Uncovering the *Cannabis sativa* methylome through Enzymatic Methyl-seq

Brittany S. Sexton[1], Louise Williams[1], V.K. Chaithanya Ponnaluri[1], Keerthana Krishnan[1], Luo Sun[1], Yvonne Helbert[2], Lei Zhang[2], Christine J. Sumner[1], Fiona J. Stewart[1], Bradley W. Langhorst[1], 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

Cannabis sativa is an industrial crop producing more economic value than the top five crops in California combined. In the medical field, cannabinoids are used to treat and alleviate symptoms for a wide range of diseases, including nausea in cancer patients undergoing chemotherapy. Sex determination in cannabis is an important economic factor. Pollination dramatically reduces cannabinoid expression in female plants. To express high levels of cannabinoids, female plants are grown in isolation of male pollen. However, female plants are capable of stress induced hermaphroditism that can lead to acres of pollination and low cannabinoid yields.

5-methylcytosine (5mC) is an important epigenetic mechanism that regulates many cellular processes, including sex determination. To better understand the hermaphroditic process in cannabi*s*, more robust 5mC surveying tools are required. The current gold standard for determining 5mC, whole genome bisulfite sequencing (WGBS), introduces DNA damage and GC-biased sequences resulting in skewed methylation profiles. To further complicate matters, the cannabis genome is 66% AT-rich and 83% AT-rich after bisulfite conversion. Very few technologies currently address the complexity of plant methylation signals, and to date no methylome exists for cannabis.

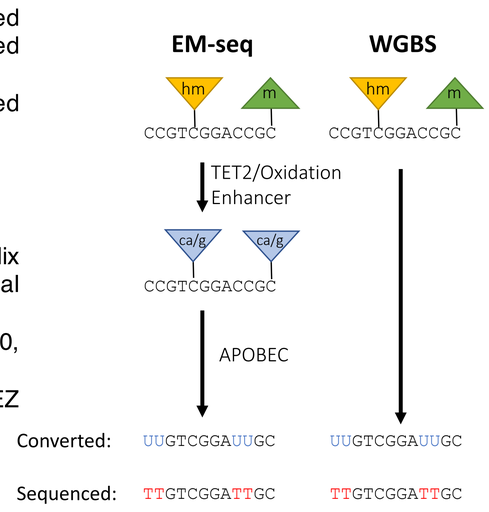
NEBNext Enzymatic Methyl-seq (EM-seq™) is a novel enzymatic method used to detect 5mC. EM-seq successfully identified the 5mCs within the cannabis genome. Libraries were generated using genomic DNA extracted from female leaf for EM-seq and WGBS. To give insight into the methylation-based regulation of the hermaphroditic process, EM-seq libraries were also prepared using genomic DNA from female flowers, female seeded flowers, and male flowers. 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. EM-seq enables the study of the previously unknown cannabis methylome and opens up the possibility to discover critical and exciting regulatory functions for the hermaphroditic process.

METHODS

SAMPLE PREPARATION

Sheared Genomic DNA → Ultra II End Prep/ Ligation → TET2/Oxidation Enhancer → APOBEC → PCR Amplification → Illumina Sequencing

- Jamaican Lion genomic DNA was extracted from female clones (leaf, seeded and unseeded flowers) and male sibling (flowers) plants.
- 50 ng of genomic DNA, spiked with control (unmethylated lambda DNA and CpG methylated pUC19) were sheared using the Covaris® S2 instrument
- Sheared DNA was end-repaired then ligated to modified sequencing adaptors
- 5mC and 5hmC were protected from APOBEC deamination by TET2/Oxidation Enhancer
- Cytosines were deaminated to uracils with APOBEC
- Libraries were amplified using NEBNext® Q5U™ Master Mix and NEBNext Multiplex Oligos for Illumina (96 Unique Dual Index Primer Pairs, E6440S)
- Libraries were sequenced using an Illumina Nextseq 500, 2x76 base paired reads.
- Bisulfite conversion was performed using Zymo Research EZ DNA Methylation-Gold™ kit



