NEW ENGLAND BioLabs® Inc.
be INSPIRED
drive DISCOVERY
stay GENUINE

FEATURE ARTICLE

Enzymatic Methyl-seq: The Next Generation of Methylome Analysis

by Louise Williams, Ph.D., Yanxia Bei, Ph.D., Heidi E. Church, Nan Dai, Ph.D., Eileen T. Dimalanta, Ph.D., Laurence M. Ettwiller, Ph.D., Thomas C. Evans, Jr., Ph.D., Bradley W. Langhorst, Ph.D., Janine G. Borgaro, Ph.D., Shengxi Guan, Ph.D., Katherine Marks, Julie F. Menin, Nicole M. Nichols, Ph.D., V. K. Chaithanya Ponnaluri, Ph.D., Lana Saleh, Ph.D., Mala Samaranayake, Ph.D., Brittany S. Sexton, Ph.D, Zhiyi Sun, Ph.D., Esta Tamanaha, Ph.D., Romualdas Vaisvila, Ph.D., Erbay Yigit, Ph.D. and Theodore B. Davis, New England Biolabs, Inc.

The identification of cytosine modifications within genomes, especially 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC), is important as they are known to have an impact on gene expression. Generally, low levels of methylation near transcription start sites are associated with higher transcription levels, while genes with regulatory regions containing high levels of cytosine modification are expressed at lower levels. The ability to analyze a complete methylome is important for studying diseases, including those associated with cancer, metabolic disorders and autoimmune diseases. Unfortunately, the current technologies for investigating 5mC and 5hmC are sub-optimal and do not permit a thorough evaluation of methylomes.

BISULFITE SEQUENCING

To date, the gold standard in methylome mapping has been bisulfite sequencing. In this method, DNA is chemically treated with sodium bisulfite, which results in the conversion of unmethylated cytosines to uracils, and the resulting uracils are ultimately sequenced as thymines (Figure 1). In contrast, the modified cytosines, 5mC and 5hmC, are resistant to bisulfite conversion, and are sequenced as cytosines (1). While the preparation of bisulfite libraries is relatively straightforward, the libraries have uneven genome coverage and therefore suffer from incomplete representation of cytosine methylation across genomes. This uneven coverage is the result of DNA damage and fragmentation, which is caused by the extreme temperatures and pH during bisulfite conversion. Sequenced bisulfite libraries typically have skewed GC bias plots, with a general under-representation of G- and C-containing dinucleotides and over-representation of AA-, AT- and TA-containing dinucleotides, as compared to a non-converted genome (2). Therefore, the damaged libraries do not adequately cover the genome, and can include many gaps with little or no coverage. Increasing the sequencing depth of these libraries can recover some missing information, but at steep sequencing costs. These bisulfite library limitations have driven the development of new approaches for studying methylomes.

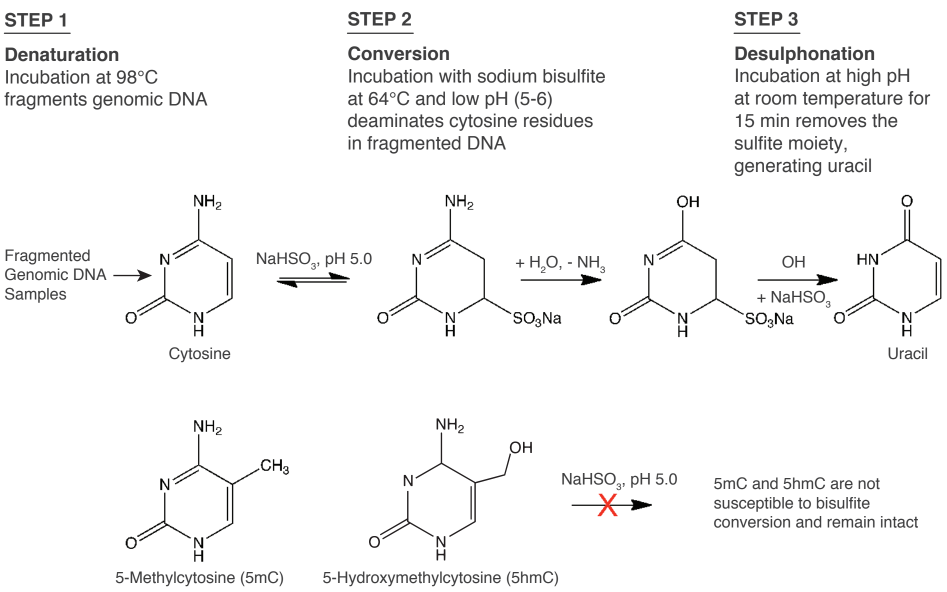
ALTERNATIVE METHODS FOR DETECTING 5mC AND 5hmC

