Whole-Genome Bisulfite Sequencing on the HiSeq® 3000/HiSeq 4000 Systems

Guidelines for optimal detection of DNA methylation using next-generation sequencing, including a performance comparison of single-read and paired-end runs.

Introduction

DNA methylation is a critical mechanism of gene regulation, and typically occurs on cytosine bases next to guanine bases (CpGs). Methylation is frequently a signal for chromatin condensation at selected genetic loci, resulting in heritable silencing of gene expression. DNA methylation can be associated with disease and development in mammals.[1] It is also influenced by environmental factors such as diet and exercise.[2,3] Recent breakthroughs in methylation sequencing have mapped approximately 38 million CpG loci in the human genome (the human methylome).[4-6] These new methods lay the groundwork for a new generation of epigenome-wide association studies investigating the link between DNA methylation, development, and disease.

DNA methylation sequencing, known as bisulfite sequencing, relies on bisulfite conversion to alter the sequence of purified DNA selectively. During this process, unmethylated cytosines are converted to uracils, which are then amplified as thymines during the PCR amplification step. Methylated cytosines are not converted to uracil. Whole-genome bisulfite sequencing (WGBS) uses next-generation sequencing (NGS) to analyze bisulfite-converted DNA and map the entire methylome in a single reaction. Using WGBS, it is possible to map differential methylation patterns between tissue types, developmental stages, or disease states.

The increased capacity of the HiSeq 3000/HiSeq 4000 Systems (Figure 1) make them an ideal sequencing solution for data-intensive methods such as WGBS. This application note demonstrates the capability of the HiSeq 4000 System to perform WGBS and provides guidance for optimal sequencing of MethylSeq libraries on these platforms. Equivalent performance is expected on the single flow cell HiSeq 3000 System.

Figure 1: HiSeq 3000 and HiSeq 4000 Systems—Support a broad range of applications generating up to 750 Gb of data with 2.5 Billion reads per flow cell.

Methods

The DNA isolation and bisulfite conversion steps have been described previously in an application note on WGBS using the HiSeq 2500 platform.[7] Genomic DNA was isolated from a member of the CEPH pedigree, NA12878, and sequencing libraries were prepared from bisulfite converted DNA using the TruSeq® DNA Methylation Kit.[8] To increase sequencing quality of bisulfite converted DNA, Illumina PhiX control DNA library[9] was added (spiked-in), so that 20% of the lane consists of an unconverted, balanced sample library. Prepared libraries were diluted to 150 pM and loaded onto the cBot™ DNA Cluster Generation System. After clustering, the flow cell was transferred to the HiSeq 4000 System and sequenced using standard conditions. Libraries generated from the TruSeq DNA Methylation Kit have a mean insert size of approximately 200 bp (standard deviation of 70 bp). Run configurations of either 1 × 150 bp single-end reads or 2 × 75 bp paired-end reads were used for optimal performance.

Table 1: Primary Analysis Sequencing Metrics from 1 x 150 bp and 2 x 75 bp cycle WGBS Sequencing Runs

Read Length (bp)	1 x 150	2 x 75
%PF	79	79
% > /= Q30 Read 1	94	96
% PhiX Read 1	22	22
PhiX Mismatch % Read 1	0.33	0.23
% > /= Q30 Read 2	N/A	76
% PhiX Read 2	N/A	21
PhiX Mismatch % Read 2	N/A	0.66

%PF = the percentage of clusters for this sample that passed filters. Q30 = Phred quality score in which base calls are 99.9% accurate. % PhiX = the percentage of reads attributed to spiked-in PhiX DNA.

For Research Use Only. Not for use in diagnostic procedures.

Primary Data Analysis

As expected, the by-cycle plot of the percent base representation shows a low abundance of cytosine and high abundance of thymine residues (Figure 2A). Changes in base representation towards the end of the sequencing read occur in a subset of clusters with shorter insert lengths, as higher cytosine content and lower thymine content occurs in the adapter sequences that flank the inserts. For the 2 × 75 bp run, the second read shows the expected distribution of low guanine abundance and high adenine abundance from the opposite strand (Figure 2B).

Both the single-read and paired-end runs are of high quality, with an average Q-score > 86% of bases at Q30 or higher across the entire run. The Q30 sequencing metric is related to the base calling error probabilities.[10] When a Q-score of 30 (Q30) is assigned to a base, this is equivalent to the probability of an incorrect base call 1 in 1000 times, or that base call accuracy is 99.9%. Read 2 quality after the paired-end turn shows a drop in Q30 percentages (Figure 3). The same pattern can be seen in the PhiX error rates. Both the single-read and paired-end runs have low Read 1 error rates, while Read 2 of the paired-end run has slightly elevated error rates at the end of the read (Figure 4).

