Enzymatic Methyl-seq: Next Generation Methylation

Fiona J. Stewart1, V K Chaitanya Ponnaluri1, Brittany S. Sexton1, Louise Williams1, Bradley W. Langhorst1, Romualdas Vaisvilas1, Yvonne Helbert2, Lei Zhang2, Timothy Harkins3, Kevin J. McKernan2, Eileen T. Dimalanta1, Theodore B. Davis1

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

Introduction

DNA methylation is an important gene regulation ability. The ability to accurately identify S-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) gives us greater insight into potential gene regulatory mechanisms. Bisulfite sequencing (BS) is traditionally used to detect methylated C, however, BS also changes ~5% of the DNA. 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, ENEMase.

Enzymatic Methyle-seq (EM-seq)1–3, for methylation detection that minimizes DNA damage, resulting in higher hybridization and minimal GC bias.

Bisulfite libraries were prepared using bisulfite and EM-seq methods with 50 ng DNA from Arabidopsis thaliana and Cannabis sativa DNA, which was sequenced on a HiSeq 2500/2000 instrument (Illumina). Reads were aligned using Bismark v.0.24.2 and methylation information was extracted from the alignments using Methyler2. Total EM-seq reads were compared to the bisulfite sequencing data from EM-seq and WGBS libraries and LCMS (Liquid Chromatography Mass Spectrometry). SimC values determined from EM-seq were lower than those from LCMS, whereas, WGBS results in an overestimation of SmC. 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 (Jansen's Line) female clones (seeded and unseeded flowers) & male stigmas (flowers) plants
- Arabidopsis thaliana

- Libraries were made using 30 ng genomic DNA, spiked with control DNA (unmethylated samples & 5mC methylated (5mC-C))
- Bismarck v.0.24.2 and methylation information was extracted from the alignments using Methyler2
- Reads were aligned using Bismarck v.0.24.2 and methylation information was extracted from the alignments using Methyler2
- Total EM-seq reads were compared to the bisulfite sequencing data from EM-seq and WGBS libraries and LCMS (Liquid Chromatography Mass Spectrometry).

- SimC values determined from EM-seq were lower than those from LCMS, whereas, WGBS results in an overestimation of SmC. 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.

Data Analysis

- Reads were aligned to Jansen’s Line reference genome (August 2018 assembly) or the Arabidopsis reference genome (TAIR 10) for Methylation analysis.
- Data were analyzed using the tools in above flowchart.

Cannabis sativa

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

Differential CpG methylation identified between Cannabis flower tissues

- EM-seq libraries outperform WGBS for Cannabis sativa input DNA. (A) Fewer PCR cycles are required for EM-seq libraries than WGBS libraries. (B) EM-seq libraries have a larger insert size than WGBS. (C) EM-seq protocol has been optimized for both standard and high-quality libraries. (D) GC distribution is more even for EM-seq than for WGBS. (E) EM-seq coverage is higher than WGBS for a similar fraction of inserts (80% of inserts 100-2000 base pairs are read for WGBS, and 22% for EM-seq (A)).

- Differential methylation across Cannabis sativa flower tissues using EM-seq data. Female flower and female seeded flower (clones) as well as the male flower (stamens) 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 0.12% for methylated lambda and 0.75% for CpG methylated lambda (data not shown). (B) Violin plot of the significant (p<0.01) differential methylation calls for the CpG context between (A) female flower and female seeded flower (B) female flower and male flower. The differential methylation calls for female flower and female seeded flower identified 50 hypomethylated CpGs but no hypermethylated CpGs. The differential methylation calls for female and male flower identified >500 hypomethylated CpGs and >100 hypermethylated CpGs. (C) Comparison of the hypomethylated CpGs with differential methylation between the female-sampled flowers and males. The statistical analysis was performed using DEseq2 (GSE141161).

Arabidopsis thaliana

Methylation profiles identified genes involved in seed & cannabinoide production

Differential methylation analysis comparing the female flowers with the seeded flowers (clones) and the male flower (stamens).

- Significant differential methylation calls (p<0.01) for CpG context (1x coverage minimum) between female flower with female seeded flower and/or male flower were identified. (A) Region 1 (BLAST) is the left) adenosine gene family of the Cannabis sativa genome. These genes are linked to the positive regulation of seed production. The female flower in CpG methylated in this region, suggesting the expression of the genes are male specific while the female seed is unmethylated at this CpG. (B) Region 2 BLAST to the THC exist sequence (THCA) gene of the Cannabis sativa genome and this gene is unmethylated in the positive regulation of THC 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 sex-specific. Male flowers produce less THC.

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 profiled

- EM-seq libraries compared to WGBS libraries had
  - Higher library yields with fewer PCR cycles
  - Lower percent duplication
  - More even base coverage

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

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

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

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

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

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

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

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

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