An end-to-end enzymatic solution for methylomes from low input DNA

V K Chaithanya Ponnaluri, Brittany S. Sexton, Laura N. Blum, Vaishnavi Panchapakesa, Daniel J. Evanich, Matthew A. Campbell, Bradley W. Langhorst, Louise Williams | New England Biolabs, Inc.

Poster # 456

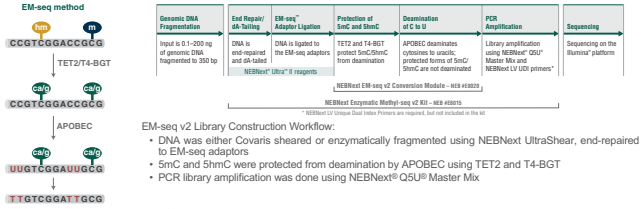
NEW ENGLAND Biolabs®

Introduction

The cytosine modifications, 5-methylcytosine and 5-hydroxymethylcytosine, are important regulatory marks, and their identification within genomes is essential in understanding gene regulation. Methylation signatures from clinical samples play a critical role in biomarker discovery and cancer diagnostics. However, the amount and quality of DNA available for diagnostics can be limiting. Historically, cytosine methylation has been detected using bisulfite sequencing. This method uses chemical conversion of cytosines into uracils and leads to DNA damage, which results in shorter DNA insert sizes as well as biases in the data. For lower input DNA, bisulfite conversion based methylomes are especially problematic due to the biases introduced and reduced genome coverage. In contrast to bisulfite sequencing, NEBNext® Enzymatic Methyl-seq (EM-seq™), leaves DNA intact and results in superior sequencing libraries with longer insert sizes, lower duplication rates and minimal GC bias. An enhanced EM-seq workflow can now accurately detect methylation in DNA samples using as little as 0.1 ng DNA. This EM-seq v2 workflow when combined with enzymatic fragmentation (NEBNext UltraShear®) enables further streamlining of library preparation by enabling end-to-end automation capability.

EM-seq v2 libraries were prepared using both acoustic and enzymatic (NEBNext UltraShear) fragmentation of 200 ng to 0.1 ng NA12878. Additionally, we also performed target capture using Twist Human Methylome Panel. All libraries generated robust yields with even GC coverage, consistent insert size profile, low duplication rates and high mapping rates. Furthermore, for the NA12878 libraries, ~56 million CpGs were identified for 200 ng to 1 ng inputs and ~44 million CpGs for 0.1 ng inputs using both fragmentation methods. EM-seq v2 combined with enzymatic fragmentation provides a robust, cost effective, end-to-end library preparation solution. Combining EM-seq v2 with target capture approach provides an opportunity to evaluate a subset of CpGs with higher coverage using lower sequencing burden. Using the data from the target capture, EM-seq libraries enables variant calling using fewer reads compared to whole genome sequencing. EM-seq v2 libraries have superior sequencing metrics that establish it as the gold standard for accurate methylation profiling, particularly for lower DNA input and challenging samples.

Methods



EM-seq v2 Library Construction Workflow:

- DNA was either Covaris sheared or enzymatically fragmented using NEBNext UltraShear, end-repaired and ligated to EM-seq adaptors
- 5mC and 5hmC were protected from deamination by APOBEC using TET2 and T4-BGT
- PCR library amplification was done using NEBNext® Q5UI® Master Mix

EM-seq v2 NA12878 Library Construction

- 200 ng, 10 ng, 1 ng & 0.1 ng of NA12878 genomic DNA were spiked with control DNAs (Unmethylated Lambda and pUC19 (all CpGs are 5mC modified)) and were used in library construction
- Samples were either fragmented to an average size of 350 bp using Covaris ME220 or using NEBNext UltraShear by incubating at 37˚C for 30 mins followed by 15 mins at 65˚C

EM-seq v2 NA12878 Library Target Capture

- 200 ng of NA12878 genomic DNA were spiked with control DNAs (Unmethylated Lambda and pUC19 (all CpGs are 5mC modified)) and were used in library construction
- Samples were fragmented to an average size of 350 bp using Covaris ME220
- Twist Human Methylome Panel was used to perform target capture following manufacture's recommendations