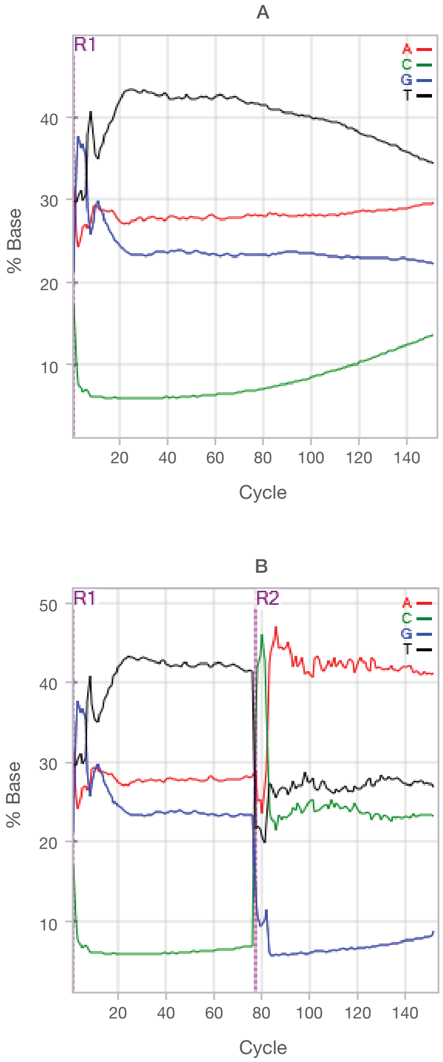

Figure 2: Average percentage of each base present in each cycle—R1 yields a high abundance of thymines and low abundance of cytosines as expected due to bisulfite conversion. R2 yields expected high abundance of adenine and low abundance of guanines from the opposite strand. (A) 1 × 150 bp single-end run, (B) 2 × 75 bp paired-end run on the HiSeq 4000 System.

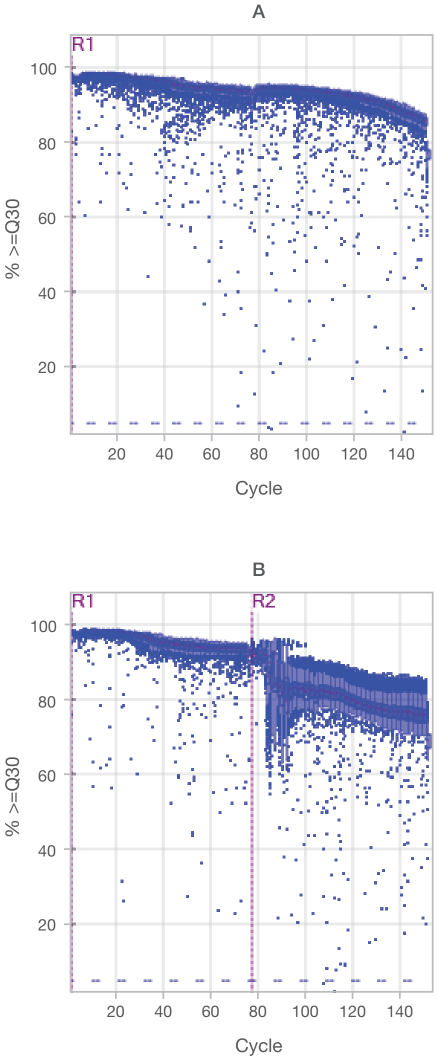

Figure 3: Box plot over all tiles of percentage of base calls greater than or equal to Q30—All clusters passing filters are included from (A) 1 x 150 bp single-end run and (B) 2 x 75 bp paired-end run on the HiSeq 4000 System. Q30 = Phred quality score in which base calls are 99.9% accurate.

For Research Use Only. Not for use in diagnostic procedures.

Primary analysis metrics from the 1 x 150 bp and 2 x 75 bp cycle sequencing runs (Table 1) demonstrate the trade-offs between the 2 run configurations. The overall quality of the 1 x 150 bp run is higher with 94% of bases greater than or equal to Q30, compared to 86% of bases greater than or equal to Q30 for the 2 x 75 bp runs. This is also reflected in the average PhiX error rate of 0.33% for 1 x 150 bp compared to 0.45% average with 2 x 75 bp runs. While performance of the single-read runs is higher than the paired-end runs for these metrics, paired-end runs have the added benefit of enabling identification of duplicates that can be removed in subsequent analyses.

