ZYMO RESEARCH
The Beauty of Science is to Make Things Simple

Genome-wide human brain DNA 5-hmC profiling using a novel sequence- and strand-specific method

Xueguang Sun[1], Adam Petterson[1], Tzu Hung Chung[1], Xi Yu Jia[1], and Pu Zhang[2]
[1]Zymo Research Corporation, Irvine, CA; [2]Yale University School of Medicine, New Haven, CT

Introduction

5-Hydroxymethylcytosine (5-hmC) is an epigenetic hallmark which has recently become central in mapping and sequencing work. While the exact function of this base is not fully understood, it is likely to regulate gene expression as a member of active DNA demethylation pathways. The levels of 5-hmC in genomic DNA vary significantly depending on the cell type, though the highest levels are found in cells of the central nervous system (CNS): These findings suggest importance of 5-hmC in gene regulation within the CNS. While several methods have been developed to profile 5-hmC at genomic scale, most are enrichment-based, utilize large amounts of genomic DNA input, and have relatively low resolution. Although efforts have been made to detect 5hmC at single-site resolution, the methods described to date still require several micrograms of DNA, require parallel or subtractive sequencing, and employ successive chemical treatments that degrade the DNA and hinder sequencing. By combining modification-sensitive restriction enzymes with massively parallel (“next-generation”) sequencing approaches, we developed a novel Reduced Representation Hydroxymethylation Profiling (RRHP) method for genome-wide 5-hmC mapping at single-site resolution from low (100 ng) DNA inputs. Importantly, the method can detect strand polarity of 5-hmC modifications, and also enables the direct identification of single nucleotide polymorphisms (SNPs) within sequencing reads. Due to the fragmentation approach, data can be directly compared with single-base DNA methylation data from Reduced Representation Bisulfite Sequencing (RRBS). Human brain 5-hmC mapping generated with this method, combined with DNA methylation profiling data, indicates unique distributions of 5-hmC modification: We confirm that several important neuronal loci, such as BDNF, NLGN2, CES1, and TAF1, demonstrate extensive 5-hmC modification. This new method of detection and mapping is a powerful tool in enhancing our understanding of the interplay of genetic and epigenetic regulations in neurobiology and other diverse biological fields.

