

NGS LIBRARY PREPARATION PRODUCES BALANCED, COMPREHENSIVE METHYLOME COVERAGE FROM LOW INPUT QUANTITIES

ABSTRACT

Next-generation sequencing (NGS) of bisulfite-converted DNA to detect methylation status with per-base resolution is currently restricted by input requirements, requiring at least 50 ng of DNA. This Application Note shows how the unique chemistry of the Accel-NGS[®] Methyl-Seq DNA Library Kit enables the construction of high complexity libraries for:

- Genome-wide methylation analysis from 5 ng of human cell-free DNA (cfDNA).
- Targeted sequencing of 1-100 ng of human genomic DNA using SeqCap[™] Epi CpGiant hybridization capture.
- Direct sequencing of 10 ng human genomic DNA.
- Direct sequencing of 1 ng of genomic DNA from the model organism *Arabidopsis thaliana*.

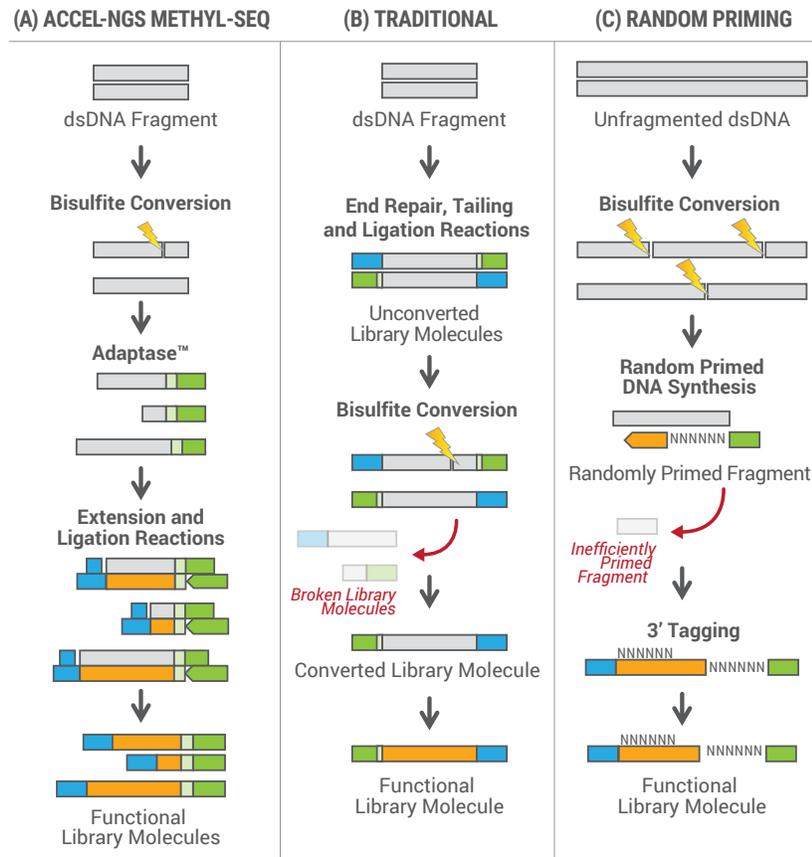
By utilizing the Accel-NGS Methyl-Seq Library Kit, scientists will no longer be forced to sacrifice data when using low quantities of starting material.

INTRODUCTION

Epigenetic research has been impeded by the amount of starting material required to perform whole genome bisulfite sequencing (WGBS) and targeted sequencing of bisulfite-converted DNA samples. Large input requirements preclude sequencing of samples with limited genomic copies, which can arise when tissue collection is constrained or the sample is restricted to a small number of cells to avoid heterogeneity. To address this limitation, Swift Biosciences has reduced the amount of input DNA by up to 100X compared to other commercially available library preparation kits. In addition to lowering input quantity requirements, the efficient repair and adaptation of template DNA by Swift Biosciences' technology also enables improved overall genomic coverage across multiple input quantities, thereby granting researchers more high quality data at any input.

