequencing.

EM-seq compared to WGBS for cfDNA:

- Higher library yields with less PCR cycles
- Lower percent duplication
- Detects more CpGs with fewer reads
- Less GC bias
- Higher correlation between replicates

Provides new method to evaluate low input cfDNA with higher concordance between the replicates for accurate methylation based biomarker detection

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