E5hmC-seq™: Detection of 5hmC at single base resolution

Daniel J. Evanich, V. K. Chaithanya Ponnaluri, Vaishnavi Panchapakesa, Laura Blum,
Matthew A. Campbell, Bradley W. Langhorst, Louise Williams | New England Biolabs, Inc.

Introduction

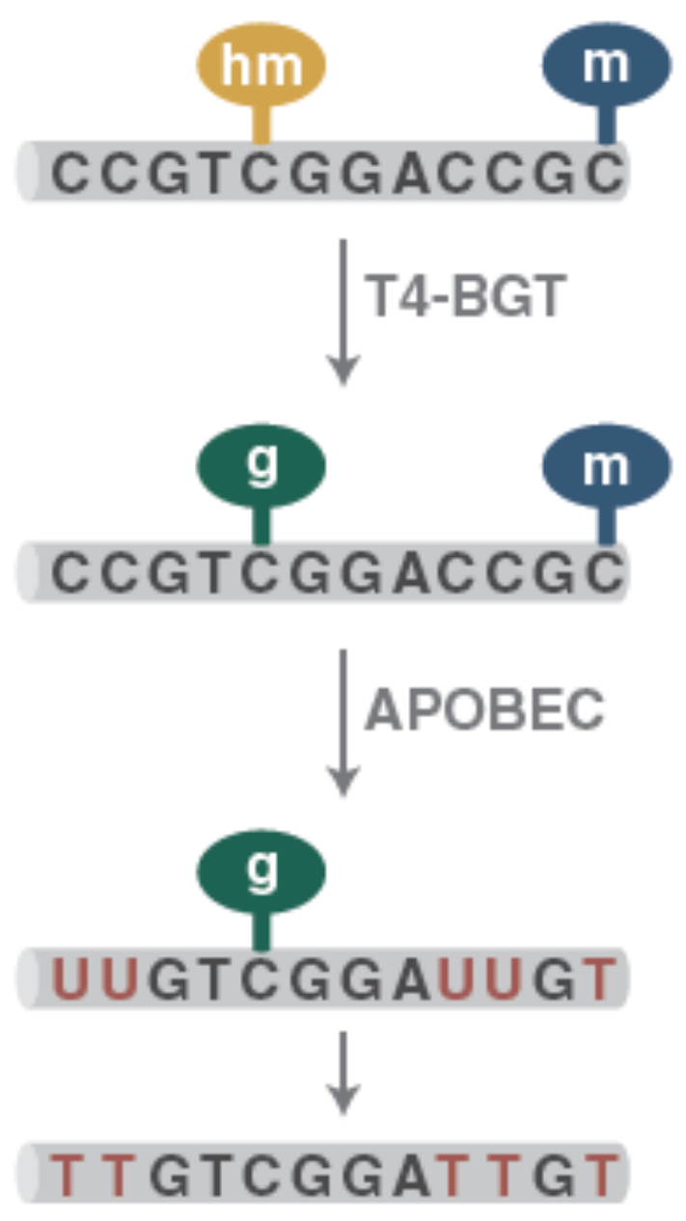
DNA methylation is an epigenetic regulator of gene expression with important functions in development and diseases such as cancer. Typically, the modified cytosines, 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC), are detected by sequencing Illumina libraries generated using an enzyme-based workflow called NEBNext® EM-seq™ or bisulfite conversion. However, these methods cannot differentiate between 5mC and 5hmC. Identification of specific 5hmC sites is important as there is increasing interest in its role in regulating gene expression in embryonic stem cells and several neuronal cell types. Methods currently exist to enable discrimination of 5mC and 5hmC (for example, oxBS-seq and TAB-seq), however these are based on modifications of bisulfite sequencing, and suffer from reduced data quality due to fragmentation and loss of DNA. Here we describe an enzymatic method that enables specific detection of 5hmC, termed NEBNext Enzymatic 5hmC-seq (E5hmC-seq™).

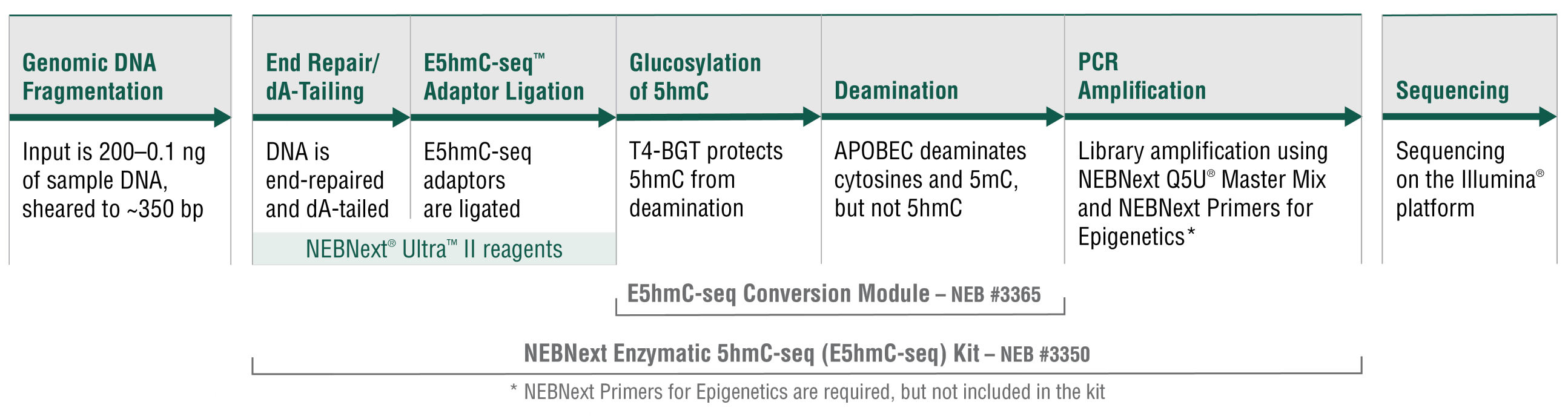
E5hmC-seq libraries were generated for 0.1 ng to 200 ng DNA isolated from human brain. Libraries were prepared using NEBNext Ultra™ II reagents followed by two enzymatic steps to detect 5hmC. In the first step, 5hmCs are glucosylated, which protects them from subsequent deamination by APOBEC. In contrast, cytosines and 5mCs are deaminated resulting in their conversion to uracil and thymine, respectively. This conversion allows discrimination of 5hmC from cytosine and 5mC. Finally, libraries were PCR amplified using NEBNext Q5U® and sequenced on the Illumina® platform.

The E5hmC-seq libraries had similar characteristics to EM-seq libraries, including expected insert sizes due to intact DNA molecules, low duplication rates and minimal GC bias. Moreover, E5hmC-seq libraries were well-correlated between inputs and replicates at higher sequencing depths. T4 phage DNA (all cytosines are 5hmC) was used as an internal control, with 98-99% of cytosines identified as 5hmC. E5hmC-seq libraries provide accurate measurements of 5hmC across a wide input range with expected insert sizes and minimal GC bias. The ability to discriminate between 5hmC and 5mC will provide key insights into the role of these cytosine modifications in development and disease.

Methods

E5hmC-seq: Enzymatic Detection of 5hmC Overview





Input is 200–0.1 ng of sample DNA, sheared to ~350 bp

DNA is end-repaired and dA-tailed

E5hmC-seq adaptors are ligated

NEBNext® Ultra™ II reagents

T4-BGT protects 5hmC from deamination

APOBEC deaminates cytosines and 5mC, but not 5hmC

Library amplification using NEBNext Q5U® Master Mix and NEBNext Primers for Epigenetics*

Sequencing on the Illumina® platform

E5hmC-seq Conversion Module – NEB #3365

NEBNext Enzymatic 5hmC-seq (E5hmC-seq) Kit – NEB #3350

* NEBNext Primers for Epigenetics are required, but not included in the kit

E5hmC-seq Library Construction Workflow:

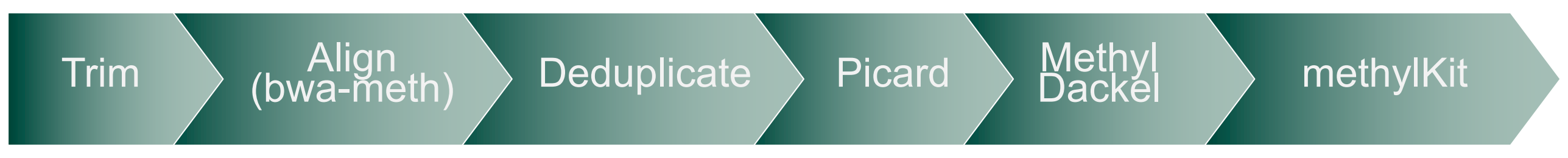
- DNA was Covaris® sheared, end-repaired and ligated to E5hmC-seq adaptors
- 5hmCs were glucosylated using UDP-glucose & T4-BGT. APOBEC then deaminates 5mC and cytosine but not 5hmC.
- Libraries were amplified using NEBNext Q5U Master Mix.

Library Construction

E5hmC-seq libraries:

- 200 ng, 10 ng, 1 ng, 0.5 ng & 0.1 ng of human brain genomic DNA were spiked with the control DNAs: unmethylated lambda and T4 phage DNA (all cytosines are 5hmC modified)
- Libraries were constructed following the E5hmC-seq workflow
- Libraries were sequenced using an Illumina NovaSeq® 6000 using 2x 150 base reads. Read numbers used in analysis were:
    - 1.9 billion reads for 200 ng, 10 ng and 1 ng inputs
    - 715 million reads for the 1 ng and 0.1 ng inputs