The Accel-NGS Methyl-Seq Kit is based on Swift's Adaptase[™] technology which enables unbiased adapter attachment to single-stranded bisulfite-converted DNA (**Figure 1A**). This preferable workflow eliminates significant (up to 90%) library loss associated with workflows where the NGS library is constructed prior to bisulfite conversion (**Figure 1B**). The template-independent adapter attachment of the Accel-NGS Methyl-Seq Kit also provides significant improvements in methylome coverage uniformity compared to methods that incorporate adapters by random priming of single-stranded DNA (**Figure 1C**).

Figure 1: Workflow for Three Library Preparations Evaluated



For both Accel-NGS Methyl-Seq and random priming kits, bisulfite conversion is performed prior to library construction. With the traditional library kit, bisulfite conversion is performed on the completed library. The lightning bolts represent bisulfite-induced fragmentation, NGS adapters are depicted in green and blue, and non-uracil containing library products are shown in yellow.

METHODS OVERVIEW

Bisulfite conversion was performed using the EZ DNA Methylation-Gold™ Kit¹ according to supplier's instructions. Bisulfite-converted DNA was quantified using the RNA setting on a NanoDrop® instrument, and final libraries were quantified by qPCR. Sequencing data was aligned to its respective reference genome using BSMAP and normalized to a specified number of reads for comparison purposes². Bisulfite-Conversion Control (BCC) or unmethylated Lambda DNA (Promega Cat. No. D1521) was added to each sample to measure bisulfite conversion efficiency. For all samples, conversion efficiency was consistently $\geq 99\%$.

Genome-Wide Hypomethylation

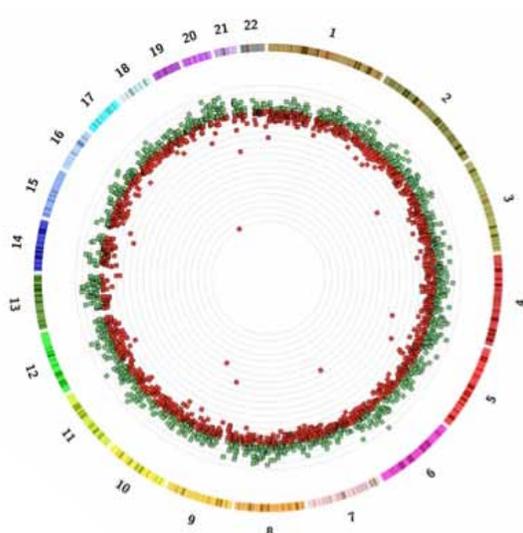
Examining whole genome methylation density (MD) from 5 ng of cfDNA, the Accel-NGS Methyl-Seq Library Kit was used to construct libraries from cfDNA obtained from the blood plasma of 8 cancer patients and 5 healthy controls. Libraries were PCR amplified with 7 cycles of PCR and sequenced on a MiSeq® using V3 chemistry and 2 x 75 read length. Each sample was normalized to 10 million total reads, which

provided sufficient coverage to determine repetitive element MD across the genome. Percent hypomethylation was calculated by comparing the MD of 1 Mb bins to the average MD of those from the 5 healthy controls. Bins were assigned as hypomethylated if the MD was > 3 SD below that of the 5 healthy controls (**Table 1** and **Figure 2**).

Table 1: Genome-wide Hypomethylation in Cancer Samples

SAMPLE	PATHOLOGY	% HYPOMETHYLATION
1	Fallopian tube high-grade papillary serous carcinoma pT3c N1 with 2 nodes involved by micrometastasis	0.4
2	5 cm ovarian "borderline" serous content (cancer-like)	1.1
3	Recurrent pT2, pN0 mammary carcinoma, 2.15 cm	2.4
4	pT1/pN1 pancreatic adenocarcinoma with neoadjuvant therapy	3.6
5	Metastatic colon cancer to the liver (previously treated)	4.4
6	14 cm ovarian "borderline" serous content (cancer-like)	18.0
7	Colon-cancer, non-resectable Adenocarcinoma T4a by imaging	18.0
8	Metastatic colorectal adenocarcinoma with liver metastasis, 2 cm primary	43.4

Percent hypomethylation of 8 cancer samples was calculated by comparing the MD of 1 Mb bins to the average of the 5 healthy control samples. Bins were assigned as hypomethylated if MD was > 3 SD below the average MD.