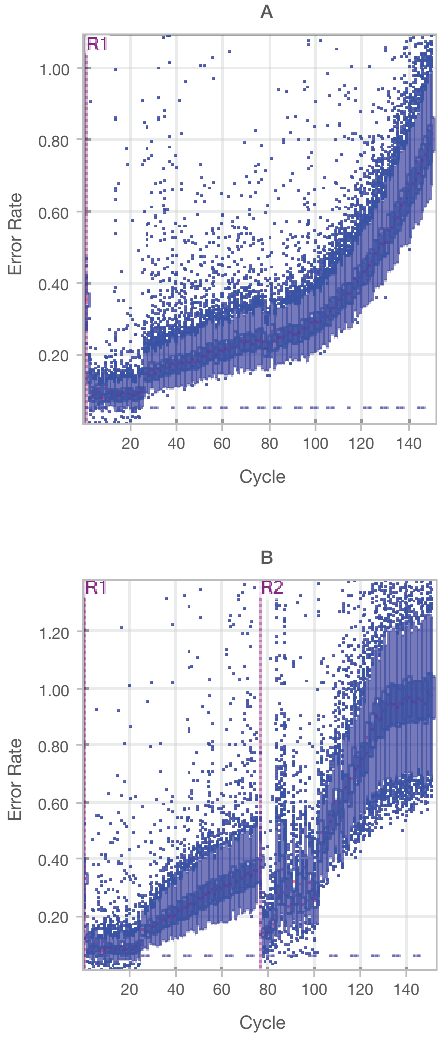

Figure 4: Box plot over all tiles of percentage of base calls that are errors in spiked-in PhiX in each cycle—All clusters passing filters are included from reads assigned to spiked-in control DNA. (A) 1 x 150 bp single-end run and (B) 2 x 75 bp paired-end runs performed on the HiSeq 4000 System.

Secondary Data Analysis

Multiple secondary analyses were performed using the MethylSeq v1.0 BaseSpace® App.[11] From each run, data from multiple lanes were pooled and analyzed to demonstrate the number of samples that can be processed per flow cell according to the desired depth of coverage.

Overall, the coverage obtained using the 1 x 150 bp sequencing run is higher than the 2 x 75 bp run for an equivalent number of lanes (Table 2). There are 2 main reasons for the increased coverage. The first is due to duplicates, which are not removed in the single-read 1 x 150 bp sequencing run. In the paired-end run, duplicate reads are easily distinguished and can be removed from subsequent analyses. The second reason for the lower coverage of the 2 x 75 bp run is the slight reduction in data quality at the end of Read 2, which yields a lower percentage of aligned reads. The other metrics for percent methylation show a high degree of reproducibility between analyses from different run configurations and from combining different numbers of lanes.

Table 2: Primary Analysis Sequencing Metrics from 1 x 150 bp and 2 x 75 bp cycle WGBS Sequencing Runs

Read Length (bp)	1 x 150 (2 lanes)	2 x 75 (2 lanes)	1 x 150 (3 lanes)	2 x 75 (3 lanes)
% Duplicates	N/A	21.0	N/A	24.4
% Aligned Reads	63.8	59.2	64.6	58.2
Coverage	21.7	15.7	32.5	22.7
% Methylated Cs in CpG	51.4	51.5	51.4	51.5
% Methylated Cs in CHG	0.7	0.4	0.7	0.4
% Methylated Cs in CHH	0.5	0.5	0.5	0.5

Guidelines

WGBS involves 2 significant challenges that result from the use of bisulfite converted DNA. First, the conversion of cytosines to thymines causes a reduction in genome diversity that impacts the quality of sequencing data. Second, the conditions required for complete bisulfite conversion result in significant DNA degradation. The first challenge is addressed with the "spike-in" addition of DNA from a well balanced genome, enabling accurate base calling on the sequencer. The TruSeq DNA Methylation Kit addresses the second challenge by capturing ssDNA fragments and reducing sample loss. Using a high-quality library from this kit, while following Illumina-recommended parameters, users can obtain highly accurate MethylSeq results with the HiSeq 3000/HiSeq 4000 Systems.

The HiSeq 3000/HiSeq 4000 Systems provide the option for single-read or paired-end runs. Due to the average size of library products from bisulfite converted DNA, read configurations are optimally set at either 1 x 150 bp or 2 x 75 bp. Both configurations were compared for accuracy as indicated by Q-scores (Figure 3 and Table 1). Low Q-scores decrease the percentage of bases aligned, and increase the amount of data being filtered in the secondary analysis pipeline. Although both run configurations provide high-quality data, the 1 x 150 bp configuration yields slightly higher data accuracy, and increased coverage. However, paired-end runs enable the removal of duplicate reads that can affect subsequent analyses.