Methodology

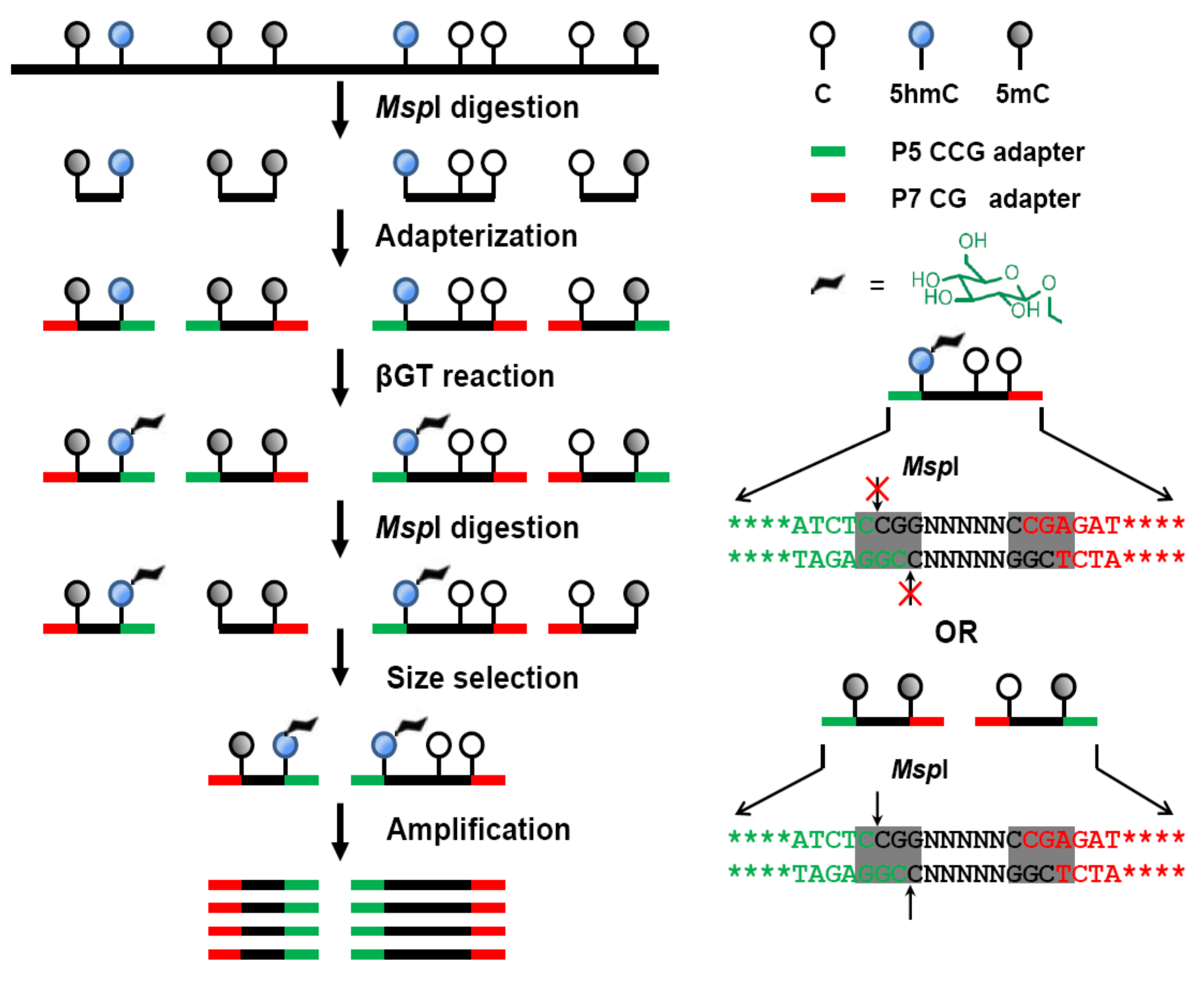
MspI digestion
Adapterization
βGT reaction
MspI digestion
Size selection
Amplification
C
5hmC
5mC
P5 CCG adapter
P7 CG adapter
=
OH
HO
HO
O
O
OH
MspI
****ATCTCCGGNNNNNCCGAGAT****
****TAGAGGCCNNNNNGGCTCTA****
OR
MspI
****ATCTCCGGNNNNNCCGAGAT****
****TAGAGGCCNNNNNGGCTCTA****

Figure 1. Schematic overview of the Reduced Representation Hydroxymethylation Profiling (RRHP) system. The assay exploits ß-glucosyltransferase (ß-GT) to selectively label 5-hmC positions at adapter junctions, thus preventing digestion of the adapter away from the fragment. Fragments lacking 5-hmC at the junction will not be labeled and the adapter can be digested away. Only fragments with intact adapters on both sides will be amplified for hybridization and sequencing.

Library amplification qualities and characteristics

We prepared six libraries from the same male cerebellum genomic source. Two libraries were prepared as negative controls, two libraries were prepared as replicates from 500 ng gDNA, and one library was prepared from 100 ng gDNA. We also prepared a library identical to the 500 ng replicates with the final digestion performed with HpaII instead of MspI: Digestion with the isoschizomer, which is sensitive to any form of methylation, results in a final library presenting 5-mC as well as 5-hmC modifications at adapter junctions, thus allowing for profiling of the total methylome of the sample.

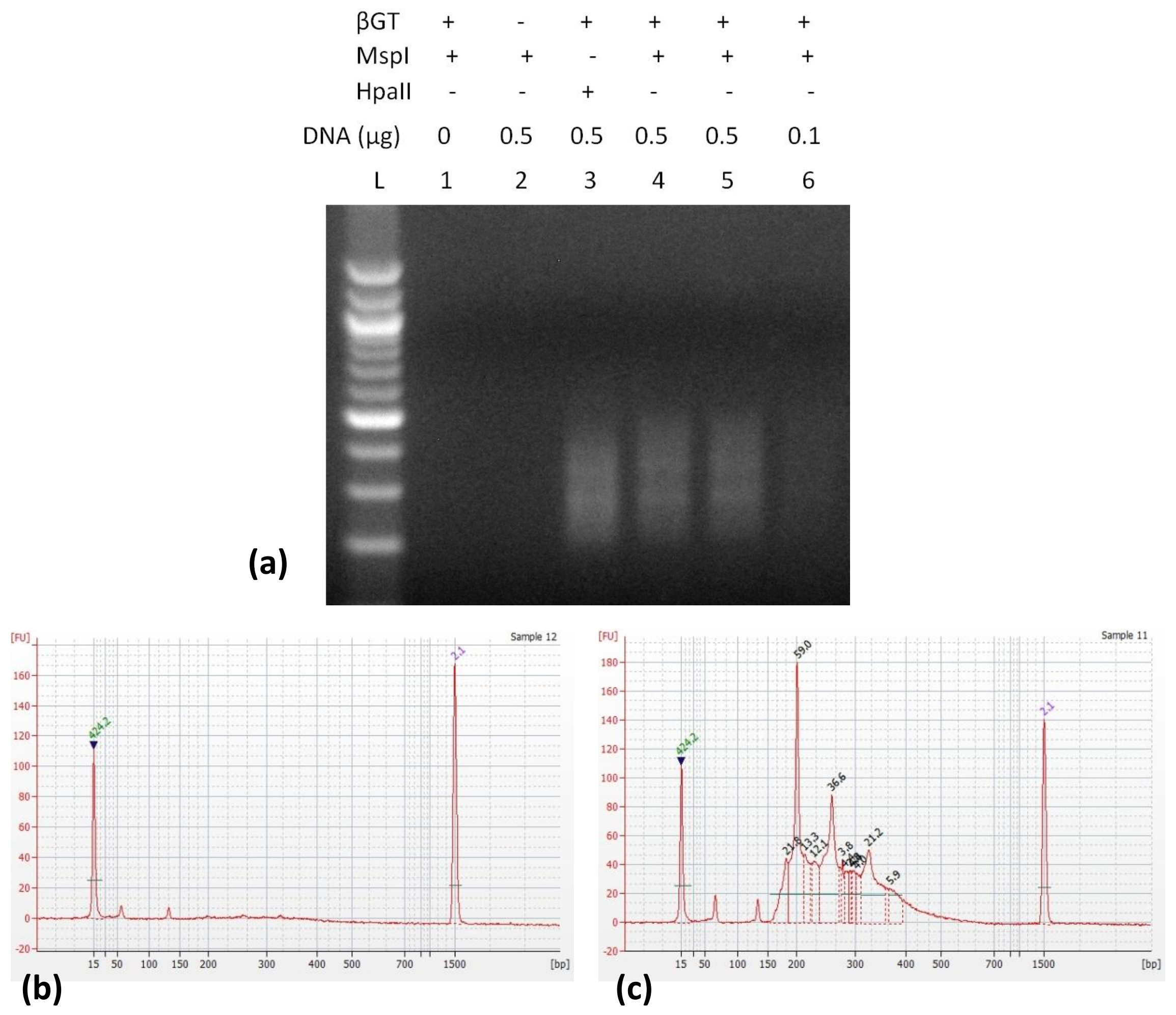

Figure 2. Genomic libraries prepared with the RRHP system demonstrate robust amplification with no detectable background. (a) The six libraries were amplified with standard P5/P7 indexing PCR primers. The library without glucosylation (2) behaves as the library without genomic input (1). Replicate libraries from 500 ng gDNA (4 and 5) demonstrate the correct range of products reproducibly, as does the library from 100 ng gDNA (6). The library digested with HpaII (3) produces higher-intensity amplification products, as the pool includes methylated fragments as well as hydroxymethylated fragments. (b) Negative control library 2 (no glucosylation) was amplified and the resultant product was run on the Bioanalyzer 2100 High Sensitivity DNA Chip, with no detectable product in the sensitive range. (c) Positive library 4 was also amplified and run on the same Bioanalyzer chip, demonstrating robust amplification product in the expected range ca. 150-500 bp.