Figure 2: Genome-wide Hypomethylation of Sample 8

This Circos plot represents the MD of 1 Mb bins across chromosomes 1-22 for Sample 8 (Metastatic colorectal adenocarcinoma with liver metastasis, 2 cm primary).

Comparative Analysis of Targeted Bisulfite Sequencing Using Human DNA

Targeted sequencing of CpG-rich regions was performed in order to compare sequencing metrics from libraries constructed by either the Accel-NGS Methyl-Seq Library Kit or the Roche-Kapa Library Preparation Kit (supplied by Roche-NimbleGen). Human HapMap DNA NA12878 from the Coriell Institute for Medical Research was fragmented to an average 200 bp by Covaris M220. The Accel-NGS Methyl-Seq Kit was used to construct libraries from 100 ng, 10 ng, and 1 ng inputs (all inputs are within specification of the Accel-NGS Methyl-Seq Kit). The Roche-Kapa Library Preparation Kit was used to construct libraries from inputs of 1 µg (within specification) and 10 ng (below specification).

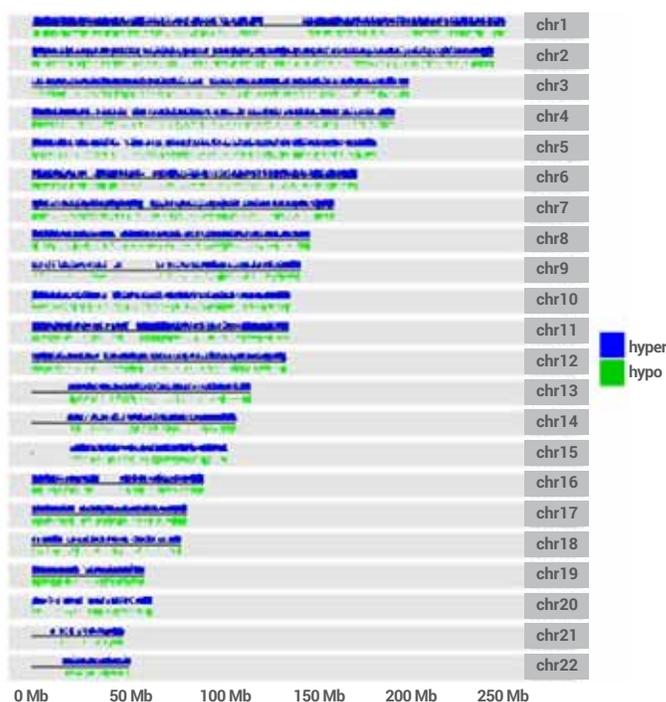
Libraries were amplified with 11, 14, and 17 cycles of PCR (Swift Biosciences) and 12 and 18 cycles of PCR (Roche-Kapa). The SeqCap Epi CpGiant Enrichment Kit (Roche-NimbleGen) was used for hybridization capture of libraries. Each sample was normalized to 80 million total reads for direct comparison of performance metrics. To analyze differentially methylated regions (DMRs), 10 ng libraries from both Swift Biosciences and Roche-Kapa (NA12878) were compared to 10 ng libraries from an H1 ES cell line and sequenced to approximately 40X coverage. Libraries were sequenced on a HiSeq® 2500 using V4 chemistry and 2 x 125 read length (Table 2 and Figure 3).

Table 2: Targeted Methylation Sequencing from 1-100 ng with SeqCap Epi CpGiant

INPUT	METHOD	% ALIGNED	% ON TARGET	% DUPLICATION	MEAN COVERAGE	% COVERED ≥ 2X	% COVERED ≥ 20X	NOT COVERED
100 ng	SWIFT	90	73	6.5	49x	98.6	78.6	0.8
1 µg	Kapa	90	80	9.4	51x	98.6	81.1	0.8
10 ng	SWIFT	91	77	26.0	35x	98.5	71.0	0.8
10 ng	Kapa	87	78	71.0	1x	24.7	0.2	47.7
1 ng	SWIFT	90	73	62.0	8x	93.6	2.3	1.0

Sequencing metrics for libraries created from Roche-Kapa library prep (1 µg, 10 ng) and Swift's Accel-NGS Methyl-Seq Kit (100 ng, 10 ng, 1 ng). All samples were normalized to 80M reads for comparison.

