NEBNext® Enzymatic 5hmC-seq Kit

NEB #E3350

FIGURE 1: **The NEBNext Enzymatic 5hmC-seq (E5hmC-seq™) workflow**

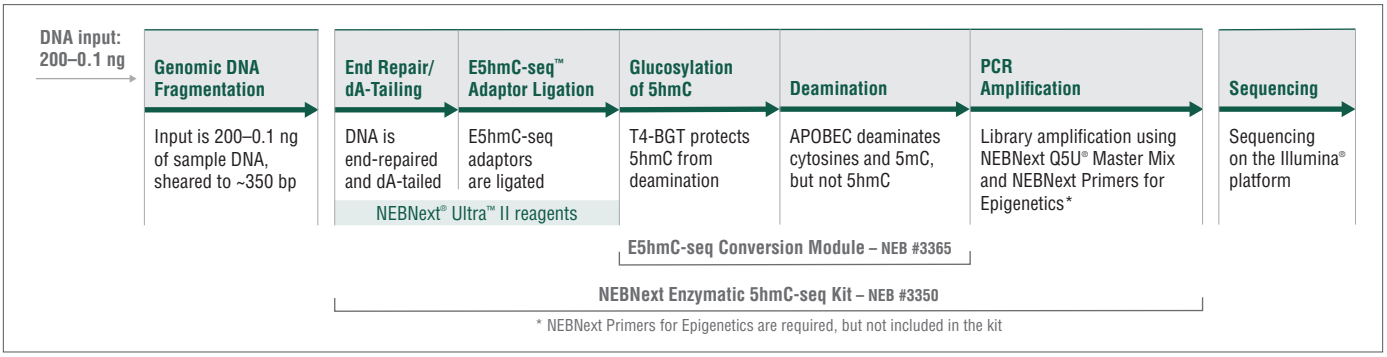
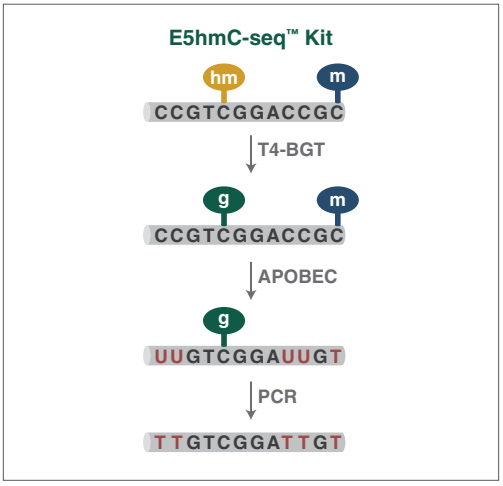
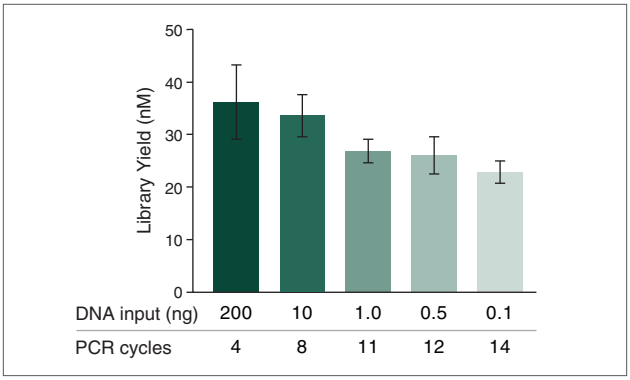