Additional approaches for investigating methylomes are available that either combine bisulfite conversion with another chemical modification or an enzymatic modification step, or that eliminate bisulfite conversion completely (Table 1).

5hmC can be detected using TET-assisted bisulfite sequencing (TAB-seq). Fragmented DNA is enzymatically modified using sequential T4 Phage ß-glucosyltransferase (T4-BGT) and then Ten-eleven translocation (TET) dioxygenase treatments before the addition of sodium

FIGURE 1:
Bisulfite conversion overview

Sodium bisulfite treatment of DNA converts cytosine to 5,6-dihydrocytosine-6-sulphonate, which is converted to 5,6-dihydrouracil-6-sulphonate, and then desulphonated to uracil. In contrast 5mC and 5hmC are not susceptible to bisulfite treatment and remain intact.


STEP 1
Denaturation
Incubation at 98°C fragments genomic DNA

STEP 2
Conversion
Incubation with sodium bisulfite at 64°C and low pH (5-6) deaminates cytosine residues in fragmented DNA

STEP 3
Desulphonation
Incubation at high pH at room temperature for 15 min removes the sulfite moiety, generating uracil

5mC and 5hmC are not susceptible to bisulfite conversion and remain intact

TABLE 1:
Summary of alternative methods of methylome analysis

SEQUENCING METHOD	CYTOSINE MODIFICATION	METHOD OF ANALYSIS	WEAKNESS
TET-assisted bisulfite sequencing (TAB-seq) (3)	5hmC	Enzymatic treatment with T4-BGT then TET followed by bisulfite treatment	DNA damage and sequencing bias
Oxidative bisulfite sequencing (oxBS) (4)	5mC	Treatment with an oxidation reagent, followed by bisulfite treatment	DNA damage and sequencing bias
APOBEC-coupled epigenetic sequencing (ACE-seq) (5)	5hmC	Enzymatic treatment with T4-BGT and APOBEC3A	APOBEC3A not commercially available
TET-assisted 5-methylcytosine sequencing (TAmC-seq) (6)	5mC	Enzymatic treatment, followed by enrichment for 5mC regions	Enriches for 5mC dense regions. Does not currently cover entire genome.

continued on page 2...

bisulfite (3). T4-BGT glucosylates 5hmC to form beta-glucosyl-5-hydroxymethylcytosine (5ghmC) and TET is then used to oxidize 5mC to 5caC (Figure 2). Only 5ghmC is protected from subsequent demination by sodium bisulfite and this enables 5hmC to be distinguished from 5mC by sequencing.

Oxidative bisulfite sequencing (oxBS) provides another method to distinguish between 5mC and 5hmC (4). The oxidation reagent potassium perruthenate converts 5hmC to 5-formylC (5fC) and subsequent sodium bisulfite treatment deaminates 5fC to uracil. 5mC remains unchanged and can therefore be identified using this method.

APOBEC-coupled epigenetic sequencing (ACE-seq) excludes bisulfite conversion altogether and relies on enzymatic conversion to detect 5hmC (5). With this method, T4-BGT glucosylates 5hmC to 5ghmC and protects it from deamination by Apolipoprotein B mRNA editing enzyme subunit 3A (APOBEC3A). Cytosine and 5mC are deaminated by APOBEC3A and sequenced as thymine.

Lastly, TET-assisted 5-methylcytosine sequencing (TAmC-seq) enrichs for 5mC loci and utilizes two sequential enzymatic reactions followed by an affinity pull-down (6). Fragmented DNA is treated with T4-BGT which protects 5hmC by glucosylation. The enzyme mTET1 is then used to oxidize 5mC to 5hmC, and T4-BGT labels the newly formed 5hmC using a modified glucose moiety (6-N3-glucose). Click chemistry is used to introduce a biotin tag which enables enrichment of 5mC-containing DNA fragments for detection and genome wide profiling.