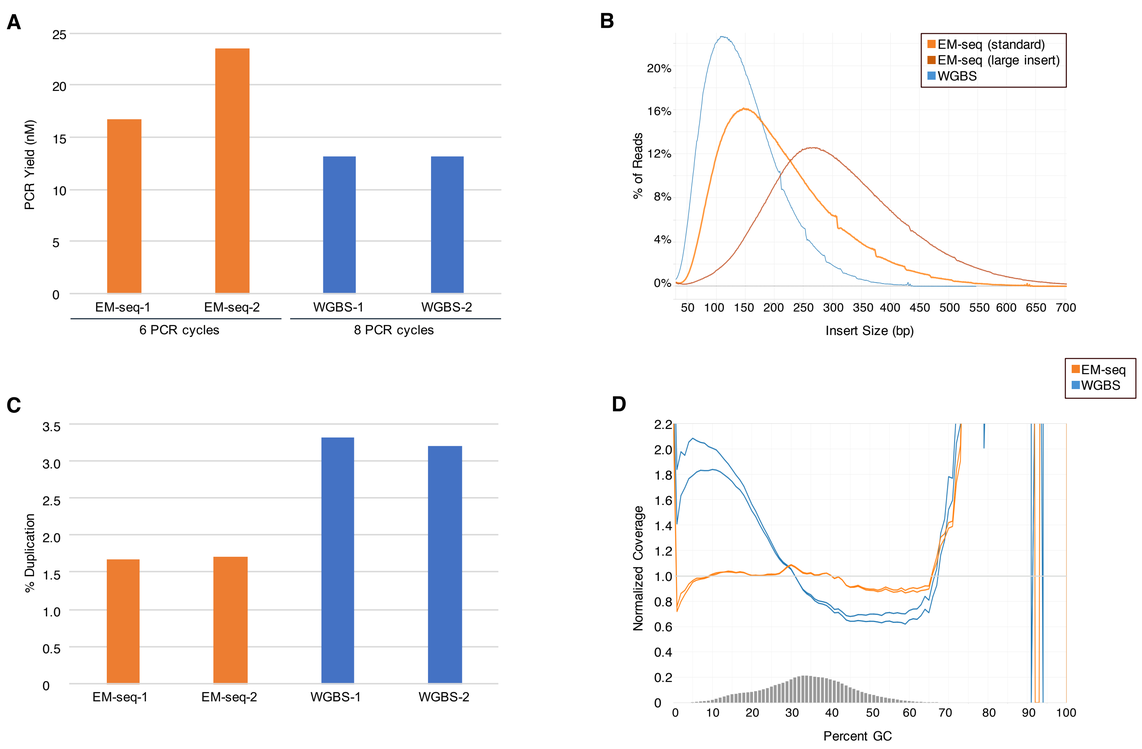
DATA ANALYSIS

Align (BWAmeth) → Samblaster Deduplicate → Picard → MethylDackel Methyl Extraction → Methyl Kit

- Reads were aligned to Jamaican Lion reference genome (August 2018 assembly) and analyzed using the tools in above flowchart.
- Four contigs from this assembly showed anomalous methylation patterns and were excluded.