For Research Use Only. Not for use in diagnostic procedures.

The addition or "spike-in" of a balanced library, such as the Illumina PhiX Control library, increases data quality by providing more uniform nucleotide representation.[9-12] Illumina recommends spiking-in a minimum of 20% balanced library for WGBS samples run on HiSeq 3000/HiSeq 4000 Systems. Spike-in values below 20% result in degradation in throughput and data quality. Values above 20% increase performance, but also reduce sample coverage. To maximize the amount of usable data, a balanced library other than PhiX can be used, as long as the percentage of unconverted DNA sample on the flow cell is 20% or greater.

Summary

The HiSeq 3000/HiSeq 4000 Systems support a broad range of applications and library prep solutions. With increased capacity and throughput, they are well matched for data-intensive applications such as human whole-genome bisulfite sequencing. Libraries prepared using the TruSeq DNA Methylation Kit can be run using either single-end or paired-end flow cells, on either the HiSeq 3000 or HiSeq 4000 System, with highly accurate results. Recommended read lengths are either 1 x 150 bp or 2 x 75 bp, with 1 x 150 bp offering slightly higher read quality. Samples require a minimum of 20% unconverted DNA spike-in, such as PhiX, or another genome balanced in base representation. With the accurate performance of HiSeq 3000/HiSeq 4000 Systems, together with quality libraries from the TruSeq DNA Methylation Kit and the analytical power of the MethylSeq BaseSpace App, Illumina provides effective tools for the preparation and analysis of WGBS samples.

Learn More

For more details on the results and analysis with the MethylSeq BaseSpace App:

WGBS Data	URL
TruSeq DNA Methylation Kit (1 x 150 bp)	Run Data basespace.illumina.com/s/FtwQQDVu91mx
	Project Analysis basespace.illumina.com/s/G2Bpw38Xtvjb
TruSeq DNA Methylation Kit (2 x 75 bp)	Run Data basespace.illumina.com/s/rJhS2Efk4Rib
	Project Analysis basespace.illumina.com/s/CIXFa92dPAG8

References

1. Gopalakrishnan S, Van Emburgh BO, Robertson KD. DNA methylation in development and human disease. *Mutat Res.* 2008;647(1-2):30-38.
2. Lindholm ME, Marabita F, Gomez-Cabrero D, et al. An integrative analysis reveals coordinated reprogramming of the epigenome and the transcriptome in human skeletal muscle after training. *Epigenetics*. 2014;9(12):1557-1569.
3. Weng X, Zhou D, Liu F, et al. DNA methylation profiling in the thalamus and hippocampus of postnatal malnourished mice, including effects related to long-term potentiation. *BMC Neurosci.* 2014;15(31) doi: 10.1186/1471-2202-15-31.
4. Soto J, Rodriguez-Antolin C, Vallespín E, de Castro Carpeño J, Ibanez de Caceres I. The impact of next-generation sequencing on the DNA methylation-based translational cancer research. *Transl Res.* 2015;S1931-5244(15). doi:10.1016/j.trsl.2015.11.003.
5. Legendre C, Gooden GC, Johnson K, Martinez RA, Liang WS, Salhia B. Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer. *Clin Epigenetics.* 2015;7(1):100. doi:10.1186/s13148-015-0135-8.
6. Roadmap Epigenomics Project. www.roadmapepigenomics.org. Accessed December 22, 2015.
7. llumina (2014) Whole-Genome Bisulfite Sequencing. (www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote-methylseq-wgbs.pdf).
8. Illumina (2014) TruSeq DNA Methylation Kit. (support.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/truseq-dna-methylation-data-sheet-770-2014-051.pdf).
9. Illumina (2014) Using a PhiX Control for HiSeq Sequencing Runs. (support.illumina.com/content/dam/illumina-marketing/documents/products/technotes/hiseq-phix-control-v3-technical-note.pdf).
10. Illumina (2011) Quality Scores for Next-Generation Sequencing. (http://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf).
11. llumina (2015) MethylSeq v1.0 App Guide. (support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/basespace/methylseq-app-guide-15069929-a.pdf).
12. Illumina (2014) Low-Diversity Sequencing on the Illumina HiSeq Platform. (www.illumina.com/documents/products/technotes/technote-hiseq-low-diversity.pdf).

For Research Use Only. Not for use in diagnostic procedures.