Libraries made from methods that combine enzymatic and sodium bisulfite identification of cytosine modifications all experience DNA damage and the inherent biases of bisulfite treatment. Furthermore, the described enzymatic methods have additional drawbacks. TAmC-seq is focused on loci and does not discriminate between methylated and unmethylated cytosines in the enriched DNA fragments. ACE-seq probes only 5hmC and requires APOBEC3A for deamination, which is not yet commercially available, making it more difficult to standardize library construction between labs.

ENZYMATIC METHYL-SEQ – A NEW APPROACH

The enzymatic methyl-seq workflow developed at NEB provides a much-needed alternative to bisulfite sequencing. This method relies on the ability of APOBEC to deaminate cytosines to uracils. Unfortunately, APOBEC also deaminates 5mC and 5hmC, making it impossible to differentiate between cytosine and its modified forms (7,8). In order to detect 5mC and 5hmC, this method also utilizes TET2 and an Oxidation Enhancer, which enzymatically modify 5mC and 5hmC to forms that are not substrates for APOBEC. The TET2 enzyme converts 5mC to 5caC (Figure 2) and the Oxidation Enhancer converts 5hmC to 5ghmC (9,10,11). Ultimately, cytosines are sequenced as thymines and 5mC and 5hmC are sequenced as cytosines, thereby protecting the integrity of the original 5mC and 5hmC sequence information.

The NEBNext Enzymatic Methyl-seq Kit (EM-seq™) combines NEBNext® Ultra™ II reagents with these two enzymatic steps to construct Illumina® libraries that accurately represent 5mC and 5hmC within the genome. Converted libraries are amplified using NEBNext Q5U DNA polymerase (Figure 3). EM-seq libraries result in a more accurate representation of the methylome, with minimal DNA fragmentation or biases when compared to whole genome bisulfite sequencing (WGBS). The combination of the Ultra II reagents for library prep and the EM-seq conversion allows for lower input amounts compared to most WGBS workflows, with a range of inputs from 10 – 200 ng.

FIGURE 3:
NEBNext EM-seq Kit Workflow

EM-seq utilizes two enzymatic steps to differentiate between modified and unmodified cytosines.

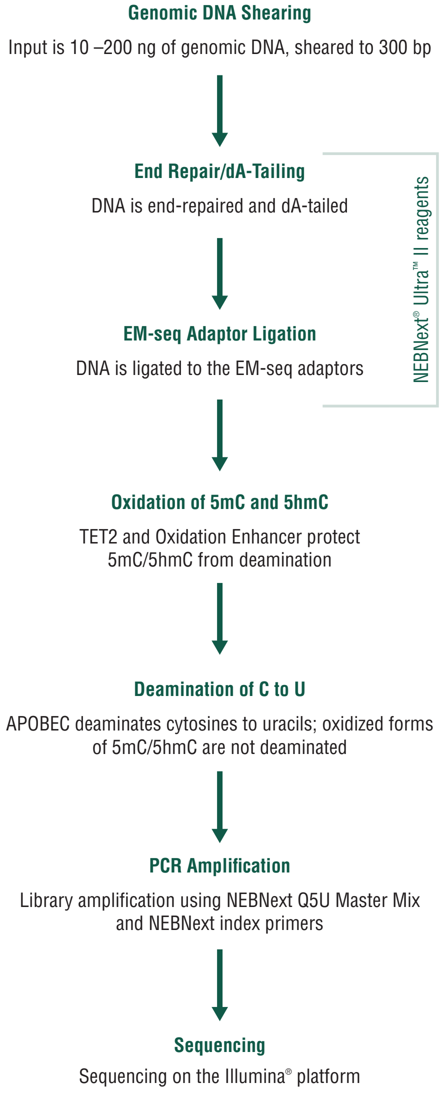
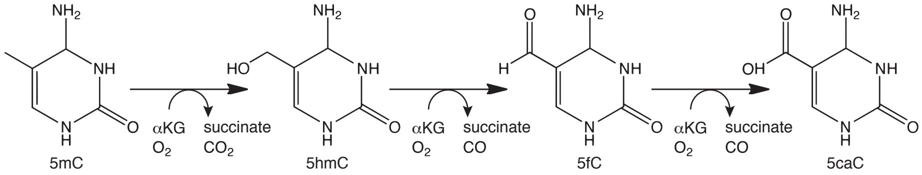