Figure 3: DMRs Called from 10 ng Libraries with Swift's Accel-NGS Methyl-Seq Kit



DMRs were identified from 10 ng libraries from an H1 ES cell line and a B-lymphocyte cell line (NA12878). Libraries created with the Accel-NGS Methyl-Seq Kit identified 294,130 DMRs (shown above). Libraries created with the Roche-Kapa kit identified only 464 DMRs (not shown).

WGBS of Arabidopsis and Human DNA

To evaluate coverage of the model organism *Arabidopsis thaliana*³, the Accel-NGS Methyl-Seq Library Kit, a traditional dsDNA library preparation kit (requiring methylated adapters), and a random priming library preparation kit were used to construct libraries from inputs of 100 ng, 10 ng, or 1 ng. DNA was fragmented to an average 350 bp using a Covaris M220, if specified. Technical replicate libraries were constructed according to suppliers' instructions and PCR amplified accordingly. Swift utilized 4, 7, and 10 cycles, respectively⁴. For traditional and random priming, the recommended cycling for 100 ng was

used plus three additional cycles for each 10X lower input. Amplified libraries were quantified by qPCR and sequenced on a HiSeq 2500 using V4 chemistry and 2 X 125 read length. Each sample was normalized to 30 million total reads for direct comparison of performance metrics (Table 3A). An Accel-NGS Methyl-Seq library was also constructed using 10 ng NA12878 DNA. This sample was processed and analyzed in the same manner as the *Arabidopsis* samples, with the exception that 183.5 million reads were analyzed (Table 3B and Figure 4).

Table 3A: Library Coverage Metrics

	SAMPLE	% READS ALIGNED	GENOME COVERAGE	% DUPLICATE READS	ESTIMATED LIBRARY SIZE (MILLIONS)	RELATIVE LIBRARY SIZE	% CPX MISSING	% CPX COVERED ≥ 10X
100 ng <i>Arabidopsis</i>	Accel-NGS Methyl-Seq	89.6	22X	1.9	714	1.00	0.6	92.2
	Traditional	80.2	21X	2.7	604	0.85	0.6	88.1
	Random Priming	71.4	16X	22.1	48	0.07	7.7	39.4
10 ng <i>Arabidopsis</i>	Accel-NGS Methyl-Seq	87.8	22X	2.7	406	1.00	0.6	90.4
	Traditional	76.7	19X	11.9	70	0.17	0.6	83.9
	Random Priming	71.9	16X	22.2	45	0.11	5.2	45.2
1 ng <i>Arabidopsis</i>	Accel-NGS Methyl-Seq	83.3	18X	18.2	38	1.00	0.6	77.1
	Traditional	80.7	10X	62.3	6	0.16	2.0	17.0
	Random Priming	73.4	12X	46.1	12	0.31	6.6	31.3

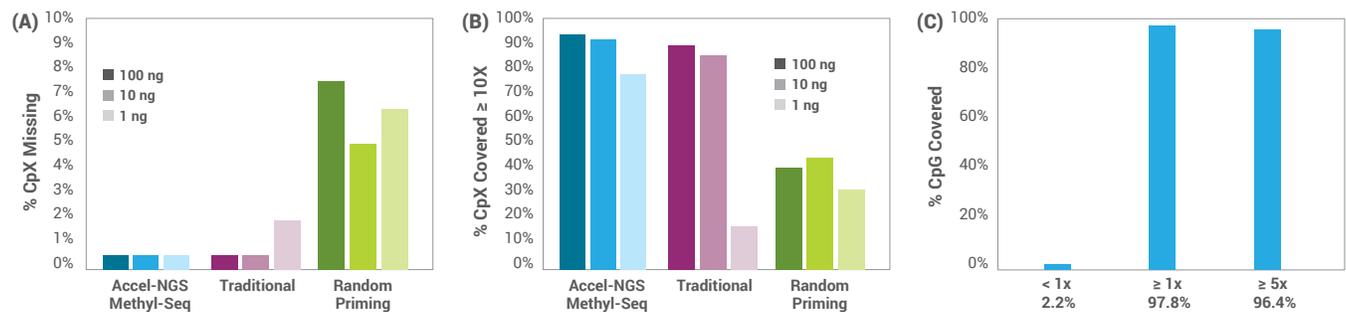
Each *Arabidopsis* file was normalized to 30.2 million reads and data reported as an average of duplicate bisulfite-converted samples. The relative library size is the ratio of alternative methods to the Accel-NGS Methyl-Seq library size.