Sequencing performance

	Simulation 40-430 bp	RRHP-HpaII (0.5 ug)	RRHP-MspI-1 (0.5 ug)	RRHP-MspI-2 (0.5 ug)	RRHP-MspI-3 (0.1 ug)	RRHP (- βgt control)	RRBS
Total reads	1,845,334	23,702,341	23,383,403	9,505,230	19,539,657	5,518	41,066,513
Mapped reads	1,845,025	20,605,538	22,271,499	9,017,016	18,482,119	4,373	15,668,523
Mappability	99.98%	86.93%	95.24%	94.86%	94.59%	79.25%	38.15%
# of tagged reads	1,845,025	17,763,501	21,081,749	8,531,903	17,328,911	3,230	NA
tagged reads %	100%	86%	95%	95%	94%	74%	NA
# of 5hmc sites	1,845,014	1,878,394	1,737,993	1,550,791	1,674,080	3,171	5,330,488

Table 1. Brief overview of several indicators of sequencing output and quality. All samples were sequenced on the Illumina HiSeq 2000 platform. Mappability for all sample preparations remains high, allowing for the identification of more unique sites even with deliberately reduced sequencing depth (RRHP-MspI-2). Samples prepared from low input (RRHP-MspI-3) demonstrate comparable mappability and unique sites to the increased inputs (RRHP-MspI-1, 2). The negative control library without glucosylation demonstrates extremely low reads, low tag rate, and the lowest number of uniquely mapped sites. Standard RRBS was prepared from the same sample for comparison.