Genomic DNA Shearing
Input is 10 –200 ng of genomic DNA, sheared to 300 bp
End Repair/dA-Tailing
DNA is end-repaired and dA-tailed
EM-seq Adaptor Ligation
DNA is ligated to the EM-seq adaptors
NEBNext® Ultra™ II reagents
Oxidation of 5mC and 5hmC
TET2 and Oxidation Enhancer protect 5mC/5hmC from deamination
Deamination of C to U
APOBEC deaminates cytosines to uracils; oxidized forms of 5mC/5hmC are not deaminated
PCR Amplification
Library amplification using NEBNext Q5U Master Mix and NEBNext index primers
Sequencing
Sequencing on the Illumina® platform

FIGURE 2:
Enzymatic modification of cytosine

TET enzymes oxidize 5mC to 5hmC then 5fC and finally 5caC.



continued on page 3...

EM-SEQ PERFORMANCE

Intact DNA

Several pieces of data suggest that the process of generating EM-seq libraries does not damage DNA in the same way as bisulfite sequencing. EM-seq libraries give higher PCR yields despite using fewer PCR cycles for all DNA input amounts (see page 6), indicating that less DNA is lost during enzymatic treatment and library preparation, as compared to WGBS. Reduced PCR cycles, in turn, translates into more complex libraries and fewer PCR duplicates during sequencing (data not shown). EM-seq libraries also have larger insert sizes than WGBS (Figure 4), which further supports the fact that DNA remains intact.

EM-seq Libraries Have Reduced Bias

The preservation of DNA integrity is also demonstrated by the GC bias graphs (Figure 5), and the dinucleotide coverage distribution graph (Figure 6). Both of these figures indicate that reduced bias is associated with the EM-seq libraries. The EM-seq libraries have a flat GC bias distribution (Figure 5) with even coverage at both GC and AT rich regions, and do not display a preference for any dinucleotide combination (Figure 6). This is in stark contrast to WGBS, which shows a skewed GC bias profile along with the previously mentioned dinucleotide biases. Reduced library bias improves the mapping and therefore coverage of CpGs.

CpG Detection

Human DNA is methylated almost exclusively in CpG contexts. EM-seq global CpG methylation levels for human NA12878 DNA are consistent with WGBS libraries (Figure 7A), indicating that EM-seq libraries accurately detect methylation. The more striking difference between EM-seq and WGBS libraries becomes apparent when the focus is shifted to CpG coverage. EM-seq libraries detect more CpGs to a higher depth of coverage than WGBS libraries (Figure 7B). The ability to detect more CpGs at a greater depth also increases confidence in the data and leads to more accurately defining methylation within a region of interest. This in turn aids in detecting methylation changes in diseased states such as cancer. In addition, increased CpG coverage has an economic impact – with more CpGs detected using the same number of reads compared to WGBS, EM-seq represents significant cost-savings.

Potential Applications

In addition to making Illumina libraries, there are other potential applications for the EM-seq technology. Many of these applications already exist, but can now be improved upon because of the intact nature of enzymatically-converted DNA and the accuracy of CpG detection. Lower input DNA is also a driving factor for some of these applications. Converted DNA can be detected on arrays, and can be used for target enrichment, reduced representation-type libraries or amplicon detection. Different types of DNA inputs, such as low input cell free DNA (cfDNA) or damaged FFPE DNA, can also be used.

FIGURE 4:
NEBNext Enzymatic Methyl-seq (EM-seq) libraries have larger inserts

EM-seq library insert sizes are larger than whole genome bisulfite sequencing (WGBS) libraries. Library insert sizes were determined using Picard 2.18.14. The larger insert size indicates that EM-seq does not damage DNA as bisulfite treatment does.

