NEBNext® Enzymatic Methyl-seq (EM-seq™)

A high-performance alternative to bisulfite sequencing for methylome analysis

Introduction

DNA methylation analysis, and specifically the identification of 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) within genomes, is important as these modifications are known to affect expression of genes. In general, low levels of methylation near transcription start sites are associated with higher transcription levels, while genes with high levels of cytosine modification in regulatory regions are expressed at lower levels. Complete and accurate methylome analysis is important in many fields: including the study of disease states such as cancer, in monitoring embryonic development, and in studies of agricultural plants. However, existing technologies for methylome analysis have significant drawbacks.

Whole Genome Bisulfite Sequencing (WGBS) has long been the gold standard for methylome analysis, but the chemical bisulfite reaction damages and degrades DNA, resulting in fragmentation and loss. Additionally, bisulfite libraries demonstrate significant GC bias and are enriched for methylated regions.

FIGURE 1: **The NEBNext Enzymatic Methyl-seq (EM-seq) workflow**

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	NEBNext® Ultra™ II reagents
EM-seq Adaptor Ligation	DNA is ligated to the EM-seq adaptors	
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	

To overcome these limitations, we developed an enzyme-based approach, NEBNext Enzymatic Methyl-seq (EM-seq), a new method for identification of 5mC and 5hmC.

The highly effective enzymatic conversion in this method minimizes damage to DNA and, with the supplied NEBNext Ultra™ II library preparation workflow reagents, produces high quality libraries that enable superior detection of 5mC and 5hmC from fewer sequencing reads. Conveniently, the EM-seq method results in the same converted sequence as WGBS and so the same analysis pipelines can be used.

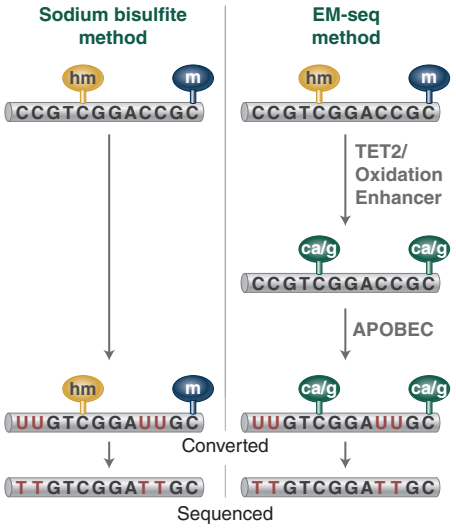
Workflow

In the EM-seq workflow (Figure 1), as with WGBS libraries, the first step is library construction from sheared DNA. For EM-seq, standard input amounts can range from 10-200 ng of sheared DNA, and a modified protocol is also available for input amounts as high as 500 ng. This is followed by two sets of enzymatic conversion steps to differentiate unmethylated cytosines from 5mC and 5hmC. Finally, libraries are PCR amplified before sequencing.

The two-step enzymatic conversion of EM-seq is shown in more detail in Figure 2, alongside bisulfite conversion.

Sodium bisulfite treatment of DNA results in the deamination of cytosines to uracils, while the modified forms of cytosine (5mC and 5hmC) are not deaminated. When bisulfite treated DNA is PCR amplified, uracils are replaced by thymines, and 5mC and 5hmC are replaced by cytosines. Once sequenced, unmethylated cytosines are represented by thymines and 5mC and 5hmC are represented by cytosines. By comparing sequences to reference sequences (C/T and G/A converted genome), the methylation status can be assessed.

FIGURE 2: **EM-seq and sodium bisulfite conversion methods**



The first EM-seq conversion step uses TET2 and an Oxidation Enhancer to protect modified cytosines from downstream deamination. TET2 enzymatically oxidizes 5mC and 5hmC through a cascade reaction into 5-carboxycytosine [5-methylcytosine (5mC) → 5-hydroxymethylcytosine (5hmC) → 5-formylcytosine (5fC) → 5-carboxycytosine (5caC)]. This protects 5mC and 5hmC from deamination. 5hmC can also be protected from deamination by glucosylation to form 5ghmc using the Oxidation Enhancer.

The second enzymatic step uses APOBEC, which deaminates cytosine but does not affect 5caC and 5ghmC.

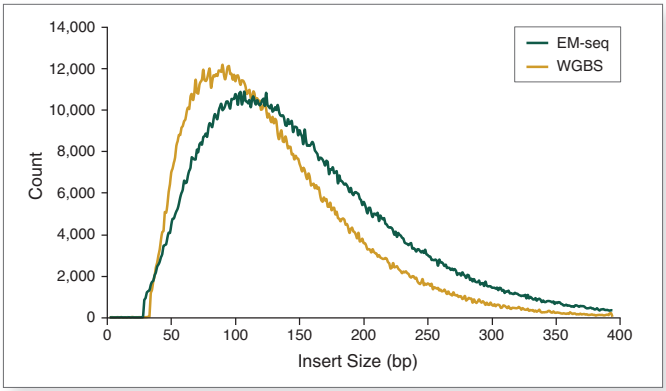
The resulting converted sequence is the same as that for bisulfite-treated DNA and so can be analyzed in the same way. Typical aligners used to analyze data include, but are not limited to, Bismark and bwa-meth.

Larger library insert sizes

The more gentle treatment of DNA by the steps in the EM-seq workflow, compared to the harsh bisulfite treatment, minimizes damage to DNA. As a result, EM-seq converted DNA is more intact than bisulfite converted DNA, resulting in libraries with a higher percentage of longer inserts, as shown in Figure 3. This enables longer sequencing reads, resulting in greater confidence in mapping, and potentially lower per-base sequencing costs depending on instrumentation and other sequencing reaction details.

FIGURE 3: **EM-seq libraries have larger insert sizes**

50 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® MiSeq® (2 x 76 bases) and insert sizes were determined using Picard 2.18.14. The normalized frequency of each insert size was plotted, illustrating that library insert sizes are larger for EM-seq than for WGBS, and indicating that EM-seq does not damage DNA as bisulfite treatment does in WGBS.



Increased library yields

The DNA damage, fragmentation and loss that result from bisulfite treatment reduce yields of bisulfite-converted libraries after amplification. In contrast, the more gentle treatment in the EM-seq workflow allows maintenance of high quality DNA libraries, which can be efficiently amplified. As a result, EM-seq library yields are not only higher than WGBS library yields, but are achieved with fewer PCR cycles (Figure 4A).

Importantly, these higher yields are not due to the presence of PCR duplicates, which can be especially apparent at low input amounts. Indeed, EM-seq libraries display consistently low levels of duplicates across a range of input amounts (Figure 4B).

FIGURE 4A: **EM-seq produces higher yields**

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. For all input amounts, EM-seq library yields were higher, and fewer PCR cycles were required, suggesting greater DNA loss in the WGBS protocol. Error bars indicate standard deviation.

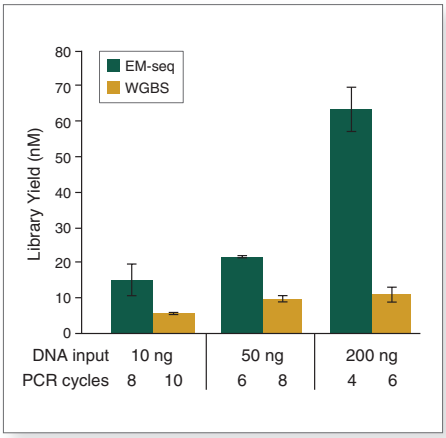
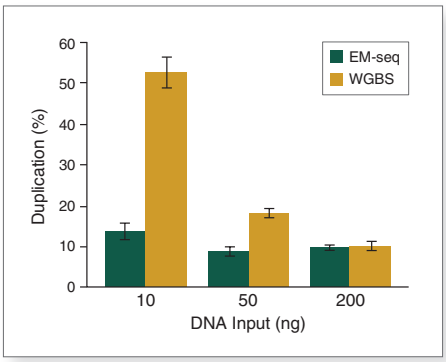


FIGURE 4B: **EM-seq results in lower duplication rates**

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), and reads were aligned to hg38 using bwa-meth 0.2.2. Duplication rates were determined using MarkDuplicates tool. Duplication levels in EM-seq libraries are lower than WGBS libraries, which is also consistent with the lower number of PCR cycles required for EM-seq libraries.



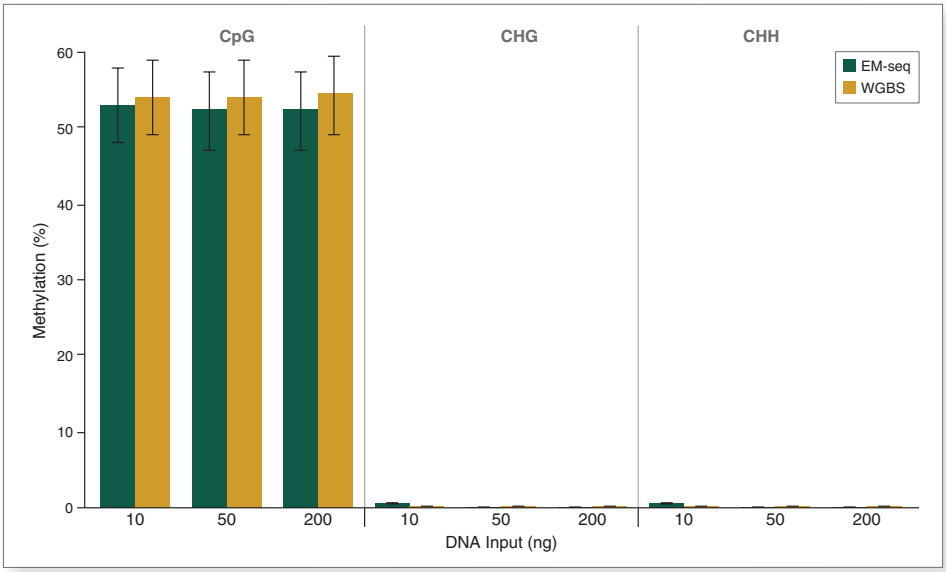
Global methylation levels

While in mammals, 5mC and 5hmC are found predominantly in the CpG context, in plants 5mC also occurs in CHG and CHH contexts, where H = A, C or T. To assess cytosine methylation detection accuracy, global methylation levels in CpG, CHG and CHH contexts were determined for human DNA. These were found to be similar between EM-seq and WGBS in each context (Figure 5).

FIGURE 5: **Total methylation detected by EM-seq and WGBS is similar**

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 for each library were aligned to hg38 using bwa-meth 0.2.2. and methylation information was extracted from the alignments using MethylDackel. Methylation levels for NA12878 are similar between EM-seq and WGBS in CpG, CHH and CHG contexts.



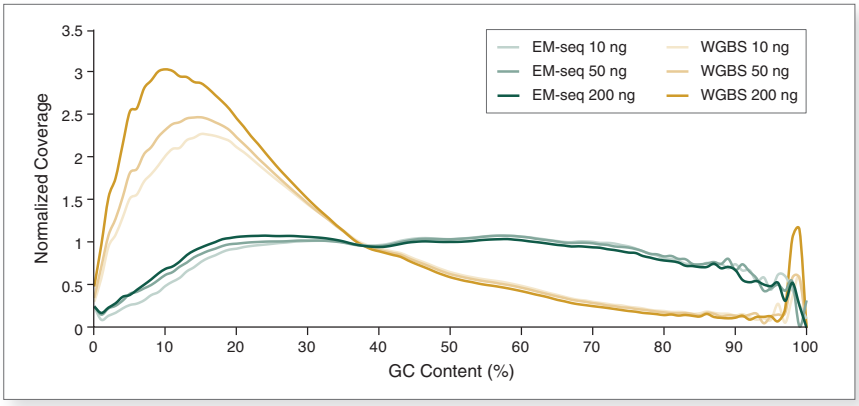
Uniformity of GC coverage

While sufficient yield of a library is required for successful sequencing, the quality of a library is also critical. A high-quality library will have uniform representation of the original sample, including uniform coverage across the GC spectrum.