FIGURE 2: **E5hmC-seq conversion method**



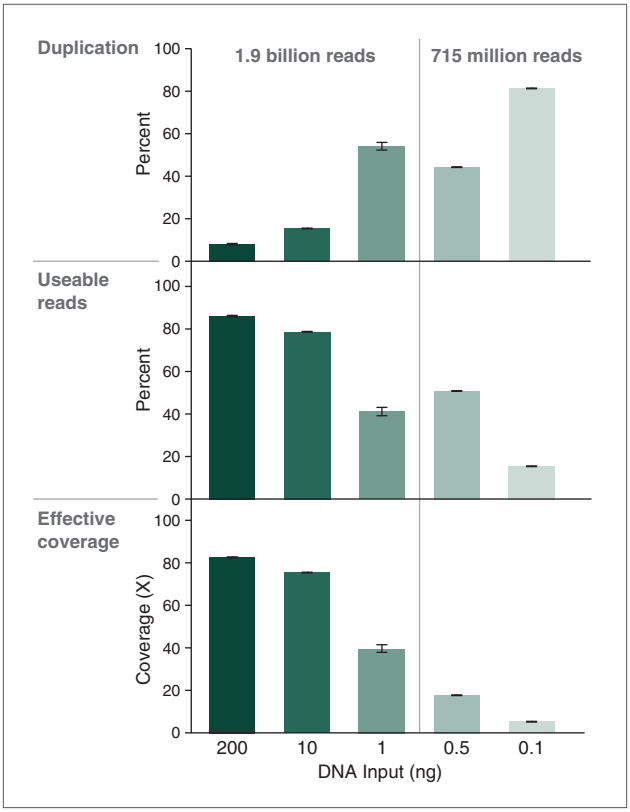
To enable specific 5hmC detection, 5hmC is first glucosylated using T4-BGT. 5mC and unmodified cytosine are then deaminated by APOBEC to thymine and uracil, respectively, while the protected 5hmC is not converted. During Illumina® sequencing, 5hmCs are represented as cytosine, while cytosine and 5mCs are represented as thymine.

FIGURE 3: **E5hmC-seq produces high library yields across a broad input range**



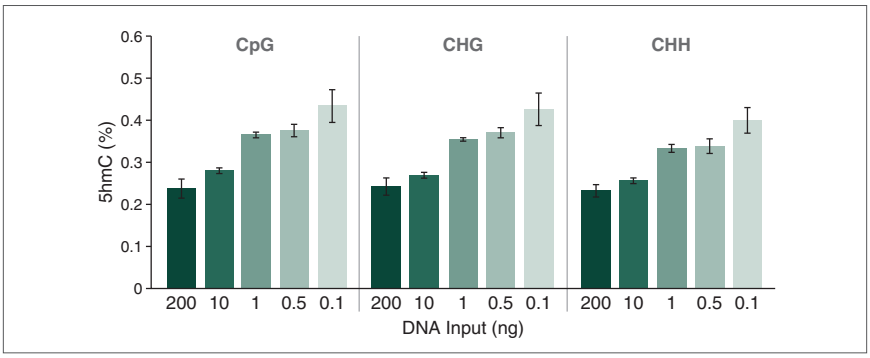
200–0.1 ng of human brain genomic DNA, sheared to 350 bp (Covaris® ME220) was used as input into the E5hmC-seq library protocol, using the number of PCR cycles shown. Library yields were determined using the Agilent® TapeStation® with High Sensitivity D1000 reagents. Values shown are the average of 4 technical replicates and error bars are standard deviation. E5hmC-seq consistently produces high-yield libraries across a wide range of inputs.

FIGURE 4: **E5hmC-seq produces high-quality libraries**



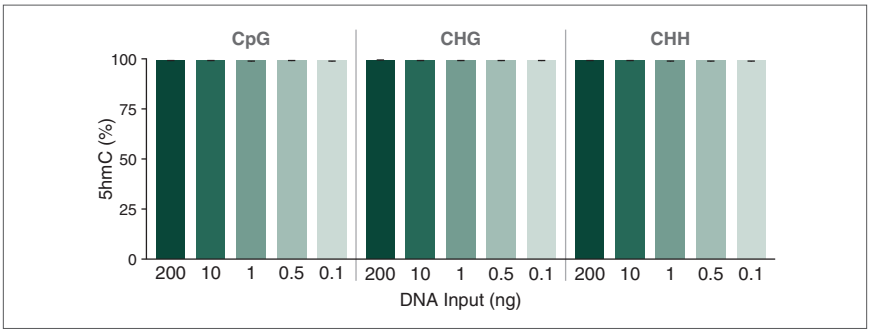
Sequencing metrics for E5hmC-seq libraries using either 1.9 billion (200 ng, 10 ng and 1 ng) or 715 million (0.5 ng and 0.1 ng) 150-base Illumina® reads. Mapping percentage for all libraries was > 99.95% (not shown), using bwa-meth to align to the reference genome (T2T+controls). Duplication: reads marked as duplicate by Picard MarkDuplicates. Usable Reads: the set of Proper-pair, MapQ > 10, primary, nonduplicate reads used in methylation calling (SAMtools view -F 0xF00 -q 10). Effective Coverage: % Usable × theoretical coverage. Theoretical coverage is calculated using the number of bases sequenced/total bases in the T2T reference.

