Whole-Genome Bisulfite Sequencing

Efficient nanogram-scale library preparation, sequencing, and analysis of DNA methylation.

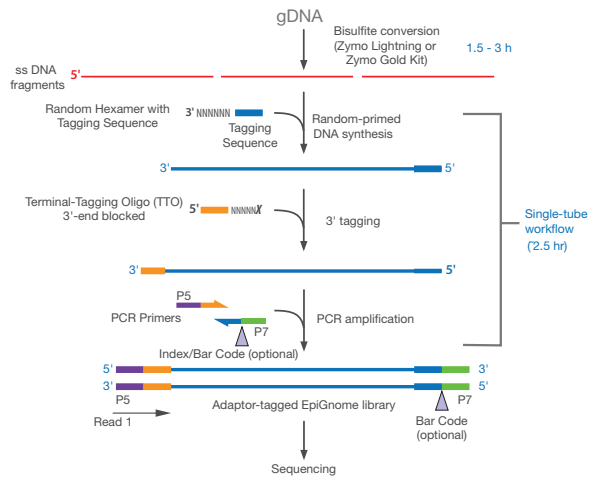
Introduction

Methylation of DNA at cytosine nucleotides forms 5-methylcytosine (5mC), which impacts various cellular processes involving gene expression and chromatin remodeling. These processes can influence health and development through methylation of promoter regions, cell differentiation, remodeling of chromatin for selective X-chromosome inactivation, and suppression of transposable elements.

Next-generation sequencing technology has enabled genome-wide analysis of 5mC nucleotides at single-nucleotide resolution. Previous methods for whole-genome bisulfite sequencing (WGBS) yielded reduced genomic representation due to required DNA shearing, ligation of methylated sequencing adapters, and bisulfite conversion of unmethylated cytidine residues prior to sequencing. One disadvantage of this workflow is that most adapter-tagged DNA fragments are degraded during bisulfite conversion and are, therefore, not included in sequencing results.

This application note describes an improved technique developed by Epicentre® (an Illumina company) where bisulfite conversion takes place prior to addition of sequencing adapters. The highly diverse libraries produced contain most of the DNA fragments as viable sequencing templates. The novel workflow reduces the amount of sample required down to 50–100 ng, enabling new types of samples to be investigated, such as cancer biopsies. The resulting libraries are sequenced on the Illumina HiSeq® 2500 System, and differentially methylated sites can be visualized on the genome at single-nucleotide resolution.

Figure 1: EpiGnome Library Preparation Workflow



The EpiGnome workflow results in di-tagged DNA that is amplified by PCR, resulting in directional libraries with the appropriate adapters for Illumina sequencing.

Methods and Results

Genomic DNA Isolation

Genomic DNA is isolated from 1 mg to 5 mg of tissue, using the MasterPure™ DNA Purification Kit (Epicentre) according to the manufacturer's instructions. This kit employs a non-enzymatic approach to cell lysis, followed by protein precipitation and subsequent nucleic acid isolation, resulting in high yields of purified, high-molecular-weight DNA. The extracted DNA is resuspended in TE buffer and quantitated by fluorometry.

Bisulfite Conversion and Library Preparation

Bisulfite treatment of genomic DNA is performed using the Zymo EZ DNA Methylation Lightning Kit[1]. With this method, non-methylated cytosine nucleotides are converted to uracil and read as thymine (T) when sequenced. Methylated cytosines protected from conversion are still read as cytosine (C). Briefly, 50–100 ng of purified genomic DNA is treated with Zymo Lightning Conversion Reagent in a thermal cycler for 8 minutes at 98ºC, followed by 60 minutes at 54ºC.

The bisulfite-treated DNA is purified on a spin column and is used to prepare the sequencing library using the EpiGnome™ Kit (Epicentre). In this procedure, bisulfite-treated single-stranded DNA is random-primed using a polymerase able to read uracil nucleotides, to synthesize DNA containing a specific sequence tag. The 3′ ends of the newly synthesized DNA strands are then selectively tagged with a second specific sequence, resulting in di-tagged DNA molecules with known sequence tags at their 5′ and 3′ ends (Figure 1). These tags are then used to add Illumina P7 and P5 adapters by PCR at the 5′ and 3′ ends, respectively, of the original DNA strand.

Only the complement to the original bisulfite-treated DNA is used as the sequencing template; thus, the resulting Read 1 will always be the same sequence as the original bisulfite-treated strands.