Since bisulfite treatment acts upon and damages unmethylated cytosines, which comprise the majority of cytosines, this harsh treatment therefore disproportionately affects GC-containing regions. This context-specific damage, breakage and loss lead to bisulfite-treated libraries being under-represented for GC content and over-represented for AT. In contrast, EM-seq libraries show uniform GC coverage, highlighting the lack of damage to DNA, and that the libraries are representative of the original sample. (Figure 6).

FIGURE 6: **EM-seq has superior uniformity of GC coverage**

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). Reads were aligned to hg38 using bwa-meth 0.2.2. 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.



Increased CpG coverage with EM-seq

In depth analysis of sequencing data allows exploration of uniformity of genome coverage and representation of CpGs.

Examination of dinucleotide coverage distribution of libraries shows the variance in coverage for dinucleotides in sequence reads when compared to unconverted library dinucleotide distribution. This was plotted across all 16 possible dinucleotide combinations, for EM-seq and WGBS. While the EM-seq libraries show uniform coverage across all dinucleotide combinations, WGBS libraries are depleted in C-containing dinucleotides, and are enriched in A/T containing dinucleotides (Figure 7).

Analysis of the number of CpGs detected by each method, at different levels of stringency, allows a comparison of the sensitivity of CpG detection by EM-seq and WGBS. Figure 8 demonstrates that EM-seq libraries identified more CpGs than bisulfite libraries using the same number of reads (324 million) for 10 ng, 50 ng and 200 ng inputs, and EM-seq libraries show a higher percentage of CpGs covered, at the lower minimum coverage depths. For example, when the number of unique and common CpGs identified by EM-seq and WGBS libraries at 1X and 8X minimum coverage for a range of input levels is measured, EM-seq covers at least 20% more CpGs at 1X minimum coverage threshold compared to WGBS. This difference in CpG coverage increases to two-fold at 8X minimum coverage threshold.

FIGURE 7: **EM-seq libraries show uniformity in coverage for all dinucleotides**

Reads were aligned to hg38 using bwa-meth 0.2.2. Dinucleotide coverage distribution was plotted for all 16 possible combinations, comparing variance in coverage in the reads for EM-seq or WGBS with unconverted Ultra II library reads. EM-seq libraries show uniformity in coverage for all dinucleotides, while WGBS libraries are depleted in C-containing dinucleotides and enriched in A/T containing dinucleotides.

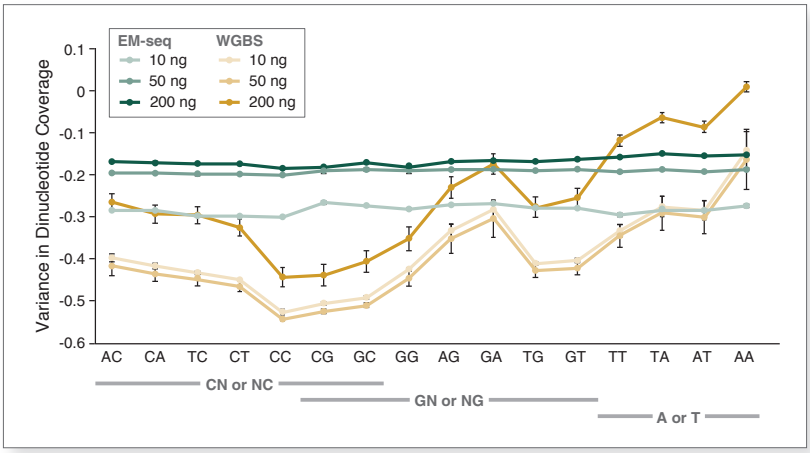
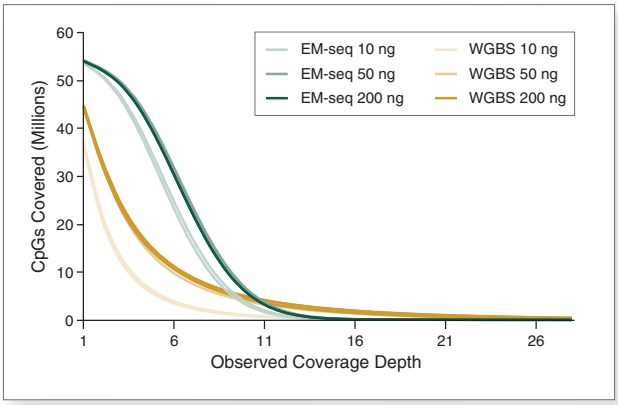


FIGURE 8: **EM-seq identifies more CpGs than WGBS, at lower sequencing coverage depth**

Reads were aligned to hg38 using bwa-meth 0.2.2. Coverage of CpGs with EM-seq and WGBS libraries was analyzed using 324 million paired end reads. 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.

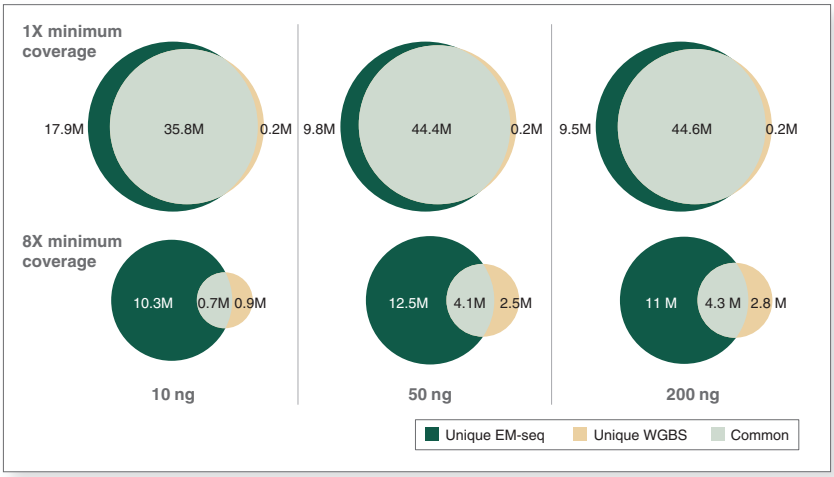


Plotting the number of unique and common CpGs detected by EM-seq and WGBS using the same number of sequence reads, for different DNA inputs and at different stringency levels, also highlights that EM-seq enables detection of more CpGs, and this difference is more striking at higher stringency (Figure 9). At 1X coverage, while EM-seq identified ~54 million CpGs using 10, 50 and 200 ng input DNA, WGBS identified ~36 to 46 million. Using 8X read coverage threshold, the number of CpGs identified using EM-seq ranged from 11 to 16.6 million, compared to 1.6 to 7.1 million for WGBS. This data demonstrates that more relevant data (CpG coverage) can be achieved with fewer sequence reads using EM-seq compared to WGBS.