FIGURE 5: **E5hmC-seq libraries have a high deamination efficiency**



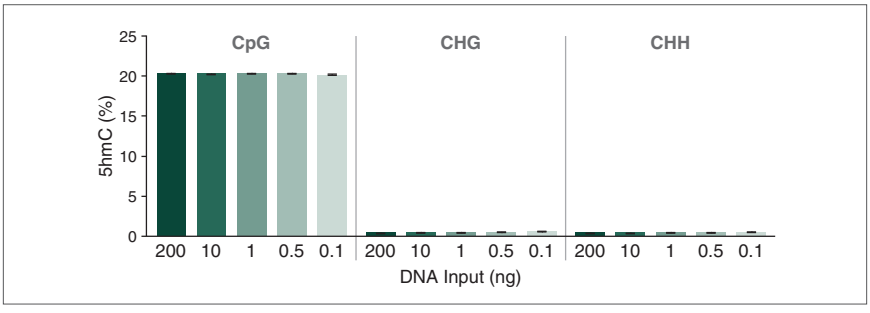
Control unmethylated lambda DNA was spiked in when preparing E5hmC-seq libraries using 200–0.1 ng of human brain DNA, sheared to 350 bp (Covaris® ME220). E5hmC-seq libraries were sequenced on an Illumina® NovaSeq® 6000 (2 x 150 bases). Approximately 1.9 billion (200 ng, 10 ng and 1 ng) or 715 million (0.5 ng and 0.1 ng) reads for each library were aligned to a composite human T2T, lambda and T4 genome using bwa-meth, and methylation information was extracted from the alignments using MethylDackel. Percent 5hmC detected in each library is ≤ 0.5% indicating a deamination efficiency of ≥ 99.5%. Values shown are the average of two technical replicates and error bars show standard deviation.

FIGURE 6: **E5hmC-seq™ libraries have robust protection of 5hmC**



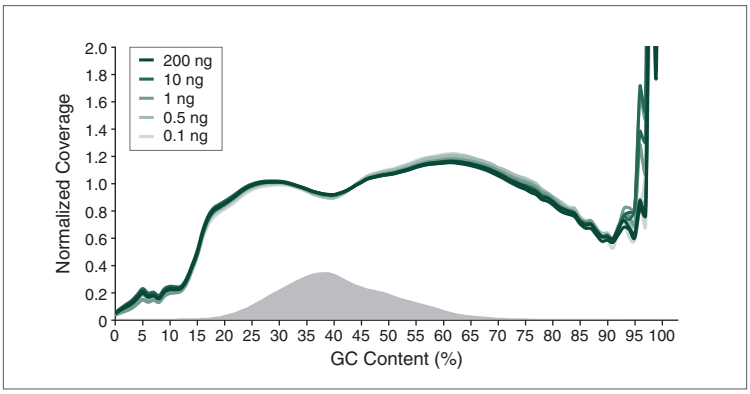
Control T4 DNA that is fully hydroxymethylated at all cytosines was spiked in when preparing E5hmC-seq libraries using 200 ng to 0.1 ng of human brain DNA. DNA was sheared to 350 bp using the Covaris® ME220 instrument, E5hmC-seq libraries were prepared and sequenced on an Illumina® NovaSeq 6000 (2 x 150 bases). Approximately 1.9 billion (200 ng, 10 ng and 1 ng) or 715 million (0.5 ng and 0.1 ng) reads for each library were aligned to a composite human T2T, lambda and T4 reference genome using bwa-meth. Methylation information was extracted from the alignments using MethylDackel. Values shown are the average of two technical replicates and error bars show standard deviation. Percent 5hmC detected for control T4 DNA in the CpG, CHG and CHH contexts was ≥ 98.9%.