Table 3B: Library Coverage Metrics

	SAMPLE	% READS ALIGNED	GENOME COVERAGE	% DUPLICATE READS	ESTIMATED LIBRARY SIZE (MILLIONS)
10 ng Human	Accel-NGS Methyl-Seq	86.4	8.9X	7.9	1,393

To assess coverage for the human genome, an Accel-NGS Methyl-Seq library was sequenced using HapMap NA12878 DNA. The human sequencing data was normalized to 183.5 million reads.

Figure 4: Coverage of Unique CpX and CpG Dinucleotide Sequences



(A, B) To assess methylome-specific coverage, 17.5 million unique CpX (CpG + CpH; where H is A, T, or C) dinucleotides were assessed from the *Arabidopsis* TAIR10 reference, including both completeness of coverage (A) and uniformity of coverage (B). (C) To assess methylome-specific coverage of the human genome, 28.1 million unique CpG dinucleotides were assessed for both completeness of coverage and coverage uniformity at 1X and 5X (where the average depth was 8.9X).

RESULTS

Determination of Genome-Wide Hypomethylation from 5 ng of cfDNA

Bisulfite sequencing of plasma cfDNA has recently proven useful for non-invasive detection and monitoring of cancer burden⁵. Detection of cancer-associated genome-wide hypomethylation from patient blood plasma samples has the potential to serve as a high sensitivity and specificity diagnostic test for multiple cancer types. To examine this application with the Accel-NGS Methyl-Seq Library Kit, we performed WGBS on cfDNA from 8 cancer patient samples and 5 normal controls. 5 ng of input cfDNA and 10 million mapped reads per sample provided sufficient coverage to determine genome-wide hypomethylation status in cancer patient samples. This methylation status ranged from 0.4-43.4% hypomethylation for various cancer pathologies (see **Table 1**). Cancer samples exhibiting a low percentage of cfDNA hypomethylation may shed less tumor DNA into the blood plasma or have insufficient tumor burden. Sample 8 demonstrated the highest percentage of hypomethylated DNA, supported by a pathology report that indicated an advanced stage colorectal cancer. When comparing to the normal control samples, a Circos plot for Sample 8 illustrated the specific genomic regions exhibiting hypomethylation (see **Figure 2**). These data support the correlation of plasma cfDNA hypomethylation status and tumor burden. This correlation potentially provides a non-invasive means to detect and monitor tumor burden in patients.

Targeted Methylation Sequencing from 1-100 ng with SeqCap Epi CpGiant

Library enrichment by hybridization capture enables methylation analysis of a targeted region of the genome, reducing sequencing costs compared to WGBS. The SeqCap Epi CpGiant Enrichment Kit (Roche-NimbleGen) captures 80.5 Mb of the human genome, which contains greater than 5.5 million CpG dinucleotides. When compared, sequencing metrics for libraries prepared with the Accel-NGS Methyl-Seq DNA Library Kit outperformed those from the Roche-Kapa library preparation kit supplied by Roche-NimbleGen at every input.

Duplicate reads cripple hybridization capture methylation studies and must be removed prior to analysis. A proficient library preparation minimizes PCR duplicates by adapting input DNA molecules with high efficiency. Roche-Kapa libraries constructed from 1 µg of DNA exhibited duplication rates comparable to those from tenfold less DNA with the Swift Accel-NGS Methyl-Seq Kit (9.4% and 6.5%) (see **Table 2**). Similarly, Roche-Kapa libraries constructed from 10 ng of DNA demonstrated duplication rates that were comparable to those from Swift 1 ng libraries (71% and 62%). Fewer duplicate reads reduce sequencing costs and increase coverage of unique library molecules, potentially increasing the number of samples that can be multiplexed on a single sequencing run.