Sequencing performance, continued

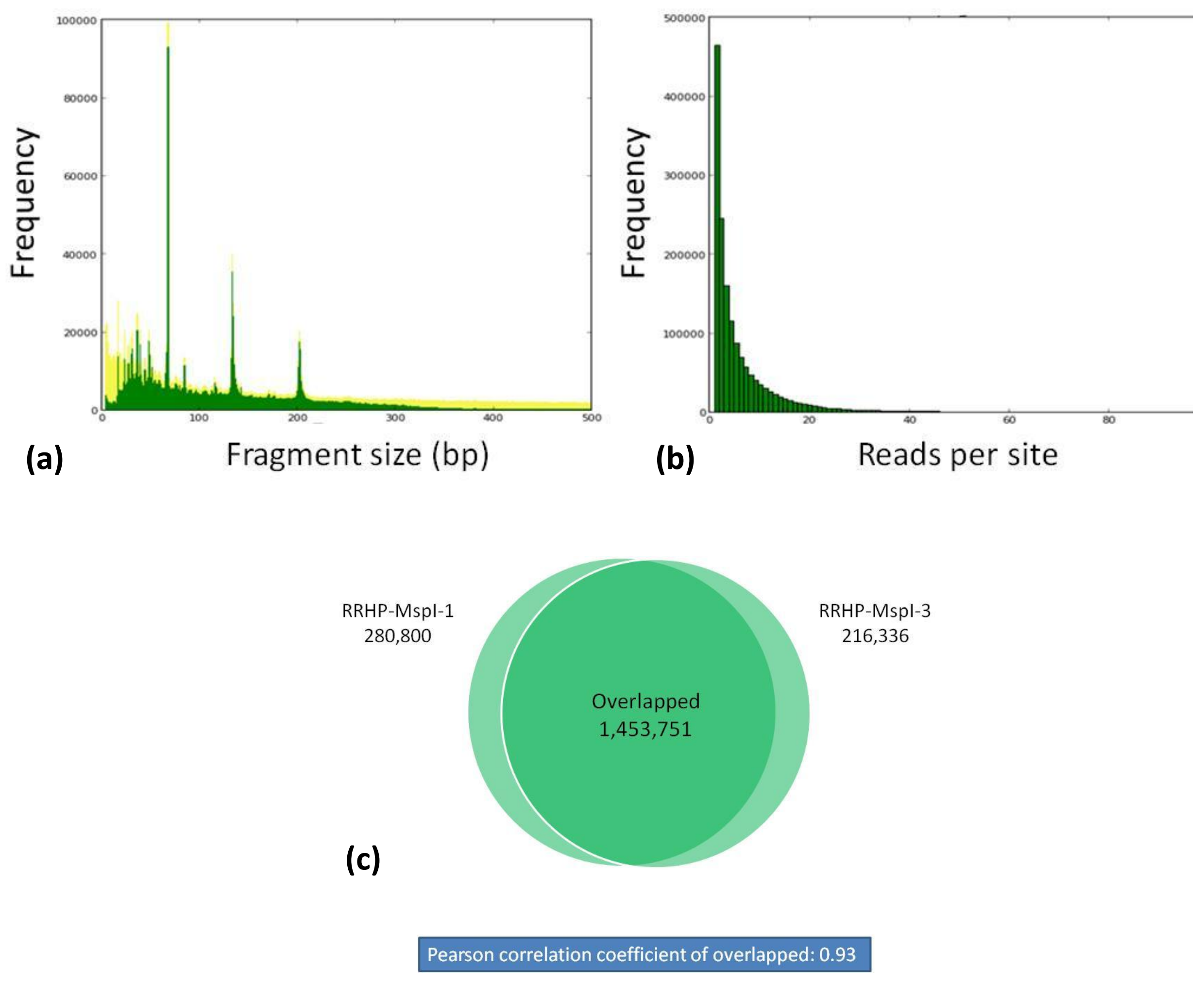

Figure 3. RRHP libraries demonstrate close correlation to *in silico* predictions as well as between one another. (a) The observed fragment size distributions from sequencing (green) are laid over the size distributions predicted from *in silico* digestion (yellow). The sizes mirror those observed in Bioanalyzer traces (shifted up for addition of adapters). (b) Reads per site indicate detection of sites with few intact fragments, suggesting proportionality between 5-hmC density and generated reads. (c) Correlation study between high input (RRHP-MspI-1) and low input (RRHP-MspI-3) libraries indicates the ability of the assay to discover and map the same unique sites from reduced inputs.

Functional analysis of mapped 5-hmC sites

	RRHP-MspI-1 (0.5 ug)		RRHP-MspI-3 (0.1 ug)	
Total	1737993		1674080	
CpG island	171982	10%	144413	9%
Promoter	133938	8%	109537	7%
5' UTR	441174	25%	416594	25%
Coding exon	167857	10%	157064	9%
Intron	1147615	66%	1108226	66%
3' UTR	177478	10%	170934	10%
Bivalent	206811	12%	177243	11%
H3K4me3	988758	57%	762340	46%
H3K27me3	429361	25%	390551	23%
HCP	125982	7%	110205	7%
ICP	34957	2%	32966	2%
LCP	0	0%	0	0%
7X Regulatory potential	683973	39%	644236	38%

Table 2. Breakdown of 5-hmC sites profiled by RRHP into specific annotated genomic elements and binding regions. Mapped and aligned fragments from high- and low-input RRHP libraries from the same sample were analyzed per annotated functional regions of the genome. Consistent annotation rates are observed between the two libraries. Despite the high bias of the assay for CG-dense regions, only 10% of the aligned fragments are mapped back to CpG islands (CGIs), indicating appropriateness for interrogating a wide scope of the genome. Nevertheless, >88% of annotated CGIs demonstrate at least one 5-hmC position from RRHP.

