Enzymatic Methyl-seq: Next Generation Methylomes

Fiona J. Stewart[1], V K Chaithanya Ponnaluri[1], Brittany S. Sexton[1], Louise Williams[1], Bradley W. Langhorst[1], Romualdas Vaisvila[1], Yvonne Helbert[2], Lei Zhang[2], Timothy Harkins[2], Kevin J. McKernan[2], Eileen T. Dimalanta[1], Theodore B. Davis[1]

[1] New England Biolabs, Inc., Ipswich, MA 01938 [2] Medicinal Genomics Corp., Woburn, MA 01801

Introduction

DNA methylation is important for gene regulation. 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 Cs, however, BS does have its drawbacks. DNA is commonly damaged and degraded by the chemical bisulfite reaction resulting in libraries that demonstrate high GC bias and are enriched for methylated regions. To overcome these limitations, we developed an enzymatic approach, NEBNext® Enzymatic Methyl-seq (EM-seq™), for methylation detection that minimizes DNA damage, resulting in longer fragments and minimal GC bias.

Illumina libraries were prepared using bisulfite and EM-seq methods with 50 ng DNA from Arabidopsis thaliana and Cannabis sativa DNA. Libraries were sequenced using Illumina's NextSeq 500 (2x75). Reads were aligned using BWAMeth 0.2. and methylation information was extracted from the alignments using MethyDackel. Total 5mC levels were compared between the sequencing data from EM-seq and WGBS libraries and LCMS (*Liquid Chromatography Mass Spectrometry*). 5mC levels determined by EM-seq are close to those from LCMS, whereas, WGBS results in an overestimation of 5mC. Additionally, EM-seq libraries produce higher quality sequencing metrics such as longer inserts, lower duplication rates, a higher percentage of mapped reads and less GC bias compared to bisulfite converted libraries. We conclude that EM-seq is superior to WGBS and delivers higher library yields, more accurate methylation information, reduced DNA damage, increased sequencing length, and decreased GC-bias.

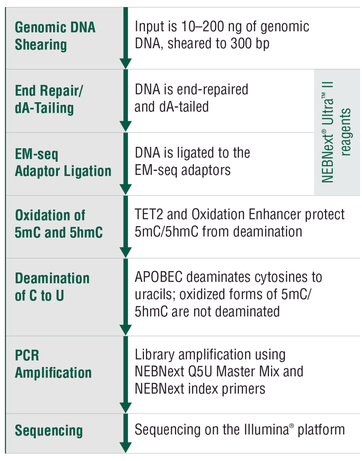
Methods

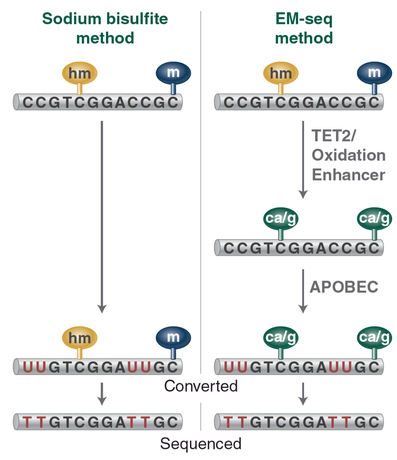
Sample Preparation

- Two plant DNAs were used to make EM-seq libraries
  - *Cannabis sativa* genomic DNA (Jamaican Lion): female clones (leaf, seeded and unseeded flowers) & male sibling (flowers) plants
  - *Arabidopsis thaliana*
- Libraries were made using 50 ng genomic DNA, spiked with control DNA (unmethylated lambda & CpG-methylated pUC19)
- Libraries were sequenced using an Illumina NextSeq 500, 2x76 base paired reads. 5caC is sequenced as C and deaminated C as T.
- Bisulfite conversion was performed using Zymo Research EZ DNA Methylation-Gold™ kit

Data Analysis

- Reads were aligned to Jamaican Lion reference genome (August 2018 assembly) or the *Arabidopsis* reference genome (TAIR10) (for Jamaican Lion, four miscellaneous contigs were removed from methylation analysis)
- Data were analyzed using the tools in above flowchart.

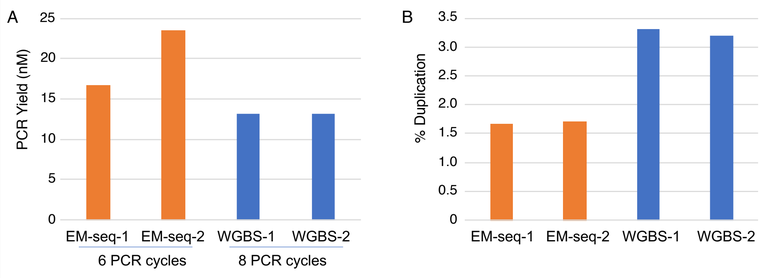




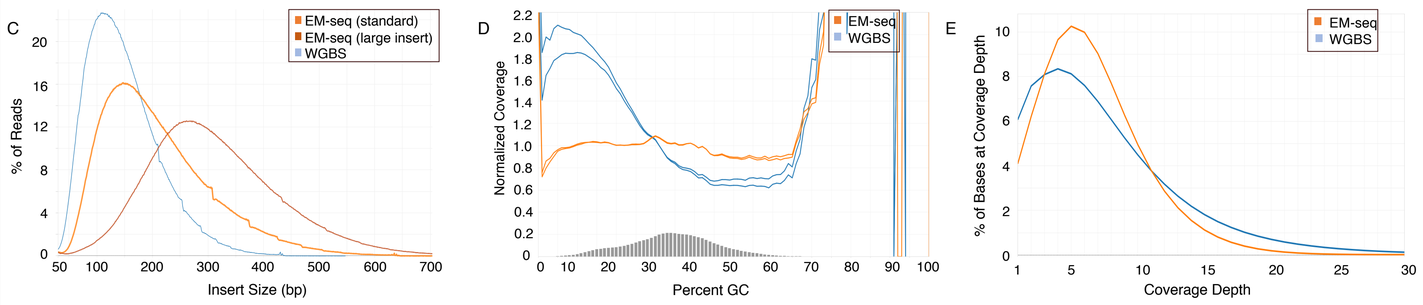
Methods

Cannabis sativa

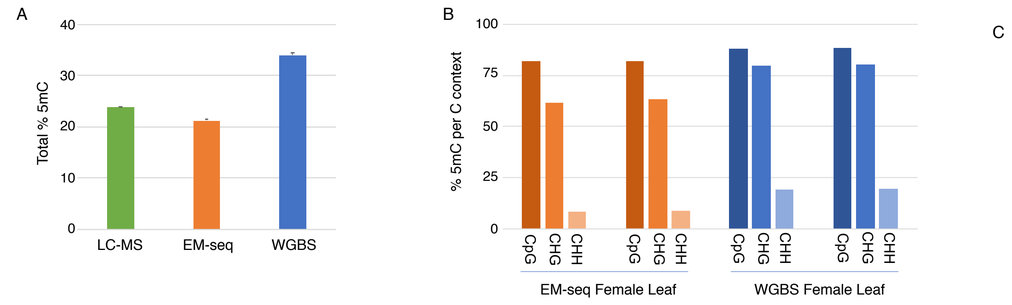
Cannabis sativa: Higher Quality Sequencing Metrics with EM-seq compared to WGBS



EM-seq libraries outperform WGBS for *Cannabis sativa* input DNA. (A) Fewer PCR cycles are required for EM-seq. (B) EM-seq libraries have lower duplication than WGBS. (C) EM-seq libraries have larger library insert sizes than WGBS. In addition, the EM-seq protocol has been optimized for both standard and large insert libraries. (D) GC distribution is more even for EM-seq than WGBS. (E) EM-seq has higher coverage depths than WGBS. 50 million 2 x 76 base reads were used for analysis (Figures A-D) and 130 million 2 x 76 base reads were used for Figure E.

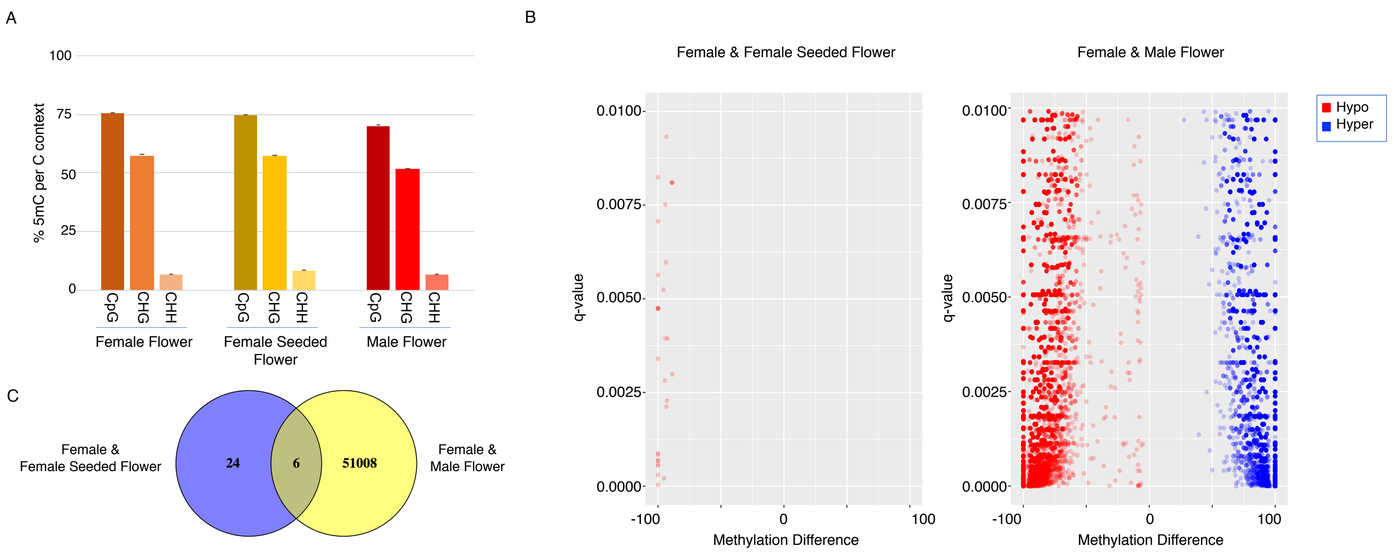


Cannabis sativa female leaf EM-seq libraries are superior to WGBS



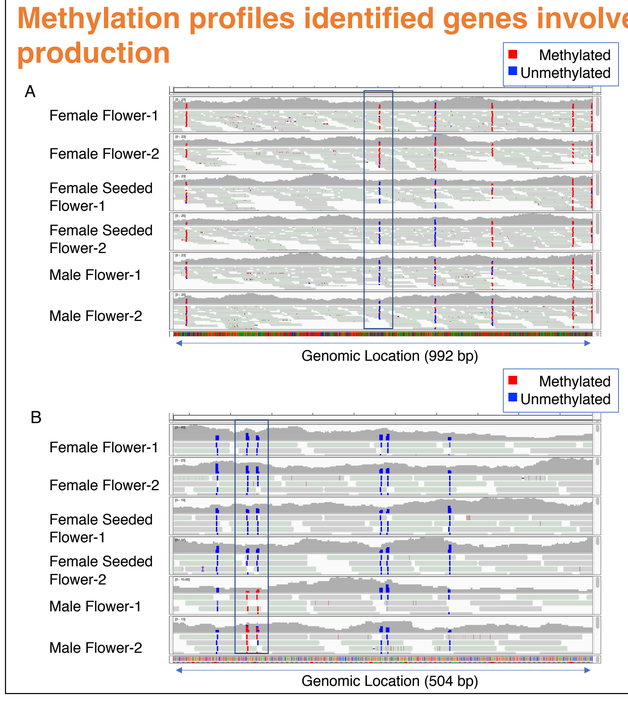
(A) Comparison of the total percentage of 5mC determined using LC-MS, EM-seq and WGBS for female leaf. The percentage of methylated cytosines using LC-MS is calculated by determining the amount of methylated cytosines and the total cytosines ((5mC/(total C)x100). 5mC percentages for EM-seq and WGBS were determined by combining 5mC in the three cytosine contexts (CpG, CHH, CHG). EM-seq cytosine methylation numbers are closer to LCMS methylation values than WGBS. (B) Cytosine methylation in the CpG, CHG, and CHH contexts for EM-seq and WGBS for female leaf. Control cytosine methylation for unmethylated lambda DNA was <0.5% for EM-seq and <2% for WGBS, and for CpG methylated pUC19 was >97.5% for both EM-seq and WGBS (data not shown). (C) CpG context correlations of EM-seq and WGBS libraries (1x minimum coverage). Both EM-seq and WGBS libraries were highly correlated between replicates and methods (CHG and CHH context data not shown). 50 million, 2 x 76 base reads were used for methylation analysis.

Differential CpG methylation identified between *Cannabis* flower tissues



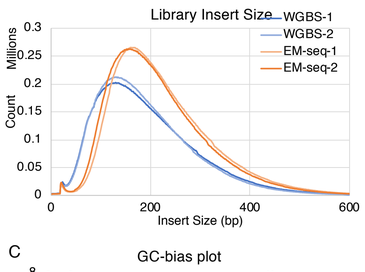
Differential methylation across *Cannabis sativa* flower tissues using EM-seq data. Female flower and female seeded flower (clones) as well as the male flower (sibling) were studied. (A) Percent cytosine methylation in the CpG, CHG, and CHH contexts. Female methylation levels for both flower and seeded flower were higher than for the male flower indicating methylation patterns are potentially determined by sex. Control DNA methylation levels were <0.5% for unmethylated lambda and >97.5% for CpG methylated pUC19 (data not shown). (B) Volcano plot of the significant (q<0.01) differential methylation calls for the CpG context between (1) female flower and female seeded flower (2) female flower and male flower. The differential methylation calls for female flower and female seeded flower identified 30 hypomethylated CpGs but no hypermethylated CpGs. The differential methylation calls for female and male flower identified >50,000 hypomethylated CpGs and >11,000 hypermethylated CpGs. (C) Comparison of the hypomethylated CpGs with differential methylation between the female flower samples & female and male flowers.

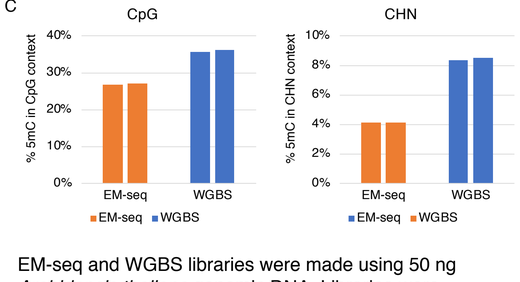
Methylation profiles identified genes involved in seed & cannabinoid production

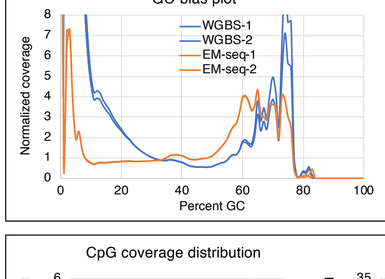


Differential methylation analysis comparing the female flower with the seeded flower (clones) and the male flower (sibling). Significant differential methylation calls (q<0.01) for CpG context (1x coverage minimum) between female flower with female seeded flower and/or male flower were identified. (A) Region 1 BLASTs to the edestin gene family of the *Cannabis sativa* genome. These genes are linked to the positive regulation of seed production. The female flower is CpG methylated in this region, suggesting the expression of the genes are turned off, while the female seeded flower is unmethylated at this CpG. (B) Region 2 BLASTs to the THC acid synthase (THCA) gene of the *Cannabis sativa* genome and this gene is linked to the positive regulation of THCA production. The female flower is CpG unmethylated in this region, suggesting the expression of the gene is turned on, while the male flower is methylated at this CpG, suggesting the expression of the gene is turned off. Male flowers produce log scale less THCA.

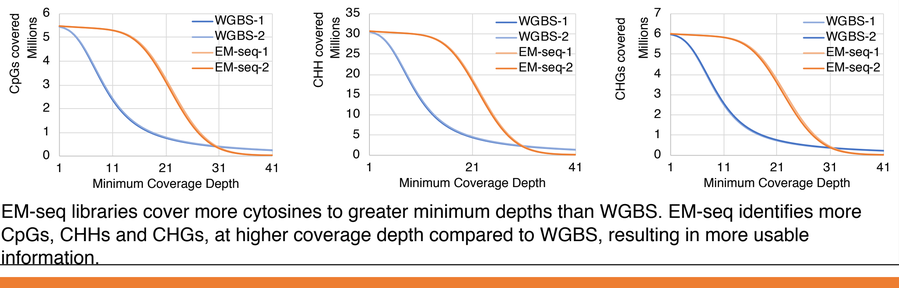
Arabidopsis thaliana



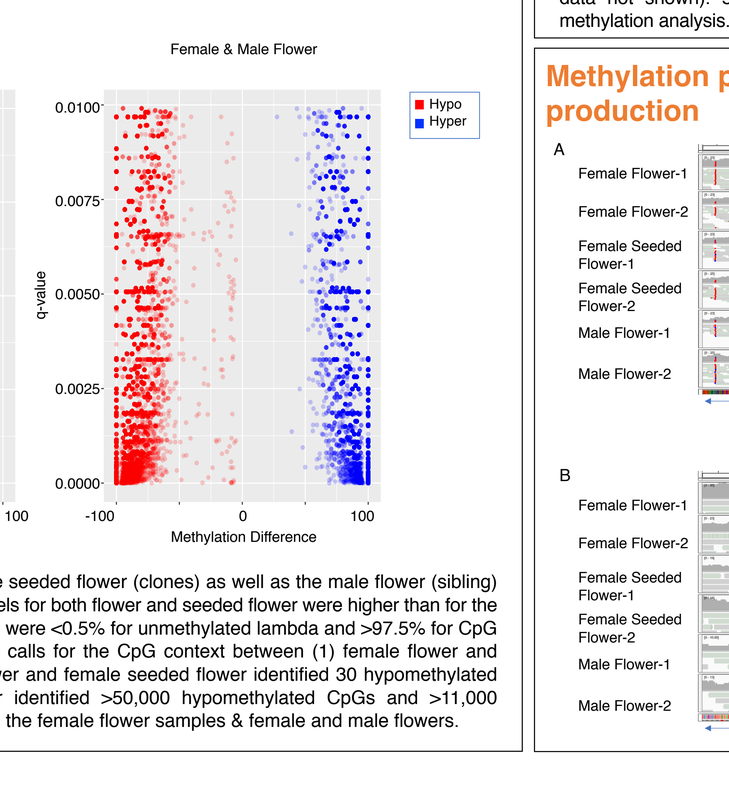




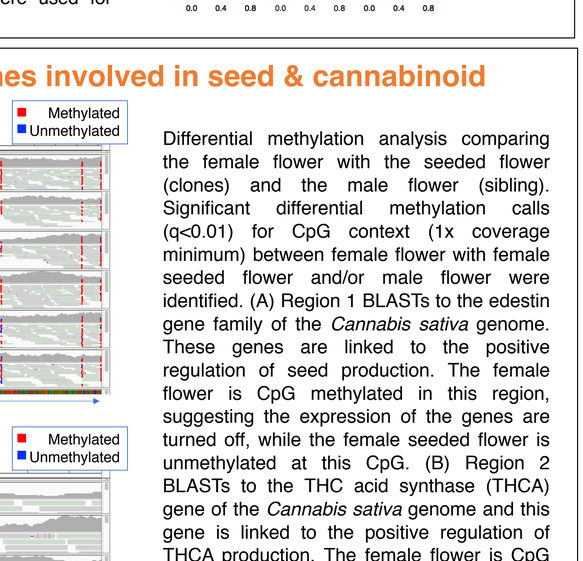
EM-seq and WGBS libraries were made using 50 ng *Arabidopsis thaliana* genomic DNA. Libraries were sequenced on an Illumina NextSeq 500 (2 x 75 bases). 125 million paired end reads for each library were aligned to TAIR10 using bwa-meth 0.2.2. CpG, CHH and CHG sites on both strands were counted independently. (A) EM-seq libraries have larger insert sizes. (B) Cytosine methylation in the CpG and CHN contexts for EM-seq and WGBS. (C) GC Bias plot for EM-seq and WGBS libraries. EM-seq libraries show less bias across the GC content.



EM-seq libraries cover more cytosines to greater minimum depths than WGBS. EM-seq identifies more CpGs, CHHs and CHGs, at higher coverage depth compared to WGBS, resulting in more usable information.



Histograms showing CpG, CHG and CHH coverage of EM-seq (A) and WGBS (C) 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. Percent methylation for libraries made using EM-seq (B) and WGBS (D) are shown in the CpG, CHG and CHH contexts. EM-seq and WGBS data both show that most cytosines in the CHG and CHH contexts are not methylated. Methylation is found in the CpG context for both EM-seq and WGBS libraries.



Cytosine correlations of EM-seq and WGBS libraries (1x minimum coverage). Both EM-seq and WGBS libraries were highly correlated between replicates and methods). 125 million, 2 x75 base reads were used for methylation analysis. Correlations in the CpG, CHG and CHH contexts are shown

Conclusion

EM-seq can be used to investigate plant genomic DNA

- analysis of the *Cannabis sativa* methylome identified genes involved in seed and THC production
- the *Arabidopsis* methylome was successfully probed

EM-seq libraries compared to WGBS libraries had:

- Higher library yields with fewer PCR cycles
- Lower percent duplication
- More even base coverage
- Larger library insert sizes
- Less GC bias
- Similar percentage methylation as LC-MS

Acknowledgements
Thank you to Laurie Mazzola and Danielle Fuchs for their technical support.

www.NEBNext.com