Coverage metrics from Roche-Kapa libraries highlight a significant deficit in coverage of the target region. This deficit in coverage increases as template amount decreases. When comparing 10 ng input libraries from Swift and Roche-Kapa, mean coverage with the Roche-Kapa library was only 1X, compared to 35X with Swift, from an equivalent read depth. While the Swift library covers 98.5% of the target region at least 2X, the Roche-Kapa library covers only 24.7% of the target 2X. At 20X, the Swift library covers 71% of the target region while the Roche-Kapa library covers a mere 0.2%. Finally, while the Roche-Kapa library only covers 52.3% of the target region at any depth (1X or greater), the Swift library covers 99.2% of the target region at least 1X. Importantly, the coverage metrics from Swift 1 ng libraries outperform those from the Roche-Kapa 10 ng libraries (see **Table 2**). Taken together, these results illustrate that the Swift Biosciences kit yields significantly more usable data than the Roche-Kapa kit.

As epigenetic methylation status has been shown to correlate with transcriptional expression of genes important in cancer and other diseases, identifying CpG sites of the genome exhibiting differential methylation can be informative for both researchers and physicians. To underscore the significance of comprehensive coverage, DMR analysis was compared

between libraries constructed with either Roche-Kapa or Swift Biosciences kits. Libraries constructed from 10 ng of DNA from an H1 ES cell line were compared to the libraries from 10 ng of DNA from a B-lymphocyte cell line (NA12878). These results illustrate the utility of DMR identification from a relatively low input quantity, which diminishes noise from cellular heterogeneity. Swift's Accel-NGS Methyl-Seq Kit called a total of 294,130 DMRs between the two sample types (37,799 hypomethylated and 256,331 hypermethylated) (see **Figure 3**). In contrast, the low coverage exhibited by the Roche-Kapa kit resulted in a total of only 464 DMRs called (not shown).

Whole Genome Sequencing of *Arabidopsis thaliana* from 1 ng Input with Accel-NGS Methyl-Seq Kit

To evaluate WGBS of a small genome, three library preparation technologies were assessed at 100, 10, and 1 ng inputs of *Arabidopsis thaliana* gDNA: Accel-NGS Methyl-Seq Library Kit, a traditional dsDNA library preparation kit (requiring methylated adapters), and a random priming library preparation kit. All input quantities are within the specifications of the Swift Biosciences kit, while the 10 and 1 ng inputs are below the specifications for the traditional and random priming kits. Across all inputs, the Accel-NGS Methyl-Seq Kit exhibited superior performance for the percentage of reads aligned to the genome, the depth of coverage over the entire genome and for CpX (CpG and CpH) sites in particular, the percentage of duplicate reads, and estimated library size.

Maximizing alignment data and minimizing duplicate reads expand the usable data generated from each sequencing run. An increase in usable data enables a specific depth of coverage to be reached with fewer reads, reducing per sample sequencing costs. The Accel-NGS Methyl-Seq Kit improves average read alignment to 87% compared to the traditional kit's 79% and the random priming kit's 72% (see **Table 3A**). With tenfold lower duplication at 100 and 10 ng inputs

than the random priming kit, and 3.4-fold lower duplication at 1 ng than the traditional kit, the Accel-NGS Methyl-Seq Kit demonstrated high efficiency in adapting bisulfite-converted fragments (see **Table 3A**). This gain in sequence data was also apparent in the library complexity as measured by the estimated library size (see **Table 3A**). The Accel-NGS Methyl-Seq Kit demonstrated the highest library complexity at 100 ng, greater than tenfold higher than the random priming kit. While all three kits had reduced estimated library size at 1 ng, the traditional kit was most sensitive as its complexity fell sixfold lower than the Accel-NGS Methyl-Seq Kit. The gain in data output for the Accel-NGS Methyl-Seq Kit was also observed in the average genome coverage (see **Table 3A**).