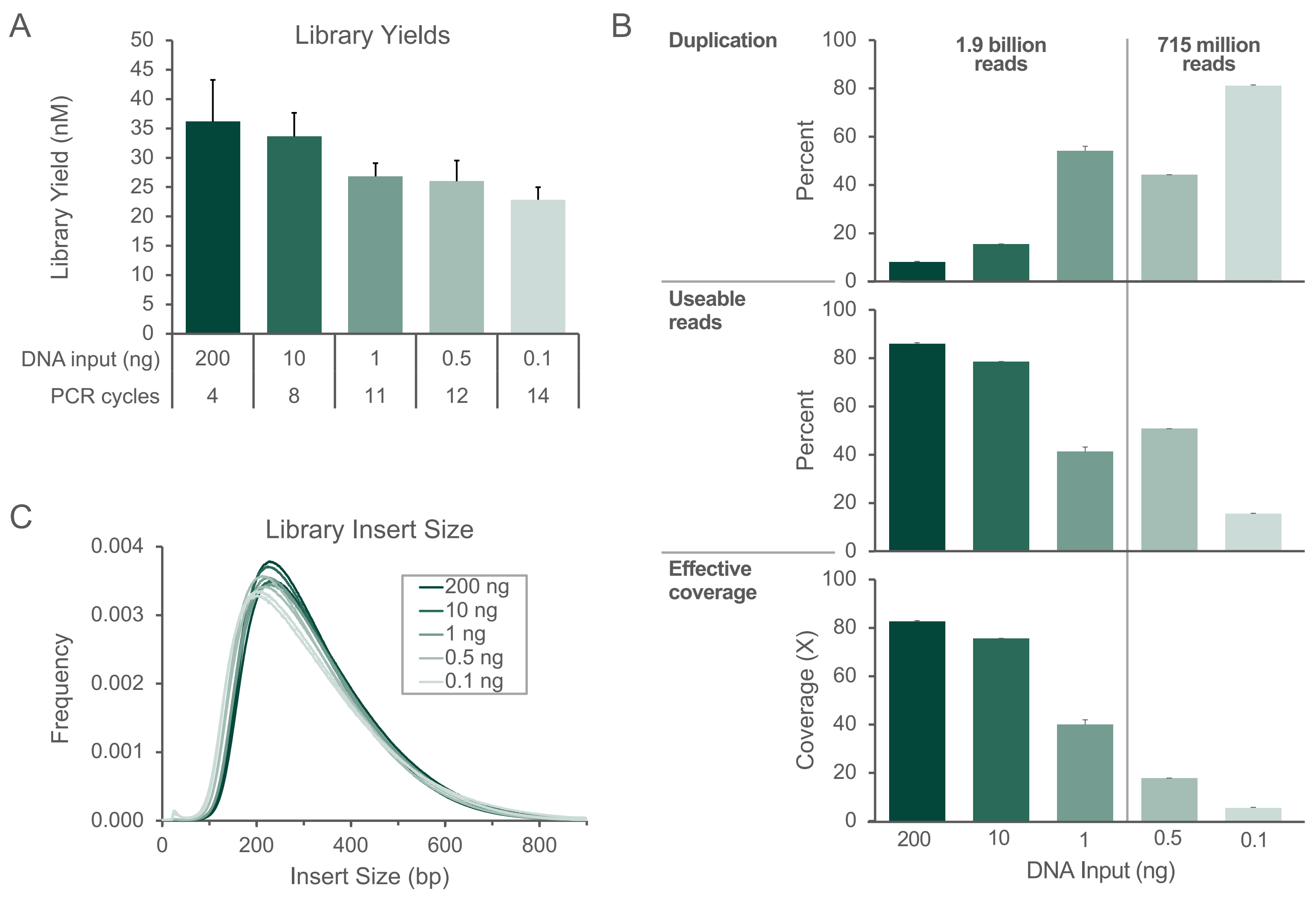
Sequencing and Data Analysis



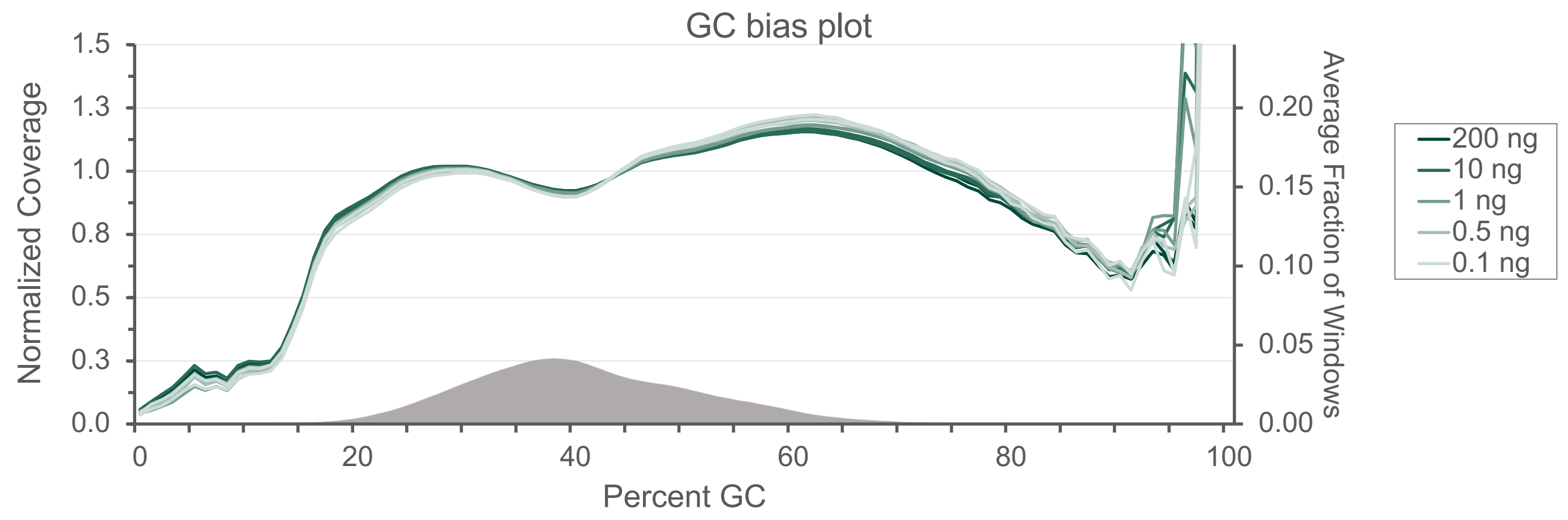
- Reads were adaptor trimmed (fastp) then aligned to a human T2T composite genomes using bwa-meth
- 5hmC information was extracted from the alignments using MethylDackel and levels were evaluated independently for each chromosome
- methylKit data was used for Pearsons correlation at 1x minimum coverage
- Picard was used to mark duplicates as well as calculate library insert size and GC bias

