Enzymatic Methyl-seq enables accurate and robust methylation detection.

Vaishnavi Panchapakesa, V. K. Chaithanya Ponnaluri, Louise Williams, Matthew Campbell, Bradley Langhorst, Eileen Dimalanta, Theodore B. Davis
New England Biolabs, Ipswich, MA 01938, USA

INTRODUCTION

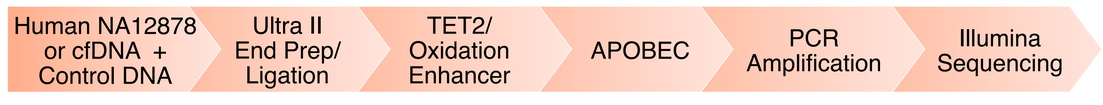
DNA methylation is one of the most important epigenetic regulatory mechanisms. The ability to accurately identify 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) gives us greater insight into potential gene regulatory mechanisms. Bisulfite sequencing (BS) is traditionally used to detect methylated cytosines, however, the chemical based conversion of cytosines to uracils leads to DNA damage which subsequently translates to shorter DNA insert sizes as well as biases in the data. To overcome these limitations, we developed NEBNext® Enzymatic Methyl-seq (EM-seq™), an enzymatic approach for detecting cytosine methylation.

EM-seq and BS Illumina libraries were prepared using 10 ng to 200 ng NA12878 DNA. EM-seq libraries have longer inserts and less GC bias compared to bisulfite converted libraries. Global methylation levels are similar between the two methods, indicating overall detection of methylated Cs is similar. However, CpG correlation plots demonstrated higher correlation coefficients indicating that EM-seq libraries are more consistent than BS across replicates and input amount. GC Bias and dinucleotide distribution showed that EM-seq has more even dinucleotide representation compared to the AT rich representation observed for BS. EM-seq libraries exhibit more even coverage allowing for a higher percentage of CpGs to be assessed and therefore leading to more consistent evaluation of methylation across key genomic features (TSS, CpG island, etc.).

There is increasing interest in the diagnostic applications of circulating cell-free DNA (cfDNA). Analysis of DNA methylation from cfDNA is challenging as the DNA is typically of low quantity and quality. EM-Seq and BS libraries were made using cfDNA. EM-seq libraries had longer inserts, lower duplication rates, higher percentages of mapped reads and less GC bias compared to BS libraries. These libraries also identified a higher number of CpGs resulting in enhanced coverage across genomic features, such as transcription start sites (TSS) and CpG islands. EM-seq is robust and reproducible, facilitating the generation of libraries with superior sequencing metrics for these challenging DNA samples.

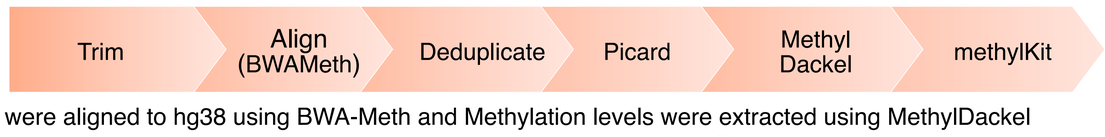
METHODS

SAMPLE PREPARATION



- NA12878 genomic DNA and cfDNA were used. cfDNA was extracted using single donor human plasma. QIAamp Circulating Nucleic Acid Kit was used to extract cfDNA from 5 ml of plasma. No carrier RNA was used during the extraction.
- 10 ng, 50 ng or 200 ng NA12878 was combined with control DNA and sheared to 300 bp. 10 ng or 25 ng of cfDNA was combined with control DNAs prior to library prep (not sheared). Control DNAs used were CpG methylated pUC19 and unmethylated lambda.
- 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 then sequenced using an Illumina NovaSeq 6000, 2x100 base paired reads
- Whole Genome Bisulfite Libraries (WGBS) were made using Zymo Research EZ DNA Methylation-Gold™ kit

DATA ANALYSIS



- Reads were aligned to hg38 using BWA-Meth and 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