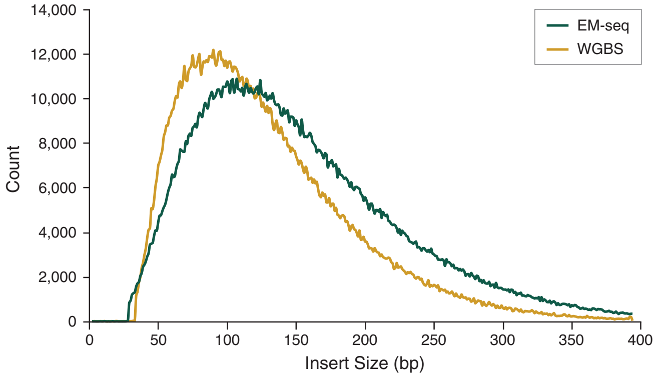
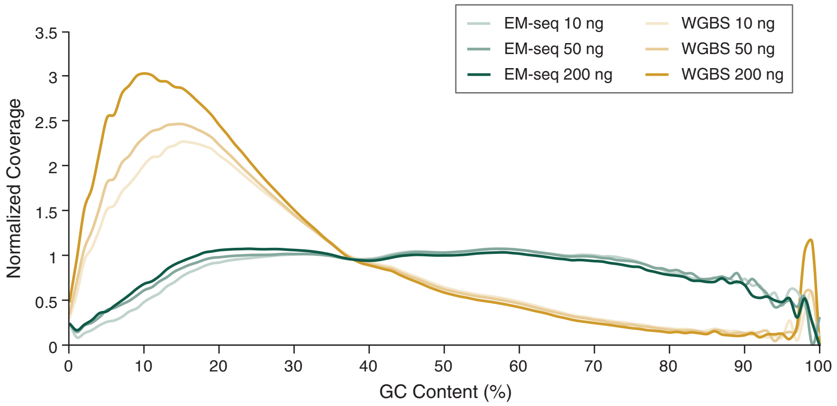


FIGURE 5:
EM-seq has superior uniformity of GC coverage

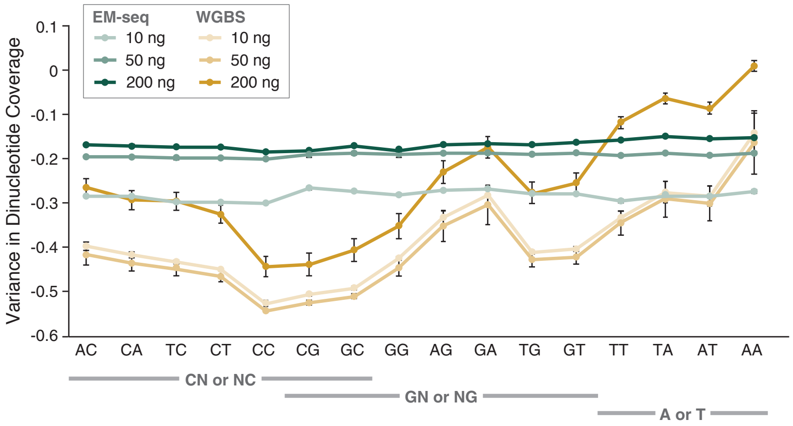
GC coverage was analyzed using Picard 2.18.14 and the distribution of normalized coverage across different GC contents of the genome (0-100%) was plotted. EM-seq libraries have significantly more uniform GC coverage, and lack the AT over-representation and GC under-representation typical of WGBS libraries.



continued on page 4...

FIGURE 6:
Dinucleotide coverage distribution

Dinucleotide coverage distribution for EM-seq and WGBS libraries showing the variance in coverage for dinucleotides in the reads when compared to unconverted Ultra II library dinucleotide distribution. EM-seq libraries show even coverage across all dinucleotide combinations compared to WGBS. C-containing dinucleotides are underrepresented in WGBS libraries and A/T containing dinucleotides are overrepresented.