Accurate cytosine methylation detection in the *Arabidopsis* genome requires comprehensive CpX dinucleotide coverage with balanced read depth across these genomic sequences; ideally, all unique CpX dinucleotides would be represented and covered at equal read depth. Two metrics were used to evaluate completeness and uniformity of CpX coverage – the percent CpX missing (see **Figure 4A**) and the percent CpX covered $\geq 10X$ (see **Figure 4B**). With a mere 0.6% missing, the Accel-NGS Methyl-Seq Kit demonstrated the most comprehensive unique CpX coverage at all inputs (see **Figure 4A**). The traditional kit had reduced representation at 1 ng (2% missing) while the random priming kit showed the highest percentage of CpX missing across all three inputs (6% missing) (see **Figure 4A**). The Accel-NGS Methyl-Seq Kit also demonstrated the most balanced CpX coverage uniformity across all three input quantities, with a remarkable advantage at 1 ng input (77% at 10X coverage compared to 17% and 31%) (see **Figure 4B**).

At all inputs tested, the Accel-NGS Methyl-Seq Kit from Swift Biosciences provided the most comprehensive and uniform methylome coverage. The traditional and random priming kits produced biased representations of the methylome with less usable data.

WGBS of the Human Methylome

While *Arabidopsis* is a useful model organism, it lacks unique elements, such as CpG islands, which can be found in the human genome. These CpG islands and other CpG-rich promoter regions can lack full representation using previously available methods. Data generated from the Accel-NGS Methyl-Seq Kit with human DNA demonstrated comprehensive coverage of unique CpG dinucleotides with only 2.2%

missing from only 9X average sequencing depth. Further, 96.4% of the methylome was covered at least 5X (see **Table 3B**, **Figure 4C**). This uniformity of coverage from low-pass sequencing validates that the Accel-NGS Methyl-Seq Kit performs with great efficiency across different genomes, including those with a complex organization of base composition.

CONCLUSION

The Accel-NGS Methyl-Seq DNA Library Kit preserves high library complexity, delivers low duplication rates, and achieves comprehensive coverage from low sample inputs. This remarkable performance with low input quantities enables analysis of cancer-associated, genome-wide hypomethylation from just 5 ng of cfDNA. For targeted sequencing, the Accel-NGS Methyl-Seq Kit demonstrated superior duplication rates and coverage metrics at all input quantities tested, maximizing the amount of usable sequencing data. Similarly, at all DNA inputs tested, the Accel-NGS Methyl-Seq WGBS libraries yield data displaying increased library complexity and methylome coverage when compared to traditional and random priming kits. These results demonstrate the comprehensive superiority of libraries constructed with the Accel-NGS Methyl-Seq Library Kit, empowering low input methylome analysis.

REFERENCES

1. EZ DNA Methylation-Gold Kit. Zymo Research.
<http://www.zymoresearch.com/downloads/dl/file/id/57/d5005i.pdf>
2. Xi et al. *BMC Bioinformatics*. 2009 Jul 27;10:232. doi:10.1186/1471-2105-10-232.
3. Lister et al. *Cell*. 2008 May 2; 133(3): 523-536. doi:10.1016/j.cell.2008.03.029.
4. Accel-NGS Methyl-Seq. Swift Biosciences.
<http://www.swiftbiosci.com/products/accel-ngs-methyl-seq-dna-library-kit>
5. Chan et al. *Proc Natl Acad Sci*. 2013 Nov 19; 110(47): 18761-18768. doi:10.1073/pnas.1313995110.



Swift Biosciences, Inc.

58 Parkland Plaza, Suite 100 • Ann Arbor, MI 48103 • 734.330.2568 • www.swiftbiosci.com

© 2016, Swift Biosciences, Inc. The Swift logo and Adaptase are trademarks and Accel-NGS is a registered trademark of Swift Biosciences. SeqCap is a trademark of Roche-NimbleGen, Inc. EZ DNA Methylation-Gold is a trademark of Zymo Research. NanoDrop is a registered trademark of Thermo Fisher Scientific Inc. HiSeq and MiSeq are registered trademarks of Illumina, Inc. 16-0831, 05/16