Display of strandedness and SNP positions

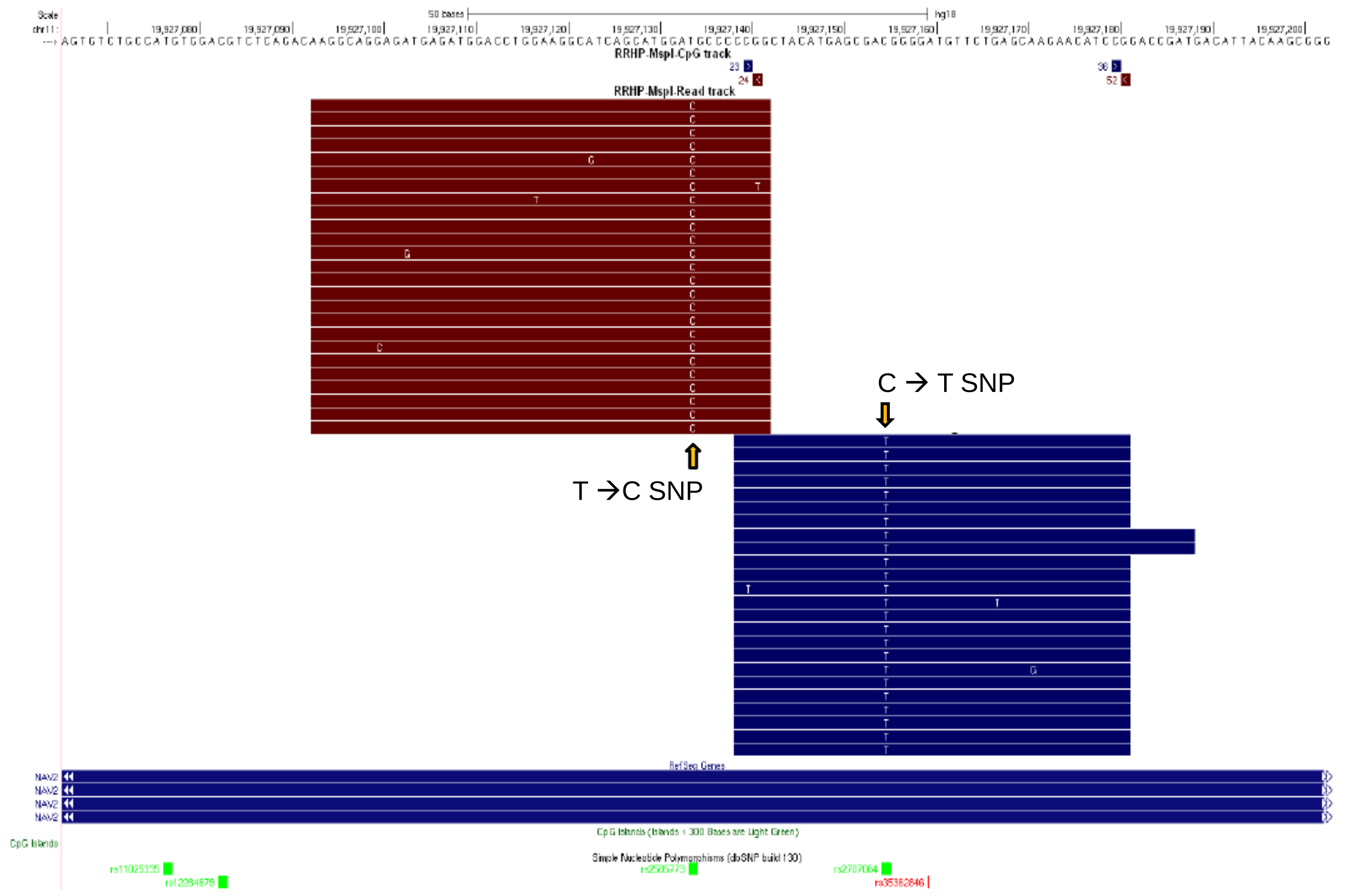

Figure 4. Visualization of RRHP data allows for direct observation of strand-specific 5-hmC distribution, as well as direct SNP identification. Forward (blue) and reverse (red) reads indicate the strand from which a 5-hmC position has been identified (library construction is directionally unbiased). Omission of chemical conversions allow for direct identification of SNPs in the native genomic DNA without the need for additional bioinformatic analysis.