CONCLUSION

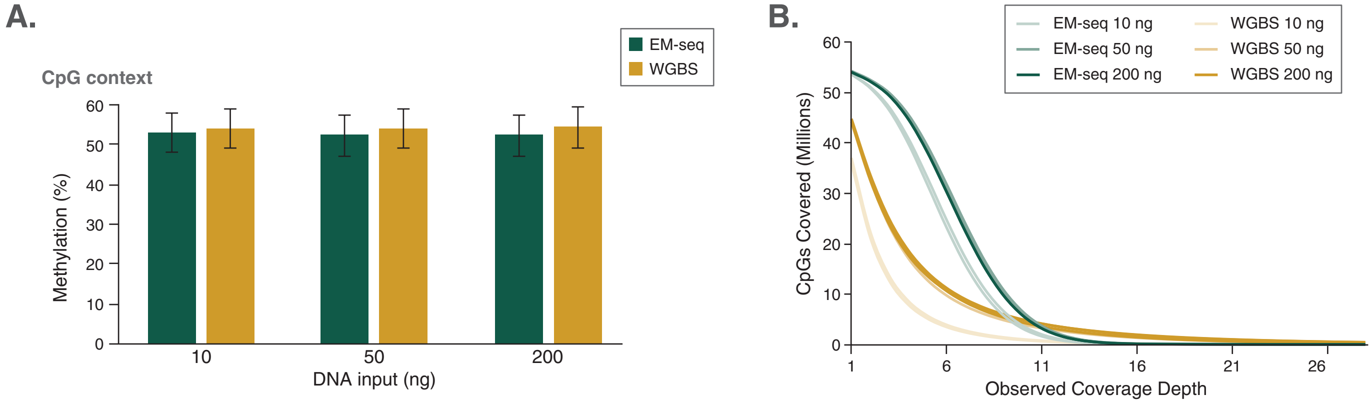
Bisulfite sequencing, while commonly used, is sub-optimal in detecting 5mC and 5hmC – large amounts of DNA are needed, DNA can be damaged, and sequences are biased towards AT-rich regions. Other methods that couple chemical or enzymatic treatment with bisulfite sequencing also share similar limitations. EM-seq provides the first commercially-available, non-bisulfite method that comprehensively addresses the limitations of bisulfite sequencing and represents a new opportunity for more complete methylome analysis. EM-seq libraries are not damaged and have longer inserts, higher PCR yields with fewer PCR cycles, and lack biases associated with GC content. More CpGs are identified with greater coverage depth using EM-seq, as compared to WGBS. These advantages all contribute to EM-seq having more usable sequencing data when comparing the same number of reads for EM-seq and WGBS, which ultimately reduces sequencing costs. EM-seq is the only commercially-available alternative to bisulfite sequencing that provides an effective method for accurate and comprehensive detection of 5mC and 5hmC across the genome, and offers a new, more accurate alternative for studying disease states.

FIGURE 7:
EM-seq identifies detect more CpGs to a higher depth of coverage than WGBS

10, 50 and 200 ng Human NA12878 genomic DNA was sheared to 300 bp using the Covaris S2 instrument and used as input into EM-seq and WGBS protocols. For WGBS, NEBNext Ultra II DNA was used for library construction, followed by the Zymo Research EZ DNA Methylation-Gold Kit for bisulfite conversion. Libraries were sequenced on an Illumina NovaSeq® 6000 (2 x 100 bases). 324 million paired end reads were aligned to hg38 using bwa-meth 0.2.2.

A: Methyl Dackel was used to determine methylation levels, which were found to be similar between EM-seq and WGBS.

B: Coverage of CpGs with EM-seq and WGBS libraries was analyzed, and each top and bottom strand CpGs were counted independently, yielding a maximum of 56 million possible CpG sites. EM-seq identifies more CpGs at lower depth of sequencing.



References

1. Harris R.A., et al. (2010) *Nat Biotechnol.* 28, 1097–1105.
2. Olova, N., et al. (2018) *Genome Biology,* 19: 33.
3. Yu, M., et al. (2012) *Cell,* 149, 1368–1380.
4. Booth, M.J., et al. (2012) *Science* 336, 934–937.
5. Schutsky, E.K., et al. (2018) *Nature Biotechnology,* 36, 1083–1090.
6. Zhang, L., et al. (2013) *Nat. Commun.* 4: 1517.
7. Carpenter, M.A., et al. (2012) *J. Biol.Chem.* 287, 34801–34808.
8. Wijesinghe, P. and Bhagwat, A.S. (2012) *Nucl. Acids Res.* 40, 9206–9217.
9. Josse, J and Kornberg, A. (1962) *J. Biol. Chem.* 237, 1968–1976.
10. Tomaschewski, J., et al. (1985) *Nucleic Acids Res.*13 (21): 7551–7568.
11. Schutsky, E.K., et al. (2017) *Nucleic Acids Res.* 45, 7655–7665.