FIGURE 9: **EM-seq identifies more CpGs than WGBS, at lower sequencing coverage depth**

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). Reads were aligned to hg38 using bwa-meth 0.2.2. Coverage of CpGs with EM-seq and WGBS libraries was analyzed using 324 million paired-end reads.

The number of unique and common CpGs identified by EM-seq and WGBS at 1X and 8x minimum coverage for each input amount are shown. EM-seq covers at least 20% more CpGs than WGBS at 1X minimum coverage threshold. The difference in CpG coverage increases to two-fold at 8X minimum coverage threshold.

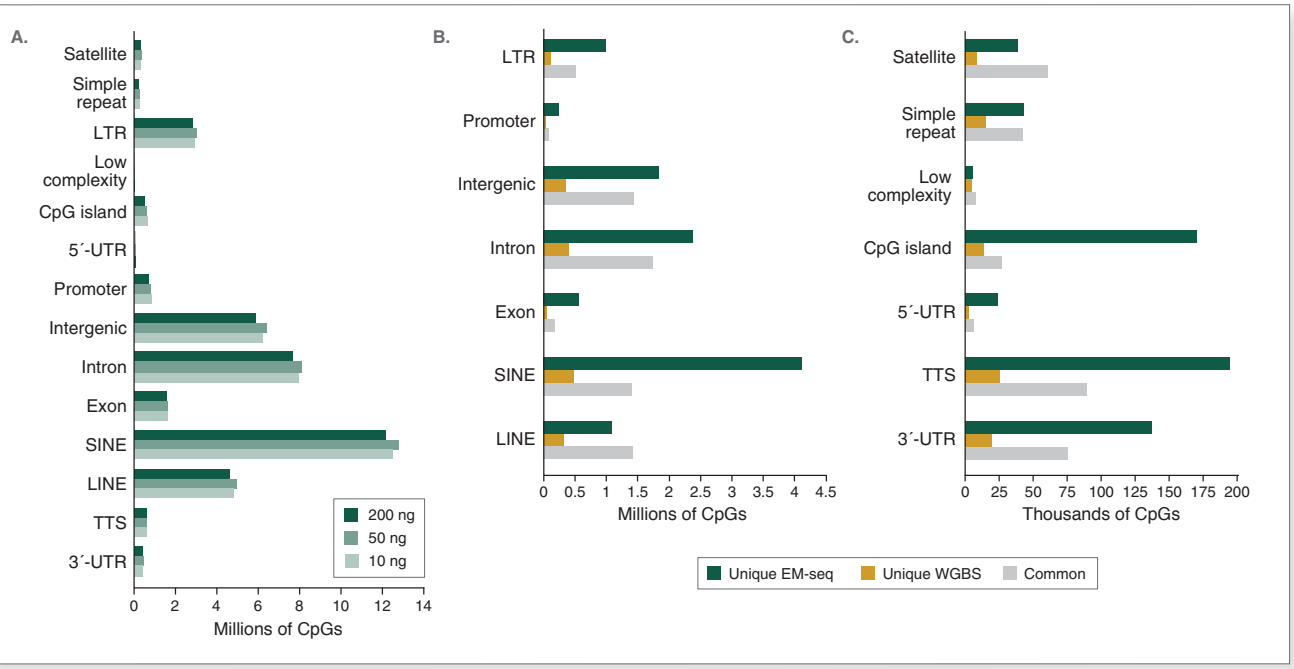


Distribution of CpGs across genomic features

Identification of genomic features should be consistent across the range of input amounts, and this is true for EM-seq (Figure 10A). Comparison of identification of these features, at stringent (8X) minimum coverage, shows that both EM-seq and WGBS identify a wide range of genomic features, including repeat elements and transcriptional regulatory regions. However, significantly more unique loci are consistently found in EM-seq at that coverage level (Figure 10B and C).

FIGURE 10: **EM-seq identifies more genomic features than WGBS, across a range of input amounts**

324 million paired reads were analyzed for each library and annotated using Homer. The number of genomic features identified is similar across a broad input range, indicating no loss in the power of CpG detection even with decreasing DNA input (A). EM-seq and WGBS identify a wide range of genomic features at 8X minimum coverage (B & C). These include repeat elements and transcriptional regulatory regions, with common loci identified within each genomic feature. However, more unique loci are found exclusively in EM-seq libraries at 8X minimum coverage threshold.

