RRBS Methods

1. Experimental Procedure

1.1 Sample Quality Control

Please refer to QC report for methods of sample quality control.

1.2 Library Construction, Quality Control and Sequencing

A total amount of 1.0 μg genomic DNA spiked with moderate lambda DNA were handled by MspI, followed by end repair and adenylation. Cytosine-methylated barcodes were ligated to DNA as per manufacturer's instructions. DNA fragments in the length of 40−220 bp were selected with gel extraction. Then these DNA fragments were treated with bisulfite using EZ DNA Methylation-GoldTM Kit (Zymo Research).
The library was constructed by Novogene Corporation (Beijing, China). Subsequently, library quality was assessed on the Agilent 5400 system （Agilent, USA） and quantified by QPCR (1.5 nM). Pair-end sequencing of sample was performed on Illumina platform (Illumina, USA).

2. Bioinformatics Analysis Pipeline

2.1 Data Quality Control

First of all, we used FastQC (fastqc_v0.11.8) 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 0.23.1). 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.

2.2 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.

2.3 Reads mapping to the reference genome

Bismark software (version 0.24.0; Krueger et al., 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 et al., 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.

2.4 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:

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

2.5 Differentially methylated analysis

Differentially methylated regions (DMRs) were identified using the DSS software （Feng et al., 2014， Park et al., 2016, Wu et al., 2015）, 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.

2.6 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 et al., 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 et al., 2005) to test the statistical enrichment of DMR-related genes in KEGG pathways.

3 References

Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. 2014;42(8):e69. doi:10.1093/nar/gku154

Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27(11):1571-1572. doi:10.1093/bioinformatics/btr167

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-359. Published 2012 Mar 4. doi:10.1038/nmeth.1923

Wu H, Xu T, Feng H, et al. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. Nucleic Acids Res. 2015;43(21):e141. doi:10.1093/nar/gkv715

Mao X, Cai T, Olyarchuk JG, Wei L. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics. 2005;21(19):3787-3793. doi:10.1093/bioinformatics/bti430

Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11(2):R14. doi:10.1186/gb-2010-11-2-r14