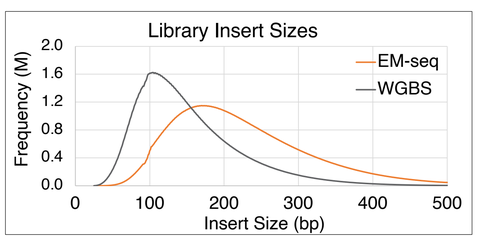
RESULTS

Human NA12878: Higher Quality Sequencing Data with EM-seq Libraries

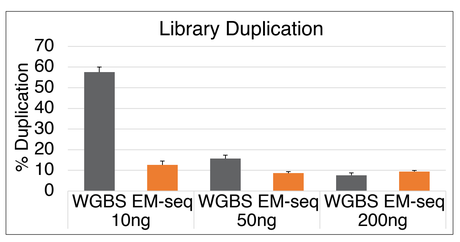
A

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

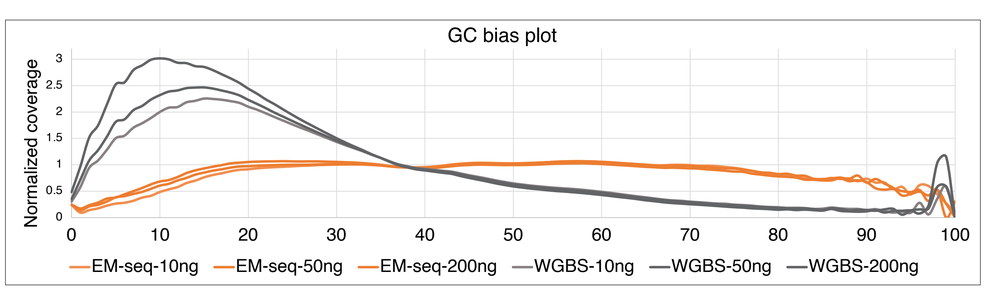
B



C

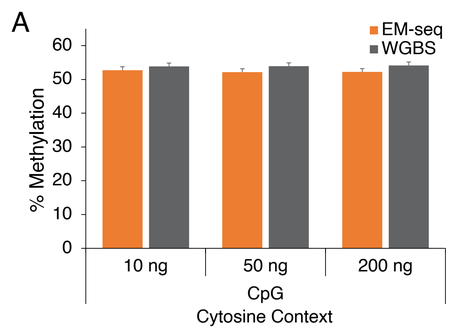


D

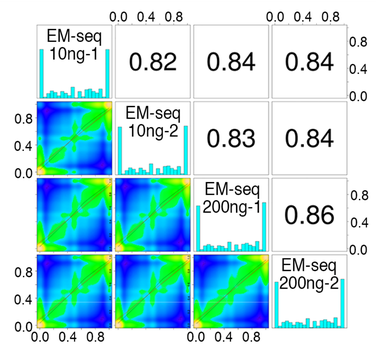


EM-seq and WGBS metrics from 10 ng, 50 ng and 200 ng NA12878 genomic DNA. Each library was sequenced using the Illumina NovaSeq 6000. 324 million, 2 x100 base reads were used for methylation analysis. (A) EM-seq libraries have higher yield but require fewer PCR cycles. (B) EM-seq library insert sizes are larger than bisulfite libraries. (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.

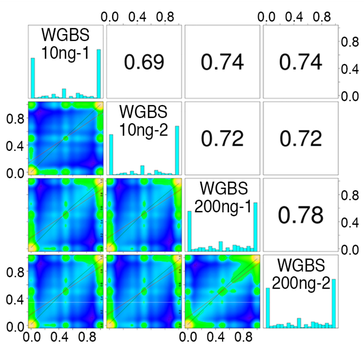
A



B

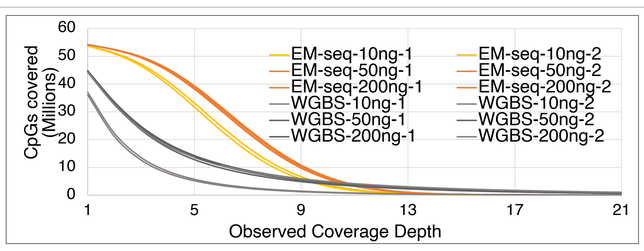


C

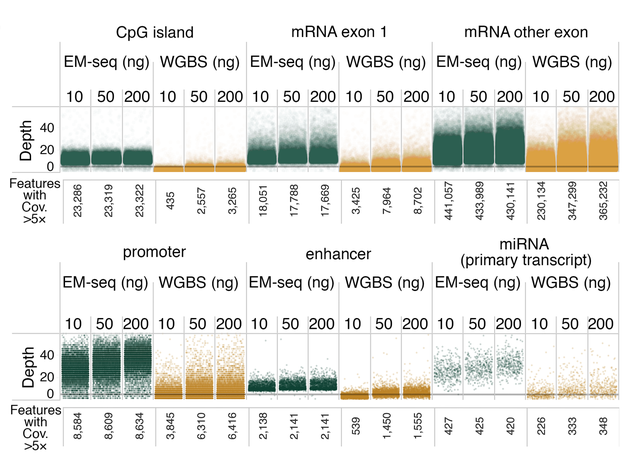


(A) NA12878 EM-seq and WGBS CpG methylation levels are similar. (B,C) Pearson's correlation plots for 10 ng and 200 ng NA12878 EM-seq (B) and WGBS libraries (C). Plots were generated using methylKit at a 1x minimum CpG coverage. EM-seq libraries have higher correlations.

A



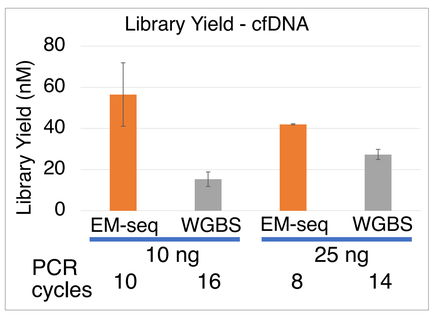
B



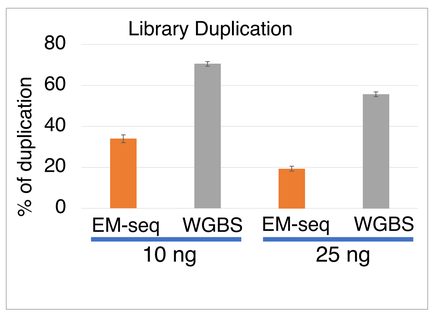
(A) The number of CpGs covered for EM-seq and WGBS libraries were calculated and graphed at minimum coverage depths of 1x through 21x. (B) Genomic features detected in EM-seq and WGBS libraries. Coverage at 5x or greater are indicated. EM-seq libraries have greater coverage for all features examined. Features from NCBI's RefSeq, the Eukaryotic Promoter, UCSC table browser and Dfam are shown. Coverages are represented with one point per region and the average coverage depth is plotted.

cfDNA: Higher Quality Sequencing Data with EM-seq Libraries

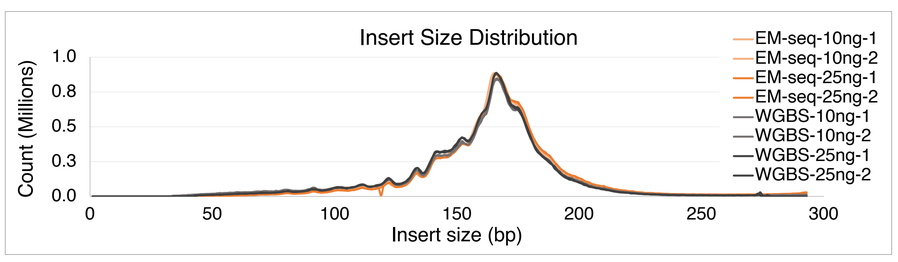
A



B



C



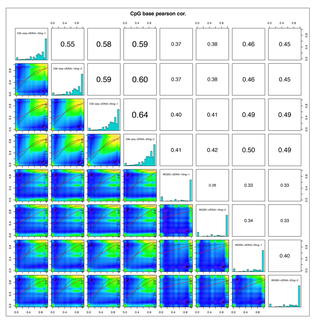
D

EM-seq and WGBS metrics from 10 ng and 25 ng cfDNA libraries. Libraries were 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 (D) EM-seq libraries show more even GC coverage distribution than bisulfite libraries. The bisulfite libraries are AT rich and have lower GC coverage.

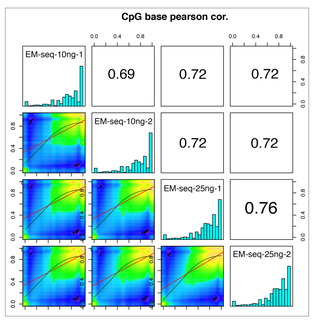
		EM-seq	WGBS
% methylation (10 ng)	CpG	76. ± 0.42	77.80 ± 0.14
	CHG	0.95 ± 0.07	0.35 ± 0.07
	CHH	0.90 ± 0.14	0.35 ± 0.07
% methylation (25 ng)	CpG	76.45 ± 0.07	78.7 ± 0.14
	CHG	0.75 ± 0.07	0.60 ± 0.14
	CHH	0.75 ± 0.07	0.60 ± 0.14

The percentage methylation for 10 ng and 25 ng cfDNA 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 (data not shown).

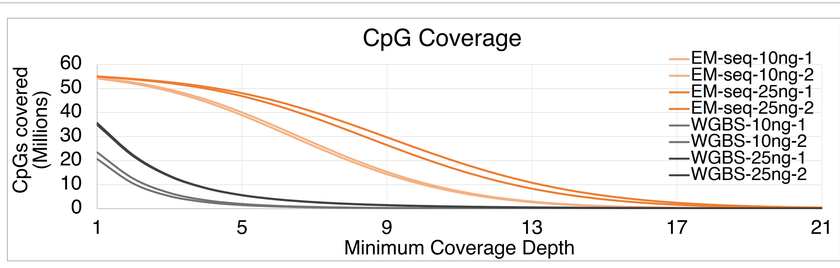
A



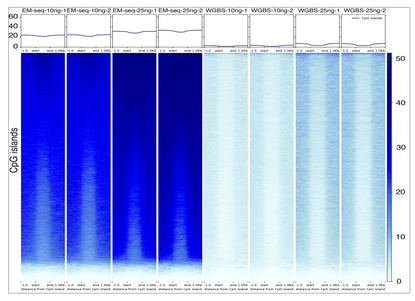
B



(A) Pearson's correlations were plotted using methylKit for 10 ng & 25 ng EM-seq and WGBS libraries at 1x minimum coverage (8 million CpGs common to all libraries). (B) Pearson's correlation of EM-seq libraries using 1x minimum coverage (53 million CpGs common to all libraries).



The number of CpGs at different coverage depths were plotted. EM-seq libraries identified more unique CpGs than bisulfite libraries for 10 ng and 25 ng inputs. EM-seq identifies more CpGs with coverage >5x providing more usable data.



EM-seq and WGBS cfDNA library CpG coverage across CpG islands. Heatmaps depict CpGs at +/- 1kb from the start and end sites of the CpG islands. 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

CONCLUSIONS

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

- Higher yields with less PCR cycles
- Larger library insert sizes
- More even base coverage
- Less GC bias
- Detects more CpG's with fewer reads

Identification of CpG's within cfDNA using the EM-seq method is robust compared to whole Genome bisulfite sequencing

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

We thank the NEB sequencing core (Laurie Mazzola, Danielle Fuchs, Kristen Augulewicz and Harold Bell) for their technical assistance.