Results

E5hmC-seq Library Profiles and Metrics for Human Brain



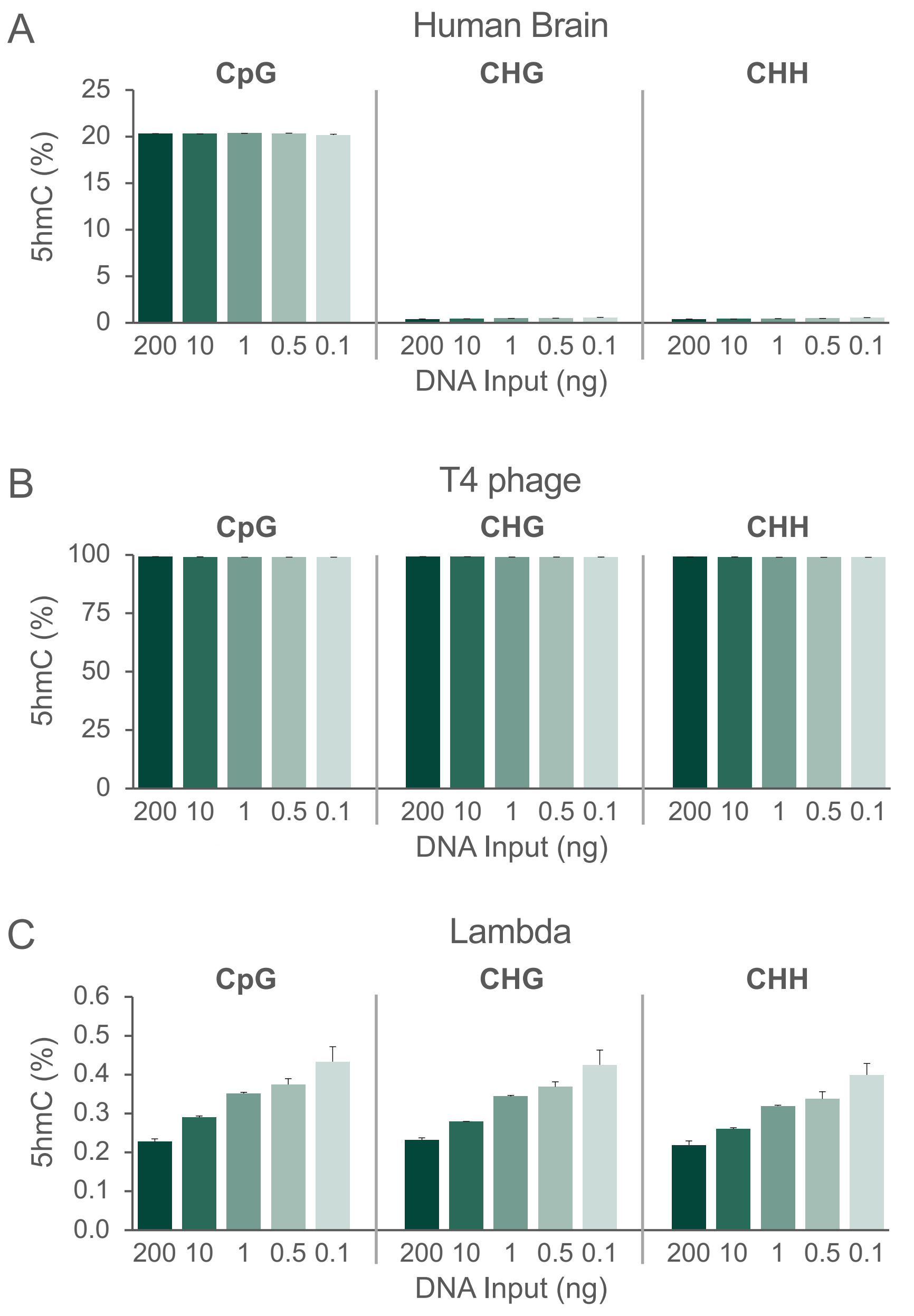
E5hmC-seq library profiles and metrics for human brain. E5hmC-seq libraries were made using 200 - 0.1 ng of human brain genomic DNA. (A) Average library yields from four replicate libraries per input. Error bars are +/- standard deviation. (B) Read numbers, mapping, duplication and useable read percentages. Reads were aligned to the human T2T genome. (C) Library insert sizes are consistent across DNA inputs.