Sequencing and Data Analysis

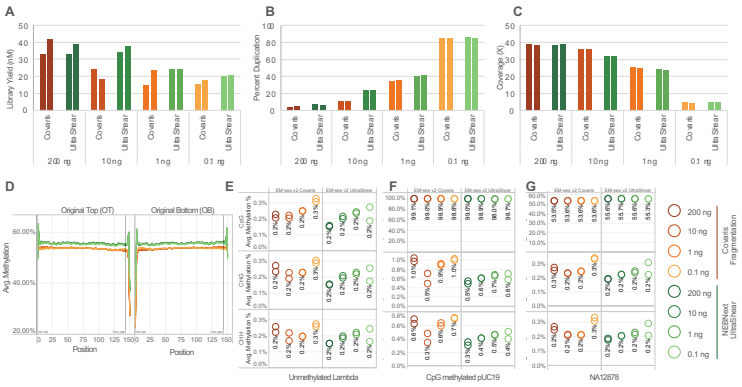
Trim → Align (bwa-meth) → Deduplicate → Picard → Methyl Dackel → methylKit

All libraries were sequenced on an Illumina NovaSeq 6000 using 2x 150 base reads and 915 million total reads were analyzed per library
For target capture libraries 128 million total reads were analyzed per library
EM-seq v2 data were processed using the following pipeline:

- Reads were adaptor trimmed (fastp) then aligned to a human T2T composite genome (including controls) using bwa-meth
- Methylation information was extracted from the alignments using MethylDackel and levels were evaluated independently for each chromosome
- methylKit data was used for Pearson correlation at 1x minimum coverage for whole genome libraries and 5x minimum for target capture comparison
- Picard was used to mark duplicates as well as to calculate library insert size, GC bias and HS metrics
- Revelio was used to strand-specifically mask base qualities to remove methylation information while retaining genotype information prior to variant calling
- Germline variants were called with Strelka2 and filtered for QUAL > 15
- Recall and precision of variant calling was assessed in targeted regions using Hap.py
- Variant calls for targeted EM-seq v2 libraries were compared with 140,958 SNPs called in a 100 ng Covaris sheared NEBNext Ultra II library (910M reads). One replicate of NEBNext Ultra II library was used as the truth set and concordance is shown to the other replicate. Recall is TP / (TP + FN). Precision is TP / (TP + FP).

Results

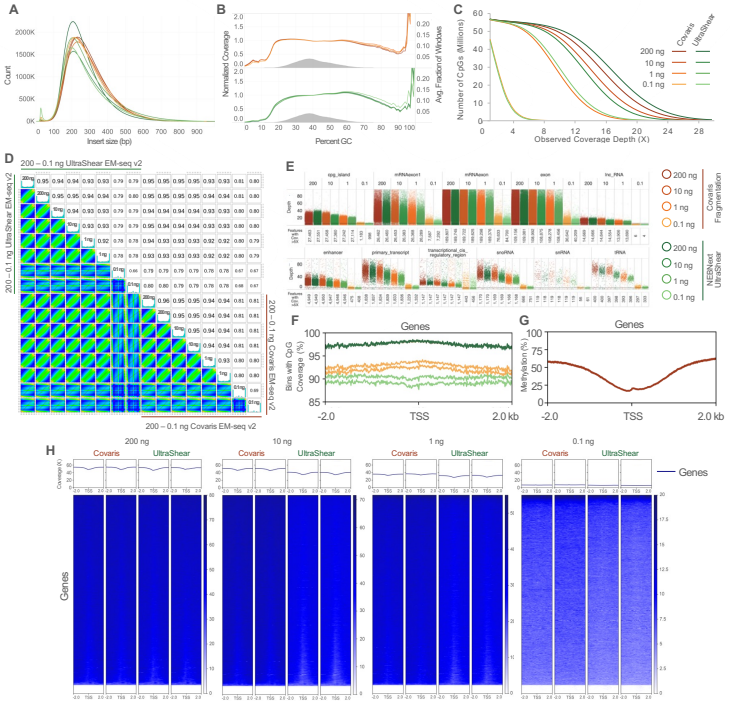
EM-seq v2 libraries using NA12878: Basic Metrics



EM-seq v2 libraries were made using 200 ng - 0.1 ng of NA12878 DNA fragmented using either Covaris or NEBNext UltraShear. (A) Library yields for input series. (B, C) Percent duplication and effective genome coverage are shown. (D) M-bias plot showing the level of methylation observed across the read in the CpG context. (E, F, G) Percent methylation detected in the CpG, CHG and CHH contexts for unmethylated lambda control (< 0.4% indicating efficient deamination), CpG methylated pUC19 control (~99.0% indicating an efficient protection reaction), and Human (NA12878) genome (consistent CpG methylation across inputs for both fragmentation methods) shown in replicates.