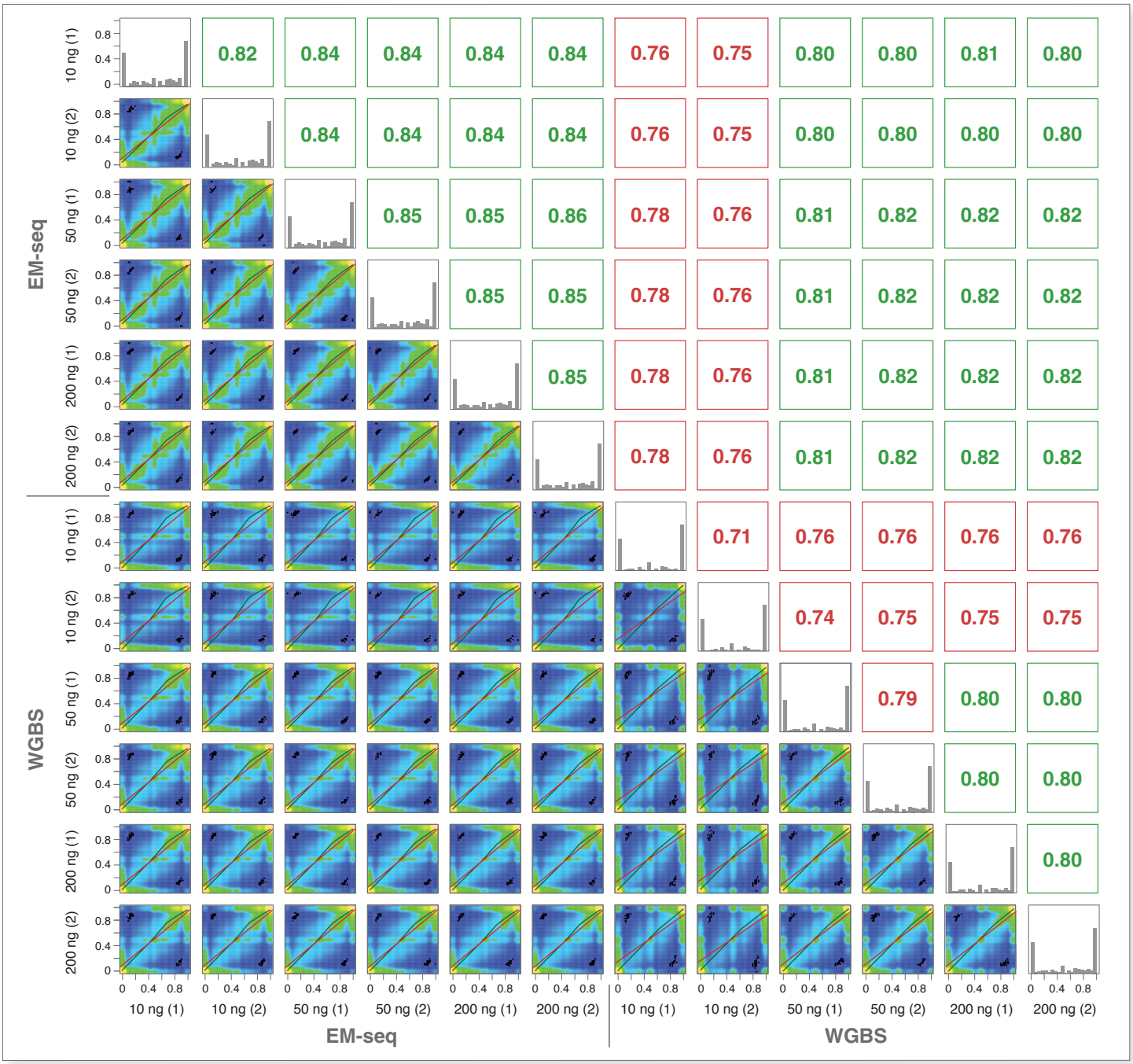


CpG coverage and consistency

Correlations between libraries from different input amounts and different methods is a useful measure of consistency as well as performance efficiency. Figure 11 shows CpG methylation correlations between EM-seq and WGBS libraries, from a range of input amounts, examining 21 million CpGs common to all 12 libraries shown. EM-seq correlations were high between all input amounts of EM-seq libraries, indicating that methylation detection sensitivity does not decrease with input amount. In contrast, correlations between WGBS libraries were the lowest. Correlations between EM-seq and WGBS libraries were highest at higher input amounts. Taken together, this data highlights the robustness of EM-seq methylation profiles.

FIGURE 11: **EM-seq libraries have higher CpG correlations than WGBS libraries**

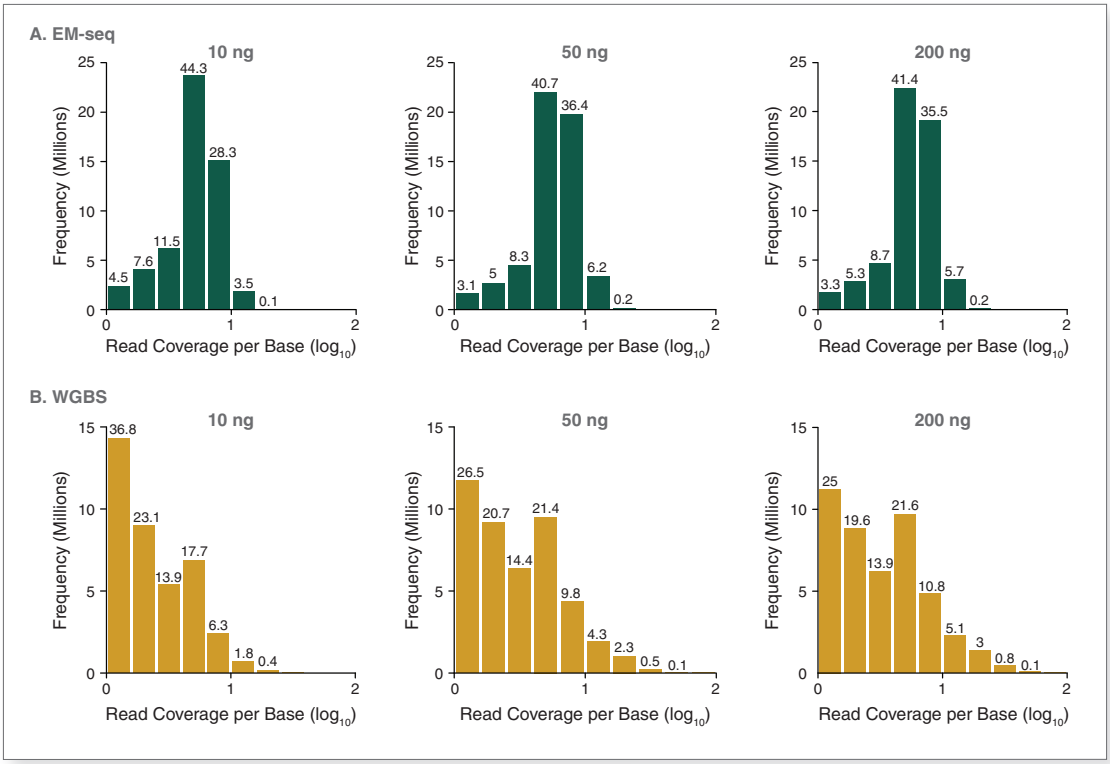
Correlations between 10, 50 and 200 ng EM-seq and WGBS libraries were plotted using methylKit with 1X minimum coverage (21 million CpGs common to all libraries). Correlations were highest between all inputs of the EM-seq libraries, with Pearson's correlation ranging from 0.82 to 0.86. Correlations between WGBS libraries ranged from 0.71 to 0.8. Comparisons between EM-seq and WGBS libraries had highest correlations with 50 ng and 200 ng WGBS inputs while the lowest correlations were observed with 10 ng WGBS DNA inputs. This data reveals that methylation profiles from EM-seq libraries are more robust, when compared both within EM-seq and across WGBS datasets.



The distribution of CpG coverage for EM-seq and WGBS is shown in Figure 12. While for WGBS, a high percentage of CpGs are covered at a lower level, especially for low input amounts, for EM-seq the majority of CpGs are covered at a higher level, and there is a tighter distribution of coverage level.

FIGURE 12: **More CpGs are covered using EM-seq than WGBS**

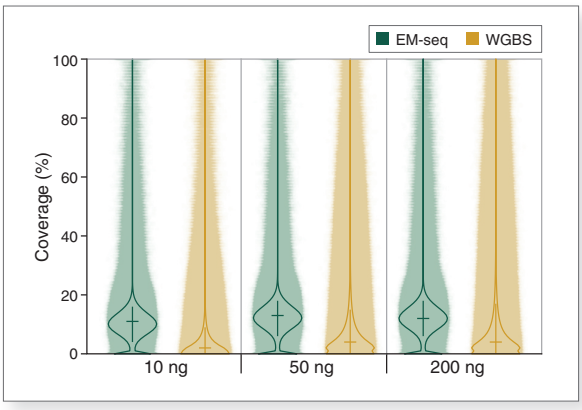
The read coverage per base for 10, 50 and 200 ng EM-seq and WGBS libraries was plotted using methylKit. The percentage of CpGs within a specific bin is displayed at the top of each bar. The histogram for EM-seq is shifted right for EM-seq compared to WGBS, showing that more CpGs are detected at higher coverage for EM-seq libraries than with WGBS, at all input amounts.



The difference in coverage level distribution is further exemplified in the violin plots in Figure 13, which shows substantially higher median CpG coverage for EM-seq compared to WGBS, at all input amounts.

FIGURE 13: **EM-seq CpG coverage is higher than WGBS across a range of input amounts**

Distribution of CpG coverage is displayed in violin plots and the median, 90th and 10th quantile of the observed CpG coverage for the distribution are represented. EM-seq data for all DNA inputs is similar, with a median CpG overage of around 11-13x. WGBS libraries had significantly lower median coverage (~4x), that decreased to around 2X for 10 ng inputs.



Transcriptional control features

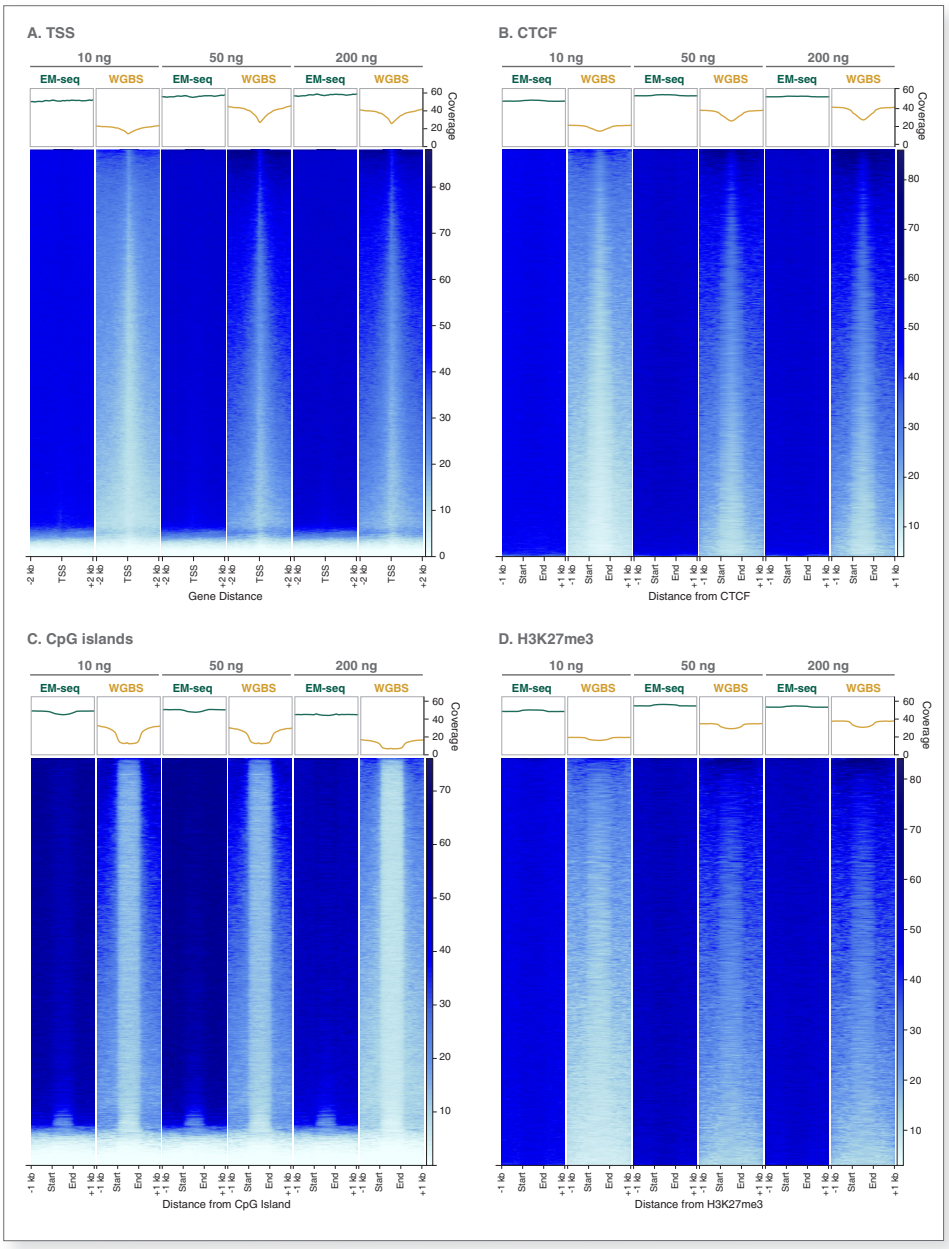
In expressed genes, the regions at transcription start sites are generally unmethylated, while genes that are expressed at lower levels generally have high levels of cytosine methylation in regulatory regions. It is therefore important to be able to accurately define transcriptional control elements. However, since, as described above, WGBS data traditionally has lower C-containing dinucleotide representation, this results in the potential loss of CpG data around important transcription control elements. The distribution of coverage 1-2kb around several important transcriptional elements was examined, for EM-seq and WGBS: Transcription start sites (TSS) (A), CTCF transcriptional repressor protein binding site (B), CpG islands (C) and H3K27me3 histone methylation site (D). The more complete landscape, provided by EM-seq methodology, around these important sites, due to both the greater levels of coverage overall, and the greater uniformity of coverage, enable more confident analysis.

FIGURE 14: **Enhanced representation of transcriptional control features using EM-seq**

The lower C-containing dinucleotide representation associated with WGBS results in the potential loss of methylated and non-methylated CpG data around important transcription control elements. Transcriptional activity is controlled by the methylation status of these elements and accurately defining transcriptional control elements is therefore important.

Heatmaps generated by deepTools show the distribution of coverage in a 2 kb window around transcription start sites (TSS) (A), and a 1kb window around CTCF transcriptional repressor protein binding site (B), CpG islands (C) and H3K27me3 histone methylation site (D).

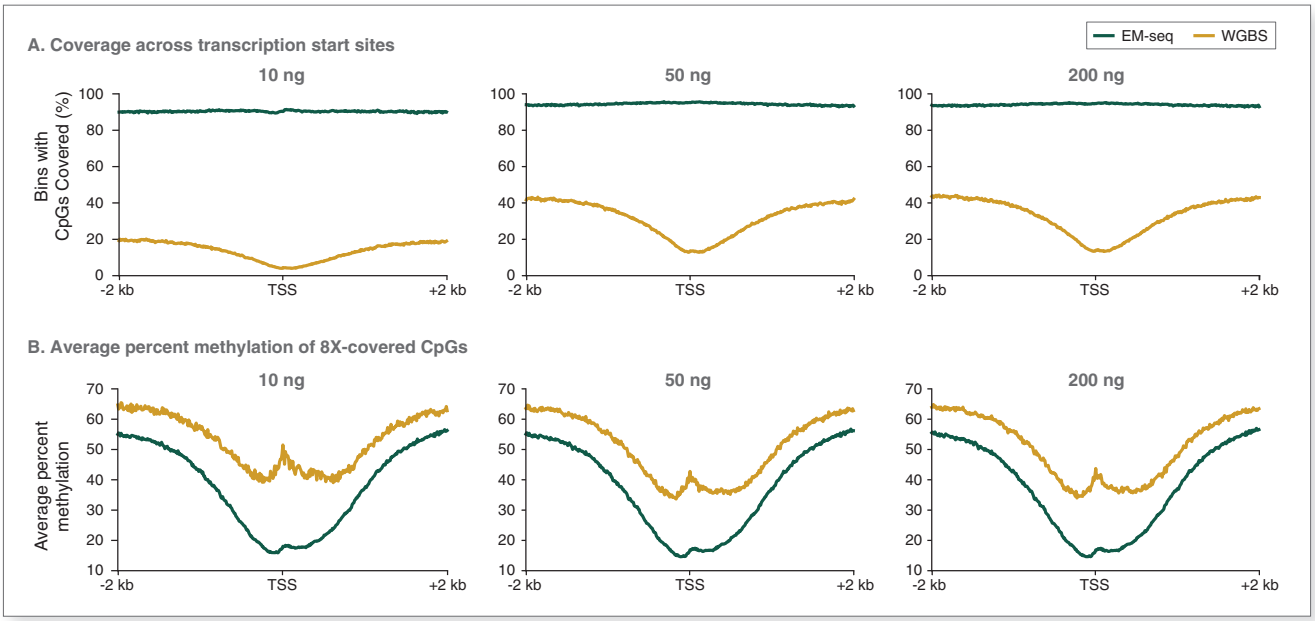
In all cases, EM-seq libraries have greater, and more uniform coverage than WGBS, and the enhanced coverage demonstrated using EM-seq results in fewer spurious methylation calls.



Examination of CpG coverage and methylation around transcription start sites is shown in Figure 15. With EM-seq, coverage of CpGs is both significantly higher than WGBS in these 4kb windows, and also strikingly uniform, lacking the dip in coverage close to the transcription start site (TSS) itself characteristic of WGBS. The information on methylation garnered from EM-seq at 8X stringency also aligns more closely with expected methylation patterns, with a reduction in methylation levels close to the TSS.

FIGURE 15: **EM-seq is superior to WGBS for detection of CpG methylation around transcription start sites**

Methylation of CpGs at each input was determined around the transcription start site (TSS). The 4 kb window around TSS was divided into 400 10 bp bins, and CpGs within these 10 bp bins with 8X or higher coverage were used to plot methylation. A: EM-seq has higher and more uniform coverage across TSSs . B: The average percentage methylation of 8X covered CpGs for EM-seq and WGBS libraries is shown. The EM-seq data is more representative of the expected methylation pattern across TSSs, with lowest levels at the TSS, and increasing methylation at the +/- 2 kb extremes.



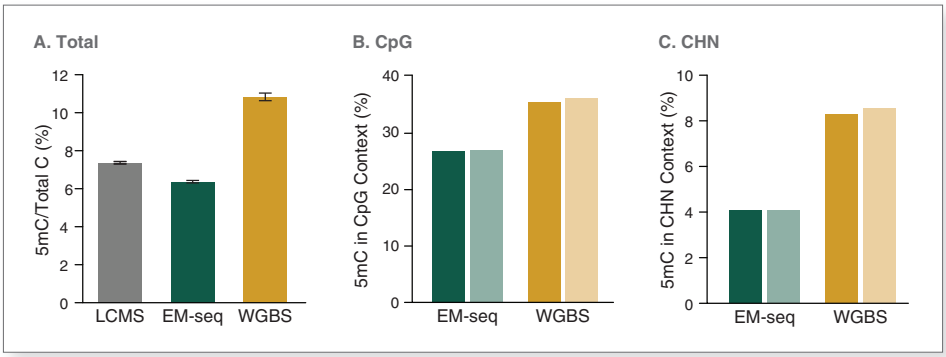
Methylome analysis in *Arabidopsis*

In plants, cytosine methylation is present in the context not only of CpG but also of CHG and CHH, where H = A, C or T. Total levels of 5mC in *Arabidopsis thaliana* were determined by LCMS (Liquid Chromatography-Mass Spectrometry), WGBS and EM-seq. LCMS data are closer to EM-seq than WGBS, and methylation percentages were also higher for WGBS for CpG and CHN contexts, suggesting over-estimation of methylated cytosines.

FIGURE 16: **EM-seq accurately represents 5mC levels within the *Arabidopsis thaliana* genome**

50 ng *A. thaliana* 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 NextSeq® 500 (2 x 75 bases). 125 million paired end reads for each library were aligned to TAIR10 using bwa-meth 0.2.2 and methylation information was extracted from the alignments using MethylDackel.

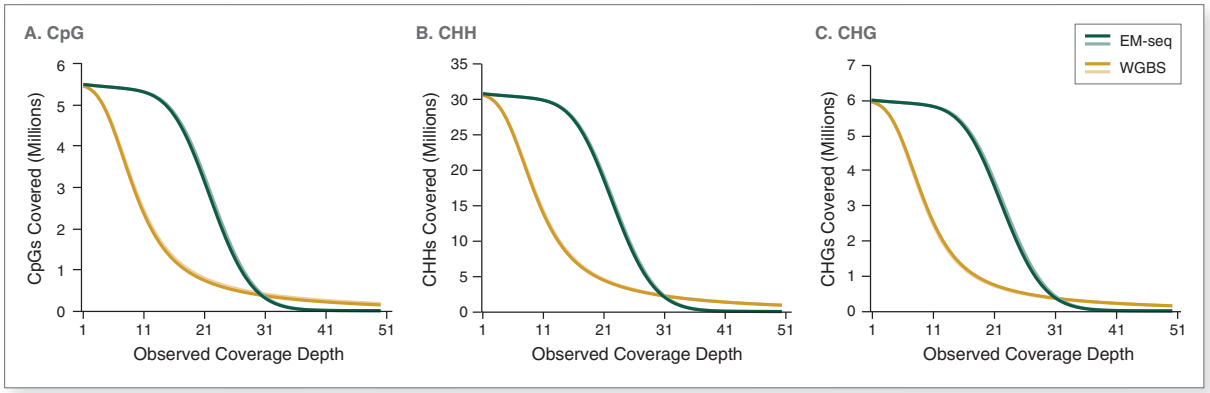
A. Total 5mC levels were compared between LCMS (Liquid Chromatography Mass Spectrometry) and sequencing data from EM-seq and WGBS libraries. EM-seq levels are close to the 5mC levels determined using LCMS. B and C. CpG and CHN methylation in Arabidopsis using EM-seq and WGBS. This data taken together indicate that WGBS suffers from over-estimation of methylated cytosines compared to EM-seq.



Sensitivity of methylation detection in *Arabidopsis thaliana* was also compared between EM-seq and WGBS, and CpG, CHH and CHG sites were identified and counted. Figure 16 demonstrates that EM-seq libraries cover more CpG, CHH and CHG sites at higher stringency than WGBS.

FIGURE 17: **With *Arabidopsis thaliana*, EM-seq identifies more CpGs than WGBS, at lower sequencing coverage depth**

50 ng *A. thaliana* 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 NextSeq 500 (2 x 75 bases). 125 million paired end reads for each library were aligned to TAIR10 using bwa-meth 0.2.2. CpG, CHH and CHG sites on both strands were counted independently. EM-seq identifies more CpGs, CHHs and CHGs, at higher coverage depth compared to WGBS, resulting in more usable information.



Conclusion

NEBNext Enzymatic Methyl-seq (EM-seq) is a new method for identification of 5mC and 5hmC. The enzyme-based conversion in the NEBNext Enzymatic Methyl-seq Kit minimizes damage to DNA and, with the supplied NEBNext Ultra II library preparation workflow reagents, produces high quality libraries that enable superior detection of 5mC and 5hmC from fewer sequencing reads.

- Superior sensitivity of detection of 5mC and 5hmC
- Greater mapping efficiency
- More uniform GC coverage
- Detection of more CpGs with fewer sequence reads
- Uniform dinucleotide distribution
- High-efficiency library preparation, with larger library insert sizes
- Utility with a range of sample types and input amounts

Ordering Information

PRODUCT	NEB #	SIZE
NEBNext Enzymatic Methyl-seq Kit	E7120S/L	24/96 rxns
NEBNext Enzymatic Methyl-seq Conversion Module	E7125S/L	24/96 rxns
NEBNext Multiplex Oligos for Enzymatic Methyl-seq (Unique Dual Index Primer Pairs)	E7140S/L	24/96 rxns
NEBNext Q5U™ Master Mix	M0597S/L	50/250 rxns