5-hmC patterns in Neuroligin-2

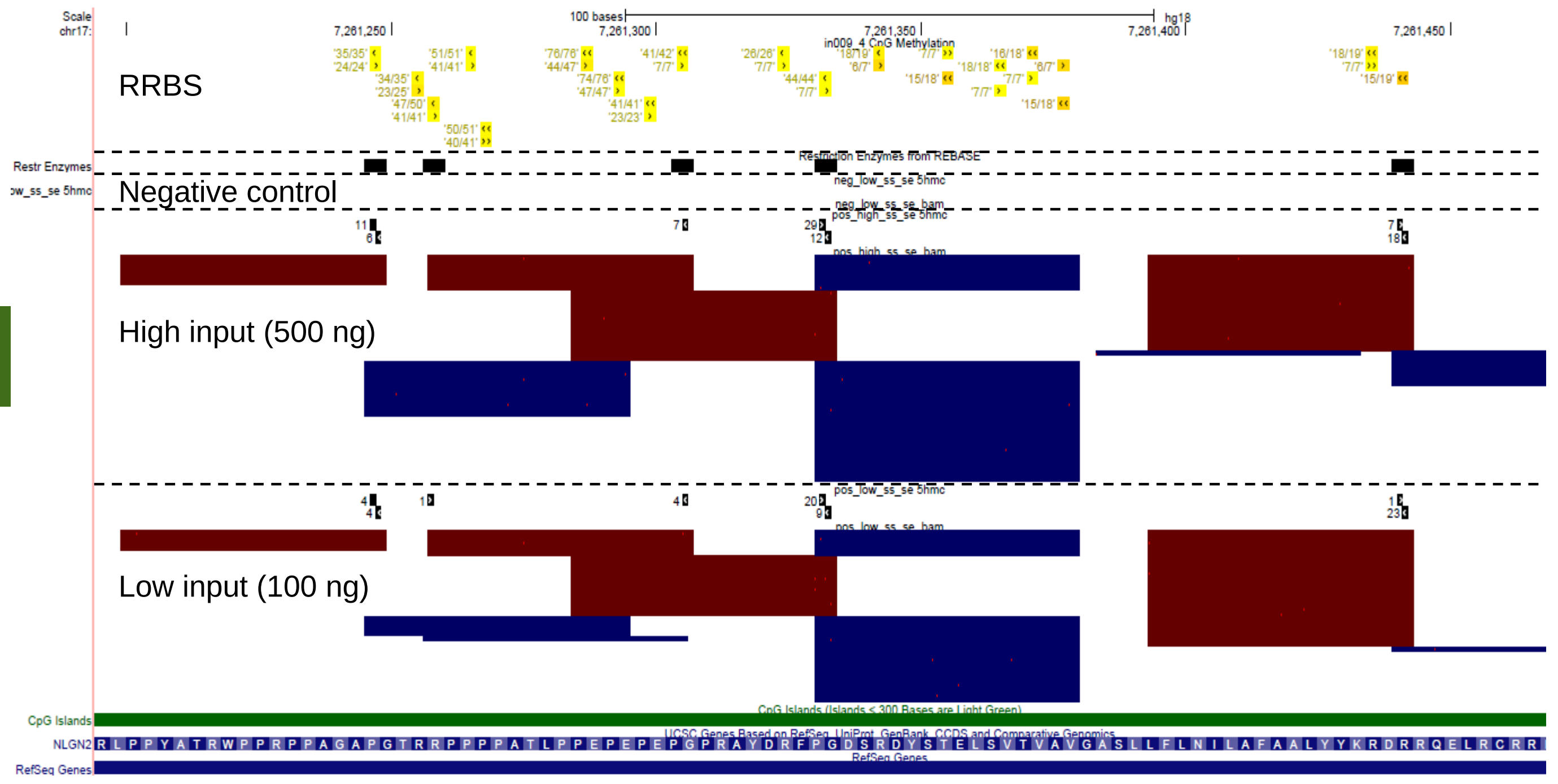

Figure 5. RRHP analysis indicates the distribution of 5-hmC in several sites over NLGN2. Visualization of RRHP reads (blue +/red -) indicates strand distribution across this coding region. Data from the high-input and low-input samples demonstrates high correlation.

Conclusions

Our RRHP method offers a strong, positive-display output for the stringent identification and mapping of 5-hmC across the genome. The fragmentation scheme employed allows for direct comparison with data generated from RRBS processing of the same sample. The method avoids harsh chemical conversion processes, allowing for better quality libraries and higher mapping efficiencies. Importantly, this also allows for the utilization of small input masses without loss in the sites that are uniquely profiled, as well as direct identification of polymorphisms, as all sequenced DNA is native and nonconverted. Due to the directionally unbiased library construction principle, the assay allows for depiction of the strandedness of the 5-hmC mark, which suggests the importance of asymmetry in its distribution. Our new method is directly applicable to any platform which requires adapterization before sequencing, and can be adapted to profile other marks (such as 6-mA) by utilizing alternative enzyme digestion schemes with compatible enzyme sensitivities.