FIGURE 7: **5hmC detected by E5hmC-seq in human brain gDNA is consistent across inputs**



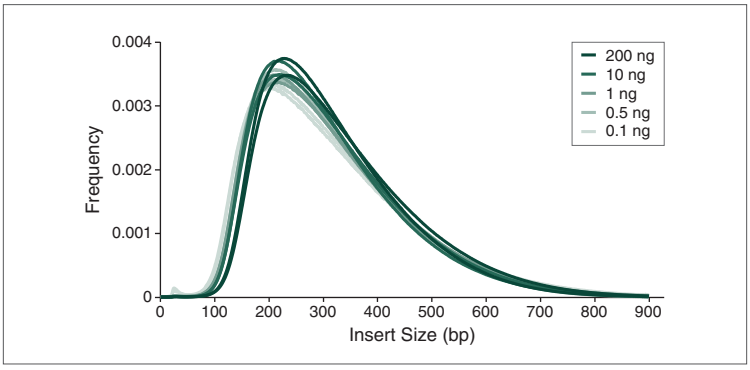
200–0.1 ng of human brain genomic DNA was sheared to 350 bp (Covaris® ME220) and E5hmC-seq™ libraries were prepared and sequenced on an Illumina NovaSeq 6000 (2 x 150 bases). Approximately 1.9 billion (200 ng, 10 ng and 1 ng) or 715 million (0.5 ng and 0.1 ng) reads for each library were aligned to a composite human T2T, lambda and T4 reference genome using bwa-meth, and methylation information was extracted from the alignments using MethylDackel. Values shown are the average of two technical replicates and error bars show standard deviation. Detected 5hmC levels are similar between all inputs in the CpG, CHH and CHG contexts.

FIGURE 8: **E5hmC-seq provides even GC coverage**



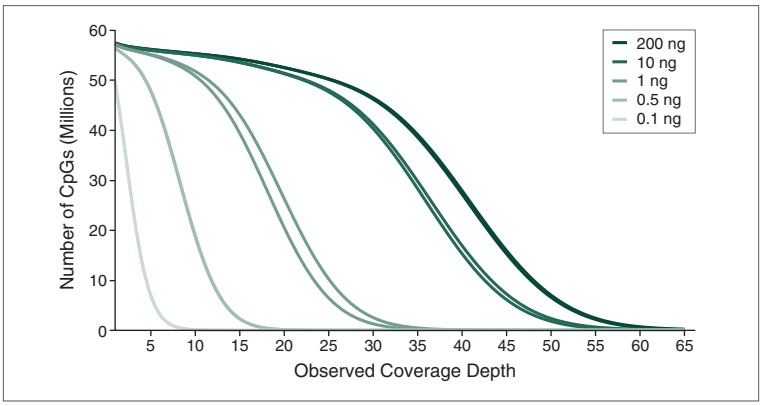
200–0.1 ng of human brain genomic DNA was sheared to 350 bp (Covaris ME220) and E5hmC-seq libraries were prepared and sequenced on an Illumina NovaSeq 6000 (2 x 150 bases). Approximately 1.9 billion (200 ng, 10 ng and 1 ng) or 715 million (0.5 ng and 0.1 ng) reads for each library were aligned to a composite human T2T, lambda and T4 reference genome using bwa-meth. GC coverage was analyzed using Picard and the distribution of normalized coverage across different GC contents of the genome (0–100%) was plotted. The GC content distribution of the human T2T genome is plotted as a histogram. E5hmC-seq libraries have uniform GC coverage across the full input range.

FIGURE 9: **E5hmC-seq libraries have a uniform insert size distribution**



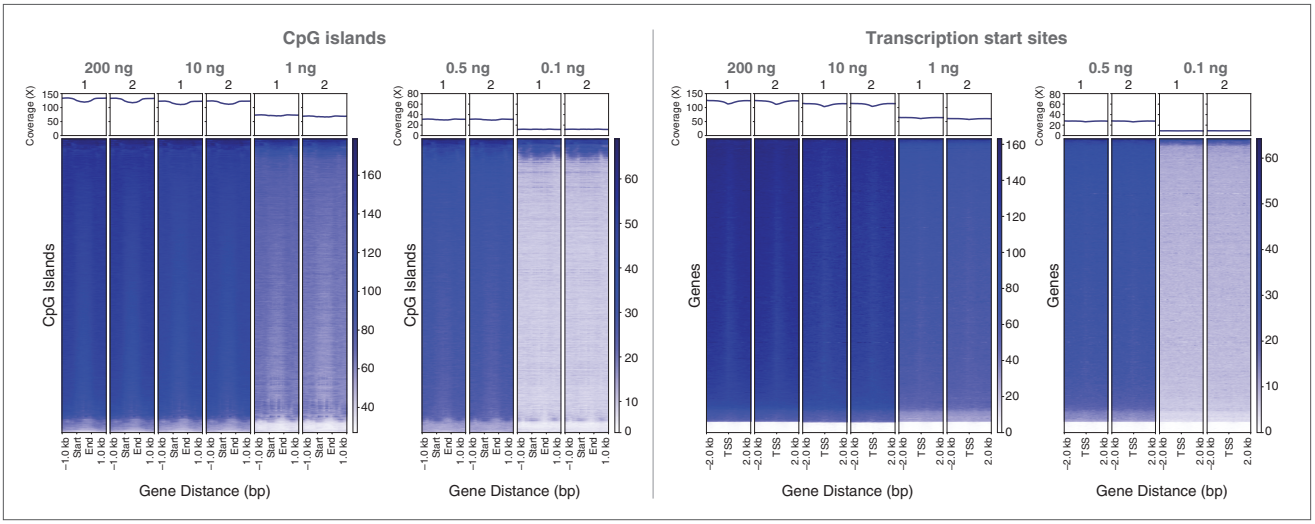
200–0.1 ng of human brain genomic DNA was sheared to 350 bp (Covaris ME220) and E5hmC-seq libraries were prepared and sequenced on an Illumina NovaSeq 6000 (2 x 150 bases). Approximately 1.9 billion (200 ng, 10 ng and 1 ng) or 715 million (0.5 ng and 0.1 ng) reads for each library were aligned to a composite human T2T, lambda and T4 reference genome using bwa-meth. Library insert sizes were determined using Picard and the normalized frequency of each insert size was plotted. E5hmC-seq libraries have consistent insert sizes regardless of the input DNA amount.

FIGURE 10: **E5hmC-seq exhibits high CpG coverage across a range of inputs**



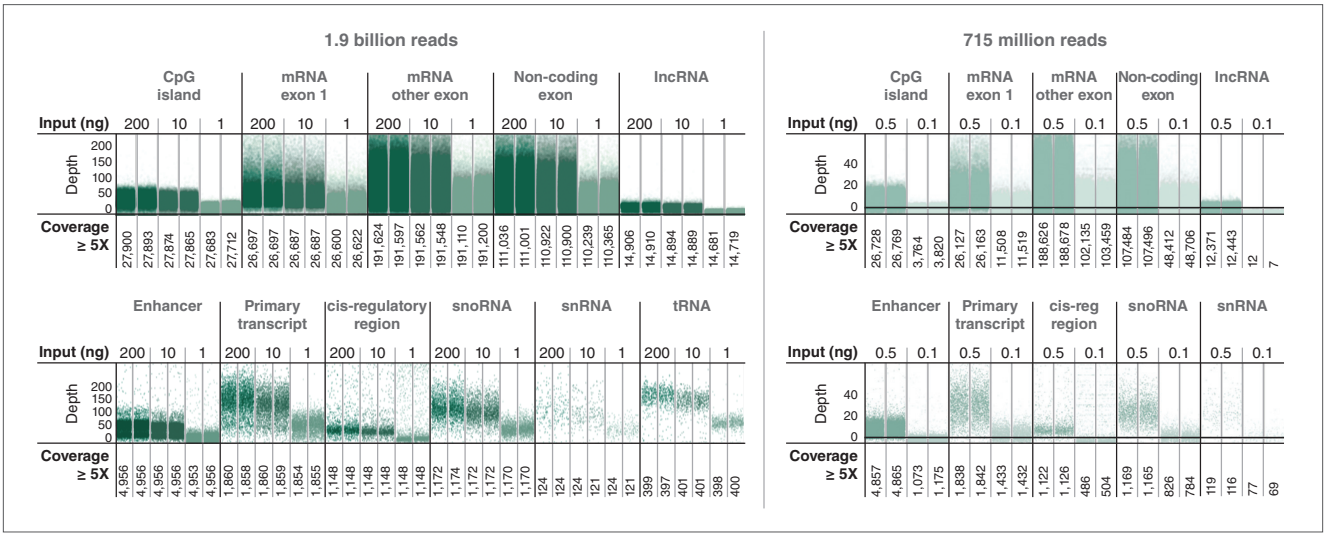
200–0.1 ng of human brain genomic DNA was sheared to 350 bp (Covaris ME220) and E5hmC-seq libraries were prepared and sequenced on an Illumina NovaSeq 6000 (2 x 150 bases). Approximately 1.9 billion (200 ng, 10 ng and 1 ng) or 715 million (0.5 ng and 0.1 ng) reads for each library were aligned to a composite human T2T, lambda and T4 reference genome using bwa-meth. Methylation information was extracted from the alignments using MethylDackel and reported in methylkit format across all three contexts. Using the CpG specific file a cumulative coverage plot was generated for CpG sites covered using E5hmC-seq libraries across all inputs. The T2T genome covers a maximum of 67.8 million CpGs when the top and bottom strands are counted independently. E5hmC-seq covered over 56 million CpG sites for 0.5 ng to 200 ng inputs and roughly 48 million CpG sites for 0.1 ng input libraries.

FIGURE 11: **E5hmC-seq libraries have even coverage across genomic features**



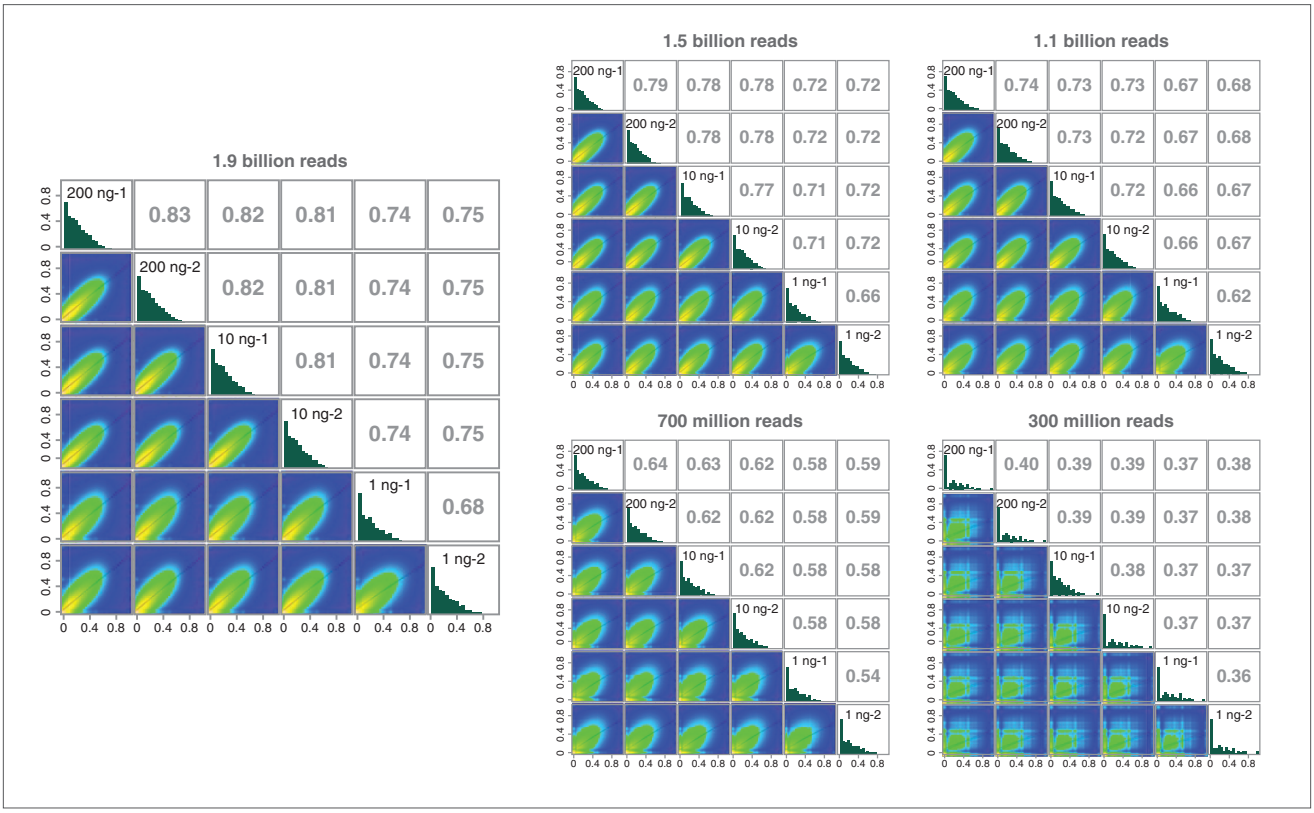
200 ng, 10 ng and 1 ng input libraries were sequenced to a depth of 1.9 billion total 150 base reads while 0.5 ng and 0.1 ng libraries were sequenced to 715 million total 150 base reads. Heatmaps were generated by deepTools, showing coverage across 1 kb windows around CpG islands (left) and coverage across 2 kb windows around transcription start sites (TSS) (right). E5hmC-seq libraries have even coverage across genomic features regardless of DNA input amount.

FIGURE 12: **E5hmC-seq provides consistent coverage of diverse genomics feature types across inputs**



E5hmC-seq libraries were prepared from 200–0.1 ng of human brain DNA. 200 ng, 10 ng and 1 ng input libraries were sequenced to a depth of 1.9 billion total 150-base reads, and 0.5 ng and 0.1 ng libraries were sequenced to 715 million total 150-base reads. The number of features with coverage greater than 5X is indicated below each plot. Coverage of genomic feature types 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.

FIGURE 13: **E5hmC-seq libraries are well correlated at higher sequencing depths**



200 ng, 10 ng and 1 ng E5hmC-seq libraries were sequenced to a depth of 1.9 billion total 150-base reads and correlations were plotted using methylKit at 1X minimum coverage (~ 56.5 million CpGs were common to all libraries). Correlations for the 200 ng and 10 ng input libraries were ≥ 0.81 between replicates. E5hmC-seq libraries for 200 ng, 10 ng and 1 ng were progressively downsampled to ~1.5 billion, 1.1 billion, 700 million and 300 million total reads, and correlation analysis was performed. We observed lower correlations across inputs at 300 million and 700 million reads compared to correlations using 1.1, 1.5 and 1.9 billion reads. This demonstrates the need for deeper sequencing of E5hmC-seq libraries due to the lower abundance of 5hmC signal in the sample.

New England Biolabs, Inc., 240 County Road, Ipswich, MA 01938-2723 Telephone: (978) 927-5054 Toll Free: (USA Orders) 1-800-632-5227 (USA Tech) 1-800-632-7799 Fax: (978) 921-1350 e-mail: info@neb.com