Sequencing

The EpiGnome libraries are diluted and loaded onto the cBot DNA Cluster Generation System. After cluster generation is complete (approximately 5 hours), the flow cell is transferred to the HiSeq 2500 System for sequencing using 75 bp paired-end reads. The HiSeq 2500 generates approximately 500 Gb of sequence data per flow cell or about 62 Gb per lane. Therefore, a single human genome library can be run across two lanes of the eight-lane flow cell to generate approximately 120 Gb of data per sample. Additional sequencing can be completed for higher coverage.

Table 1: Coverage Details

Total Cs analyzed	8,655,380,797
Methylated Cs in CpG context	276,057,623
Methylated Cs in CHG context	14,630,414
Methylated Cs in CHH context	50,567,172
Unmethylated Cs in CpG context	216,539,261
Unmethylated Cs in CHG context	2,024,464,291
Unmethylated Cs in CHH context	6,073,122,036
Percent methylation (CpG context)	56.0%
Percent methylation (CHG context)	0.7%
Percent methylation (CHH context)	0.8%

H = A, C, or T

Data Analysis

Methylation analysis[2] is performed using Bismark[3]. In summary, FASTQ files are quality-filtered and adapter sequences trimmed using Trimmomatic[4]. A bisulfite-converted UCSC HG19 reference genome file is generated using Bowtie 2[5], and the EpiGnome library sequence data are aligned to the reference genome. Methylation information is extracted from the output *.sam file (Table 1), and genome tracks are output for visualization and reporting of downstream differential methylation calculations.

The methylation extraction report should show minimal bias as depicted in the M-bias plot (Figure 2), which shows the percent methylation across positions in the read. A perfectly unbiased sequencing run would be a horizontal line.

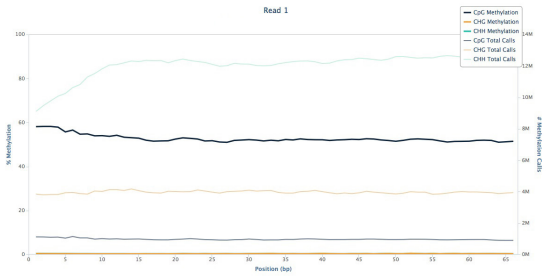
Visualization of methylated sites of the genome can be performed using the Integrative Genomics viewer (Broad Institute)[6] as shown in Figure 3. The region of chromosome 1 shown contains both hypo- and hyper-methylated regions.

The methylKit R package can calculate differential methylation and annotate differentially methylated sites[6].

Conclusions

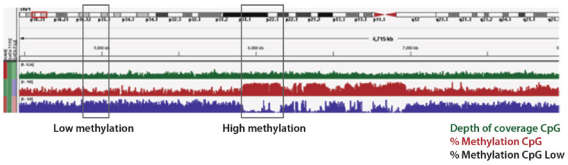
The workflow described in this application note provides an efficient method to prepare highly diverse libraries, sequence DNA, and analyze methylation of whole genomes. The post-bisulfite EpiGnome library preparation ensures that most of the genomic DNA is converted to usable sequencing templates, unlike conventional WGBS methods. The high retention of DNA fragments reduces the required starting amount and enables investigation of new sample types that were previously not possible due to high input DNA requirements. Sequencing on the HiSeq 2500 System and the optimized set of data analysis tools provide minimal bias and flexibility in displaying whole-genome methylation data.

Figure 2: M-Bias Plot



The M-bias plot shows the percent methylation along the read for each type of methylation indicated. The plot shows minimal bias for most methylation contexts including CpG, which is the most commonly studied.

Figure 3: CpG Methylation Coverage



CpG methylation pattern across a 10 Mb region of chromosome 1. CpG coverage is shown in green. Regions of high CpG methylation are shown in red and regions of low CpG methylation are shown in blue.

References

1. www.zymoresearch.com/downloads/dl/file/id/490/d5030i.pdf
2. www.epibio.com/docs/default-source/protocols/epignome-bioinformatics-user-guide.pdf?sfvrsn=2
3. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27: 1571–1572.
4. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. doi: 10.1093/bioinformatics/btu170
5. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359.
6. www.broadinstitute.org/igv/home
7. methylkit.googlecode.com/git/inst/doc/methylKit.pdf

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

© 2014 Illumina, Inc. All rights reserved.
Illumina, cBot, CSPro, Epicentre, EpiGnome, HiSeq, MasterPure, the pumpkin orange color, and the Genetic Energy streaming bases design are trademarks or registered trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners.
Pub. No. 470-2014-004 Current as of 30 July 2014

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA