Methods

Sample collection and preparation

➢ DNA quantification and qualification

Genomic DNA degradation and contamination was validated by agarose gels. DNA purity was checked using NanoPhotometer® spectrophotometer (IMPLEN, CA, USA). DNA concentration was measured using Qubit® DNA Assay Kit in Qubit® 2.0 Flurometer (Life Technologies, CA, USA).

➢ Library preparation and quantification

A total amount of 100 ng genomic DNA spiked with 0.5 ng lambda DNA were fragmented by sonication to 200-300 bp with Covaris S220. These DNA fragments were treated with bisulfite using EZ DNA Methylation-GoldTM Kit (Zymo Research), and the library was constructed by Novogene Corporation (Beijing, China). Subsequently, pair-end sequencing of sample was performed on Illumina platform (Illumina, CA, USA). Library quality was assessed on the Agilent Bioanalyzer 2100 system.

Data Analysis

The library was sequenced on Illumina Novaseq platform. Image analysis and base calling were performed with Illumina CASAVA pipeline, and finally generated 150bp paired-end reads.

➢ Quality control

First of all, we used FastQC (fastqc_v0.11.5) to perform basic statistics on the quality of the raw reads. Then, those reads sequences produced by the Illumina pipleline in FASTQ format were pre-processed through fastp (fastp 0.20.0). The remaining reads that passed all the filtering steps was counted as clean reads and all subsequent analyses were based on this. Finally, we used FastQC to perform basic statistics on the quality of the clean data reads.

➢ Reference data preparation before analysis

Before the analysis, we prepared the reference data for the species we study, including the reference sequence fasta file, the annotation file in gtf format, the GO annotation file, the description file and the gene region file in bed format. We predicted repeats with RepeatMasker, and CGI track from a genome with cpgIslandExt.

➢ Reads mapping to the reference genome

Bismark software (version 0.16.3; Krueger F, 2011) was used to perform alignments of bisulfite-treated reads to a reference genome (-X 700 --dovetail). The reference genome was firstly transformed into bisulfite-converted version (C-to-T and G-to-A converted) and then indexed using bowtie2 (Langmead B, 2012). Clean reads were also transformed into fully bisulfite-converted versions (C-to-T and G-to-A converted) before being aligned to the similarly converted versions of the genome in a directional manner. Sequence reads that produce a unique best alignment from the two alignment processes (original top and bottom strand) were then compared to the normal genomic sequence and the methylation state of all cytosine positions was inferred. The same reads that aligned to the same regions of genome were regarded as duplicated ones. The sequencing depth and coverage were summarized using deduplicated reads.

The results of methylation extractor (bismark_methylation_extractor, -- no_overlap) were transformed into bigWig format for visualization using IGV browser. The sodium bisulfite non-conversion rate was calculated as the percentage of cytosine sequenced at cytosine reference positions in the lambda genome.

➢ Estimating methylation level

Methylated sites were identified with a binomial test using the methylated counts (mC), totols counts (mC+umC) and the non-conversion rate (r). Sites with FDR-corrected p-value<0.05 were considered as a methylated site. To calculate the methylation level of the sequence, we divided the sequence into multiple bins, with bin size is 10 kb. The sum of methylated and unmethylated read counts in each window were calculated. Methylation level (ML) for each window or C site shows the fraction of methylated Cs, and is defined as:

$$ML(\mathrm{C}) = \frac{reads(\mathrm{mC})}{reads(\mathrm{mC}) + reads(\mathrm{C})}$$

➤ Differentially methylated analysis

Differentially methylated regions (DMRs) were identified using the DSS software (Hao Feng, Hao Wu, 2014; Hao Wu, 2015; Yongseok Park Hao Wu,2016) , The core of DSS is a new dispersion shrinkage method for estimating the dispersion parameter from Gamma-Poisson or Beta-Binomial distributions. According to the distribution of DMRs through the genome, we defined the genes related to DMRs as genes whose gene body region (from TSS to TES) or promoter region (2 kb upstream from the TSS) have an overlap with the DMRs.

➤ GO and KEGG enrichment analysis of DMR-related genes

Gene Ontology (GO) enrichment analysis of genes related to DMRs was implemented by the GOseq R package (Young MD, 2010), in which gene length bias was corrected. GO terms with corrected P-value less than 0.05 were considered significantly enriched by DMR-related genes. We used KOBAS software (Mao X, 2005) to test the statistical enrichment of DMR- related genes in KEGG pathways.

References

Langmead B, Salzberg SL. (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods, 9(4): 357-9. (Bowtie 2)

Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD, et al. 2013. Global epigenomic reconfiguration during mammalian brain development. Science 341:1237905

Krueger F, Andrews SR. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics, 27(11): 1571-2. (Bismark)

Hao Feng, Karen N. Conneely, Hao Wu. (2014) A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Research, 2014, Vol. 42, No. 8 (DSS)

Hao Wu, Tianlei Xu, et al. (2015) Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. Nucleic Acids Research, 2015 (DSS)

Yongseok Park and Hao Wu. (2016) Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics (DSS)

Kanehisa M, Araki M, Goto S, et al. (2008). KEGG for linking genomes to life and the environment. Nucleic Acids Res, 36(Database issue): D480-4. (KEGG)

Mao X, Cai T, Olyarchuk JG, et al. (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics, 21(19): 3787-93. (KOBAS)

Smallwood SA, Lee HJ, Angermueller C, et al. (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods, 11(8): 817- 20.

Young MD, Wakefield MJ, Smyth GK, et al. (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol, 11(2): R14. (GOseq)

Zhao L, Sun MA, Li Z, et al. (2014) The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. Genome Res, 24(8): 1296-307.