RESULTS

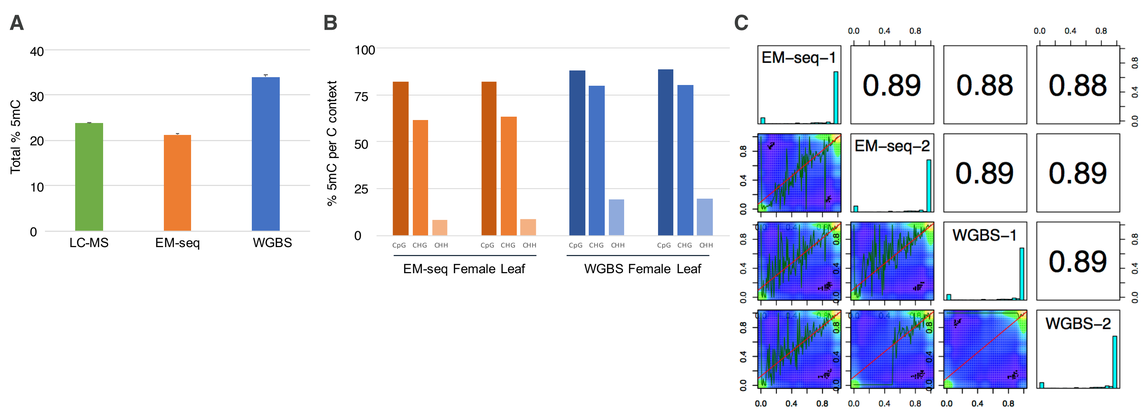
HIGHER QUALITY SEQUENCING METRICS WITH EM-SEQ COMPARED TO WGBS



EM-seq libraries outperform WGBS for *Cannabis sativa* input DNA as shown in figures A-E: (A) Less PCR cycles are required (6 vs 8 cycles respectively) for EM-seq. (B) EM-seq libraries have larger library insert sizes than WGBS. Additionally, the EM-seq protocol has been optimized for both standard and large insert libraries. (C) EM-seq libraries have lower duplication percentages. (D) GC distribution is more even for EM-seq than WGBS. (E) More even base coverage for EM-seq 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.

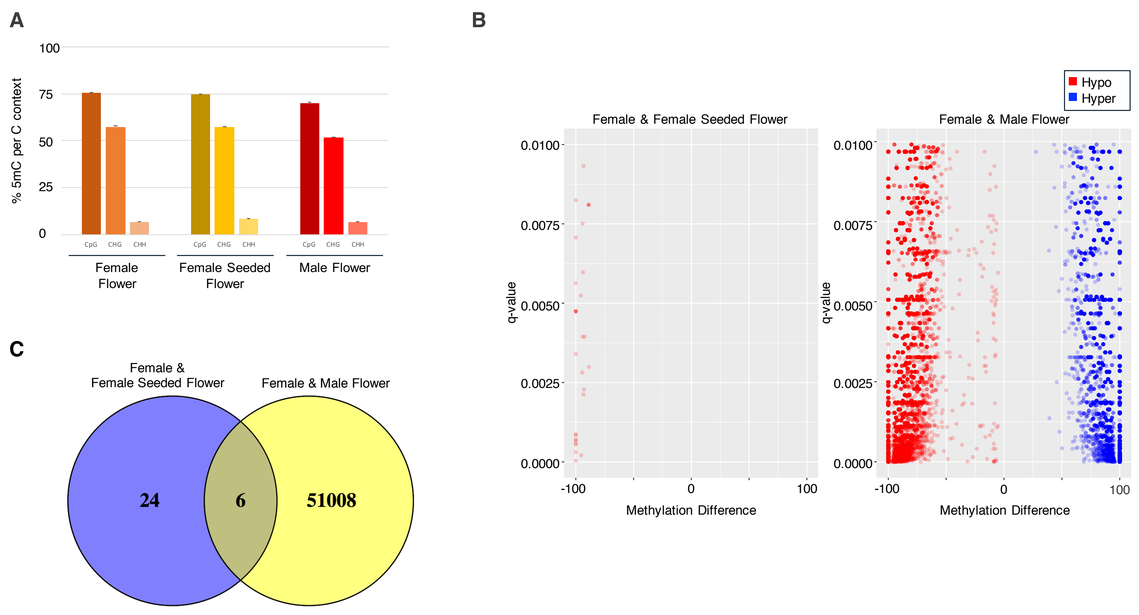
RESULTS

CANNABIS EM-SEQ LIBRARIES ARE SUPERIOR TO WGBS



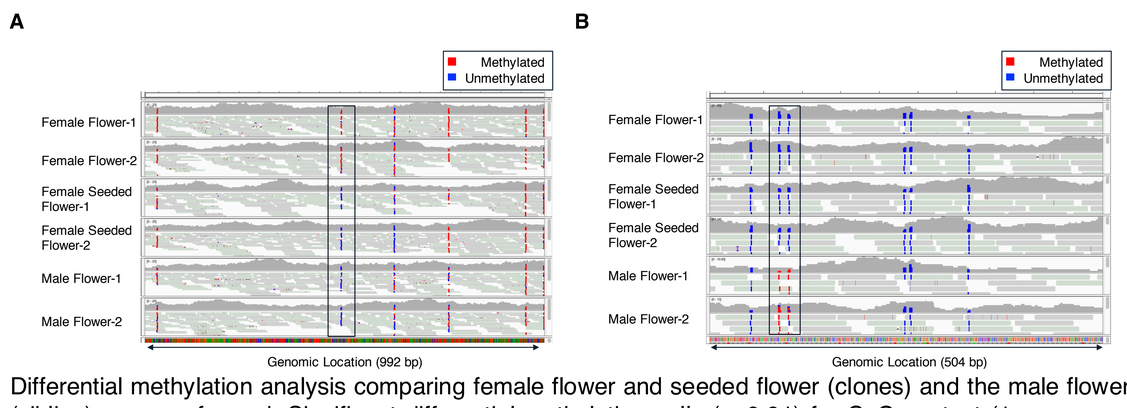
(A) Comparison of total % 5mC determined using LC-MS, EM-seq and WGBS for female leaf. The % 5mC for LC-MS is determined as a percentage of total cytosines ((5mC/(total C))x100). 5mC percentages for EM-seq and WGBS were determined by combining 5mC in the 3 methylated cytosine contexts (CpG. CHH, CHG). EM-seq cytosine methylation numbers are closer to LCMS methylation values than WGBS. (B) % 5mC levels in the CpG, CHG, and CHH contexts for EM-seq and WGBS for female leaf. Unmethylated Lambda control was <0.5% methylated for EM-seq and <2% for WGBS and was >97.5% CpG methylated 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 x76 base reads were used for methylation analysis.

DIFFERENTIAL CpG METHYLATION IDENTIFIED BETWEEN CANNABIS FLOWER TISSUES



EM-seq enables the identification of differential methylation across *Cannabis sativa* flower tissues: female flower and female seeded flower (clones) and the male flower (sibling). (A) % 5mC levels in the CpG, CHG, and CHH contexts for all flower EM-seq libraries. The female methylation levels for flower and seeded flower were higher than the male flower indicating methylation patterns are potentially determined by sex. Unmethylated Lambda control was <0.5% methylated and pUC19 was >97.5% CpG methylated (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. (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 AND CANNABINOID PRODUCTION



Differential methylation analysis comparing female flower and seeded flower (clones) and the male flower (sibling) were performed. 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. Interestingly, 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.

CONCLUSIONS

EM-seq enables the first in depth analysis of the *Cannabis sativa* methylome and differential methylation identified genes involved in seed production and THCA production.

EM-seq libraries compared to WGBS libraries had:

- Higher Library Yields with less PCR cycles
- Lower Percent Duplication
- More Even Base Coverage
- Larger Library insert Sizes
- Less GC Bias
- Similar Percentage Methylation as LC-MS

REFERENCES

1. Faust, G. G., and I. M. Hall. “SAMBLASTER: Fast Duplicate Marking and Structural Variant Read Extraction.”
2. Pedersen et al. (2014) “Fast and Accurate Alignment of Long Bisulfite-Seq Reads.”
3. Ryan D, Pederson, B “MethylDackel”
4. Broad Institute. (2016). Picard tools.