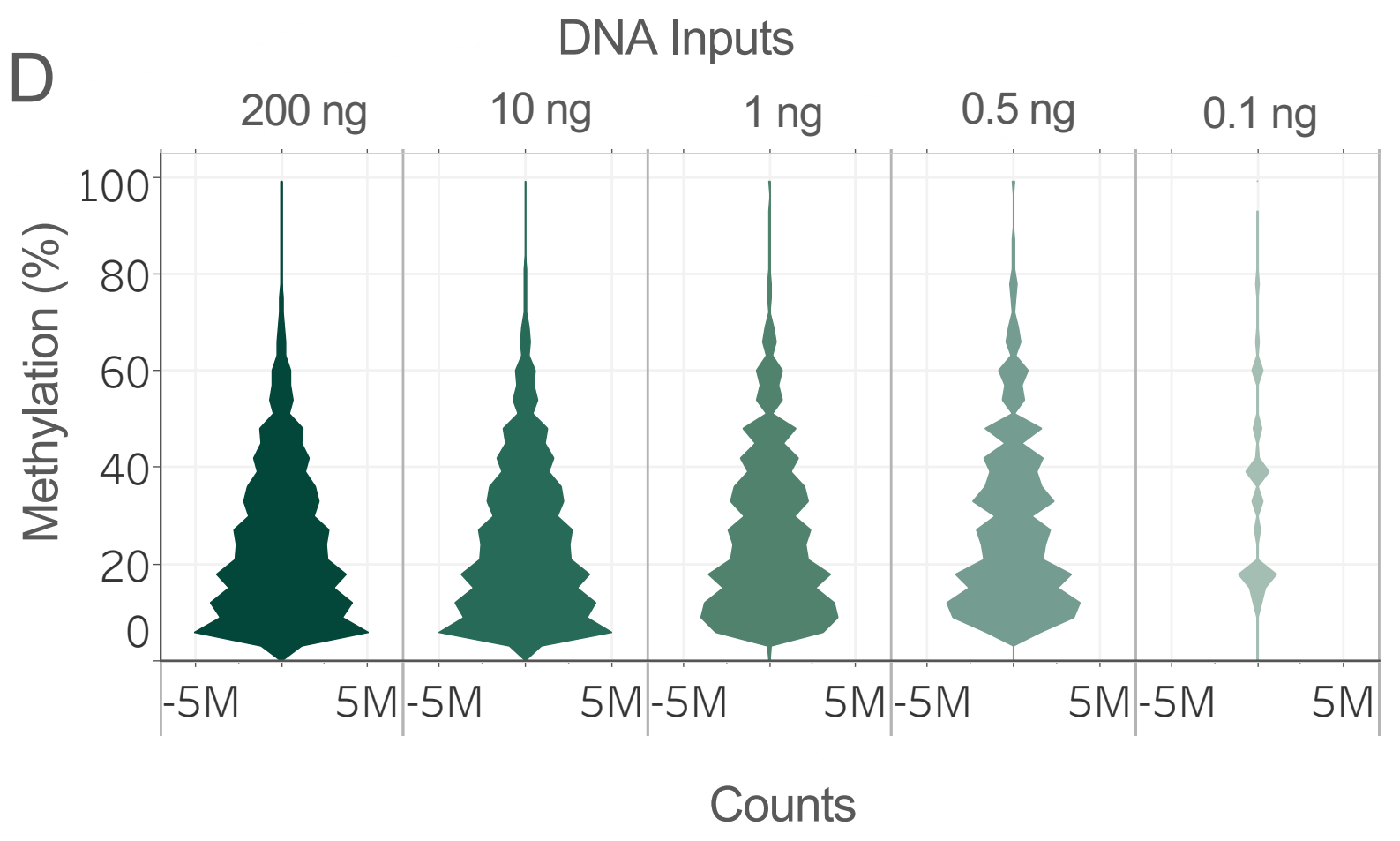


GC coverage was analyzed using Picard and the distribution of normalized coverage across different GC contents of the genome (0 – 100%) was plotted. E5hmC-seq libraries exhibit uniform GC coverage over a range of inputs.

Results

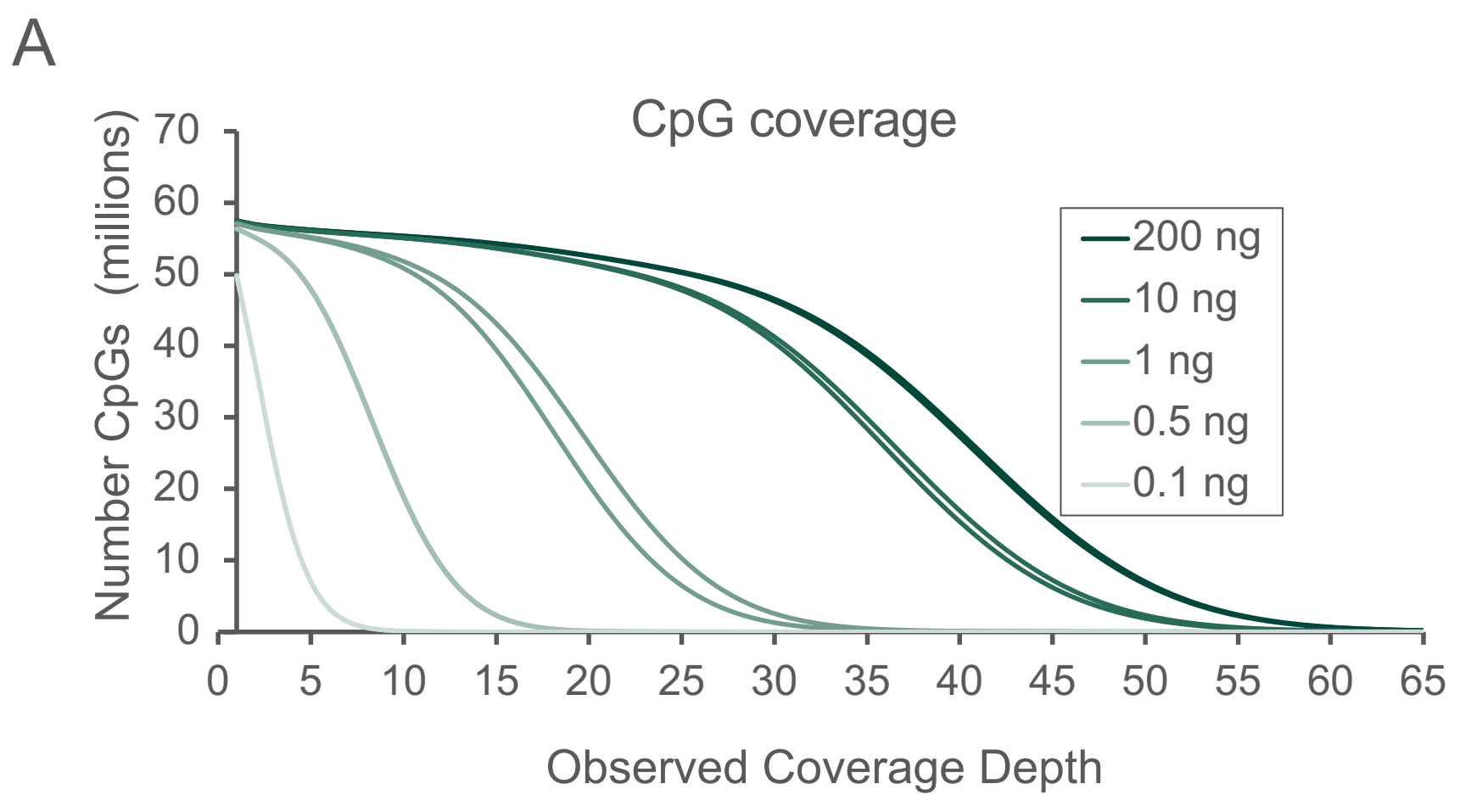
Hydroxymethylation in Human Brain

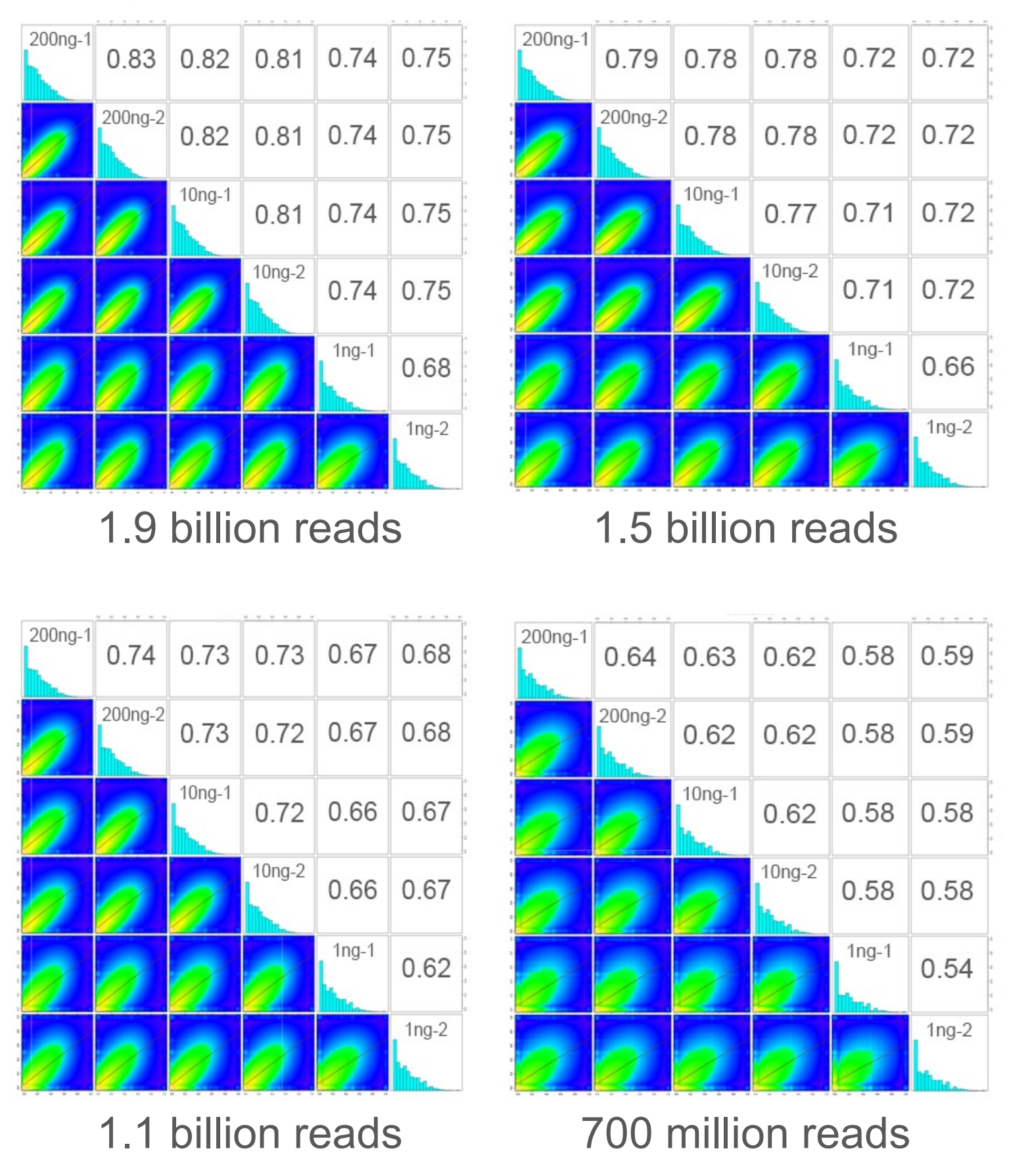




5hmC levels detected in the CpG, CHG and CHH contexts for the 200 ng to 0.1 ng human brain E5hmC-seq libraries. (A) Global 5hmC levels in the CpG context are approximately 20% and 0.4 to 0.6% for CHG and CHH contexts for all of the DNA inputs. (B) 5hmC percentages for lambda control DNA were < 0.5% indicating efficient deamination. (C) 5hmC levels in the fully hydroxymethylated T4 phage control DNA are all ~99.0% indicating efficient glucosylation and protection of 5hmC from deamination. (D) Counts of percentage hydroxymethylation per base in CpG context. Methylation is binned per 3 percentage points. Note that the majority of 5hmC is <60% per base with minimal CpGs at 100% hydroxymethylation. 5hmC at 0% is not shown, but ranges from ~ 20% to 40% of the total CpG for the 200 ng to 0.1 ng inputs respectively.

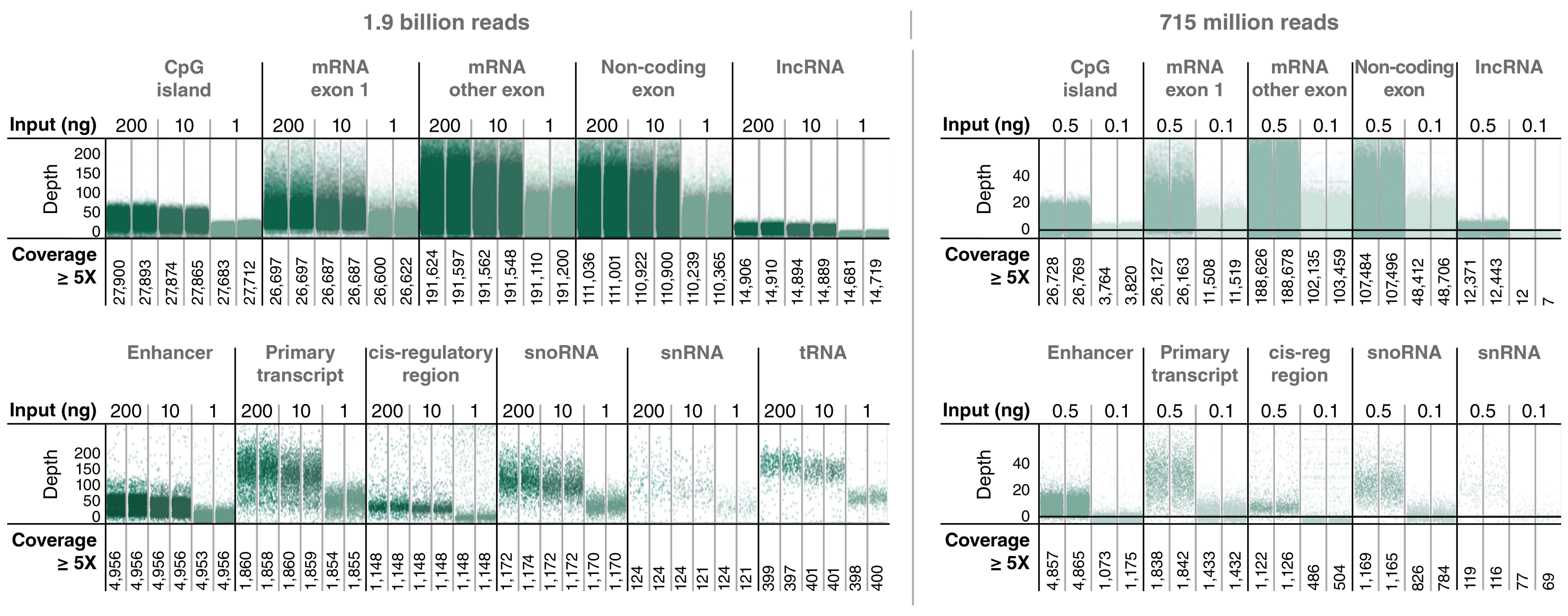
E5hmC-seq Correlation and Coverage Analysis



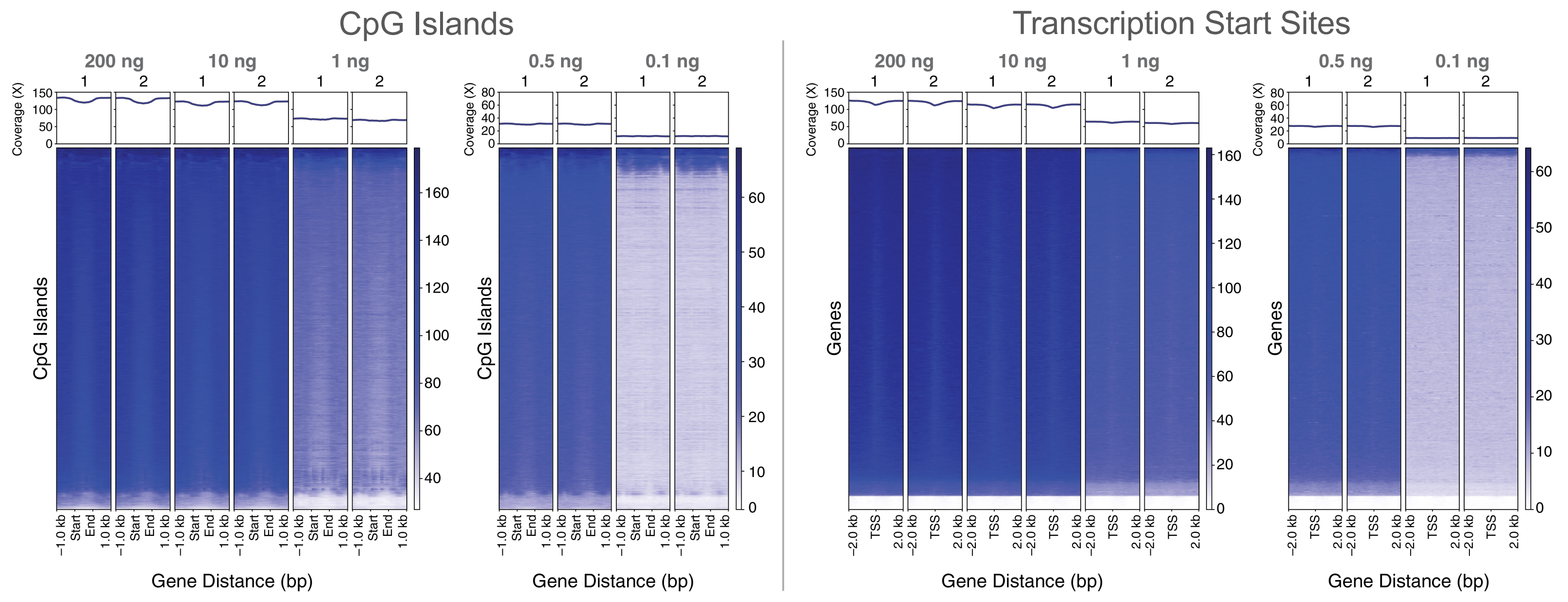


(A) At 1X depth, the 200 ng to 0.5 ng libraries detected 56 million CpGs and the 0.1 ng input detected ~48 million CpGs. Coverage was calculated using 1.9 billion reads for the 200 ng - 1 ng inputs and 715 million reads for the 0.5 and 0.1 ng DNA inputs. (B) Pearson correlations of sequencing reads for 200 ng, 10 ng and 1 ng DNA inputs. Correlations were made using 1.9 billion, 1.5 billion, 1.1 billion and 700 million reads. Approximately 56 million CpGs at 1X coverage were considered for correlation analysis at each read input.

Coverage of Genomic Features



Coverage of genomic features. Left Panel: 1.9 billion reads for the 200 ng to 1 ng inputs. Right Panel: 715 million reads for the 0.5 ng and 0.1 ng DNA inputs. The number of features with coverage ≥ 5X is indicated. Coverage of genomic feature are represented with one point per region with the vertical position representing the average coverage of the feature. Points are staggered horizontally to avoid excess overlapping.



Feature annotations are from NCBI's RefSeq browser. CpG islands were defined based on the UCSC genome browser. Heatmaps generated using deepTools showing the distribution of coverage in 2 kb and 1 kb windows around transcription start sites (TSS) and CpG islands, respectively. Dark blues indicate high coverage and light blue/white indicate little or no coverage. The heatmaps show that E5hmC-seq has even coverage at all DNA inputs across these genomic features.

Conclusions

E5hmC-seq libraries:

- streamlined and user-friendly protocol
- expected insert sizes
- minimal GC bias indicating uniform GC coverage over a range of inputs
- accurate measurements of hydroxymethylation across inputs from 200 ng – 0.1 ng
- E5hmC-seq libraries have high coverage of diverse genomic feature types

E5hmC-seq data subtracted from EM-seq data (5mC and 5hmC) gives a complete representation of 5mC and 5hmC

Acknowledgements

We thank Dora Posfai, Kristen Augulewicz, and Harry Bell for their technical support and members of NEBNext for productive discussions.