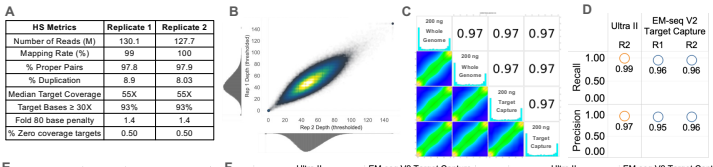
Results

EM-seq v2 libraries using NA12878: Secondary Analysis



(A) Library insert sizes are consistent across DNA inputs and fragmentation methods. (B) Uniform GC coverage over a range of inputs and fragmentation methods. (C) Cumulative coverage plot showing number of CpGs covered at different coverage depth. 200 ng - 1 ng libraries cover ~56 M CpGs at 1X coverage and 0.1 ng libraires cover ~44 M CpGs using both fragmentation methods. (D) Pearson correlation for 200 ng to 0.1 ng NA12878 DNA inputs using both fragmentation methods. Correlations were made using 24 million CpGs covered by all libraries. (E) Coverage of genomic features. The number of features with coverage greater than 5X is indicated below each plot. Coverage of genomic feature types are represented with one point per region with the vertical position representing the average coverage of the feature. Points are staggered horizontally to avoid excess overlapping. Feature annotations are from NCBI's RefSeq browser. CpG islands were defined based on the UCSC genome browser. Strong correlations and uniform coverage of genomic features across input range using both fragmentation methods demonstrating robustness of the EM-seq v2 workflow. One replicate shown for each input and fragmentation condition in A, B, C, and E. (F) Plot of the percentage of CpG containing bins (bin size: 10 bp) with coverage. The percentage of bins covered ± 2 kb around transcription start sites (TSS) showed low variability. (G) Plot of the level of methylation observed for the same bins. (H) Heatmaps showing uniform coverage ± 2 kb windows around TSS across input range for both fragmentation methods.

EM-seq v2 libraries using NA12878 coupled with Target Capture

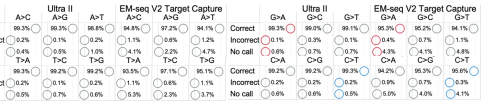


HS Metrics	Replicate 1	Replicate 2
Number of Reads (M)	130.1	127.7
Mapping Rate (%)	99	100
% Proper Pairs	97.8	97.9
% Duplication	8.9	8.03
Median Target Coverage	55X	55X
Target Bases ≥ 30X	93%	93%
Fold 80 base penalty	1.4	1.4
% Zero coverage targets	0.50	0.50

E

	Ultra II	EM-seq V2 Target Capture	
	R2	R1	R2
False Neg.	1,084	5,975	6,021
False Pos.	4,299	6,399	6,339
Total Truth	140,958	140,958	140,958

F



EM-seq v2 libraries were made using 200 ng of NA12878 DNA fragmented using Covaris. Target capture was performed using Twist Human Methylome Panel. (A) Table showing consistent HS metrics for the target captured libraries (replicates). (B) Density plot showing the concordance between the observed coverage between replicates. (C) Pearson correlation using a 5X coverage threshold demonstrates strong agreement for the detected methylation across 11.9 million CpGs between 200 ng whole genome EM-seq v2 libraries and target captured libraries. (D) Plot showing the variant calling Recall and Precision metrics for EM-seq v2 target captured libraries compared to unconverted NEBNext Ultra II DNA libraries prepared using NA12878 DNA. (E) Quantification of False Negatives and False Positives for the variants observed in the targeted region. (F) Classification of the calls across all possible mutation combinations.

Conclusions

EM-seq v2 provides:

- Streamlined, robust and accurate enzymatic conversion method to detect 5mC and 5hmC with DNA inputs ranging from 200 ng to 0.1 ng
- Consistent CpG coverage, expected insert sizes and minimal GC bias
- Consistent performance with both Enzymatic fragmentation using NEBNext UltraShear and acoustic shearing (Covaris)
- End-to-end automation compatibility with incorporation of NEBNext UltraShear
- Compatibility with target capture workflows
- Ability to perform simultaneous variant calling and methylation assessment