WGBS Methods

1.Bioinformatics Analysis Pipeline

1.1 Data Quality Control

First of all, we use FastQC (fastqc_v0.11.8) to perform basic statistics on the quality of the raw reads. Then, those reads sequences produced by the Illumina pipeline in FASTQ format were pre-processed through Trimmomatic (Trimmomatic-0.36) software use the parameter (SLIDINGWINDOW: 4:15 ; LEADING:3, TRAILING:3 ; ILLUMINACLIP: adapter.fa: 2: 30: 10 ; MINLEN:36) The remaining reads that passed all the filtering steps was counted as clean reads and all subsequent analyses were based on this. At last, we use FastQC to perform basic statistics on the quality of the clean data reads.

1.2 Reference data preparation before analysis

Before the analysis, we have to prepare the reference data for the species we study, which includes 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. As for the bed files, we predict repeats through RepeatMasker, followed by getting CGI track from a genome use cpgIslandExt.

1.3 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). Sequence reads were also transformed into fully bisulfite-converted versions (C-to-T and G-to-A converted) before they are aligned to 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) are then compared to the normal genomic sequence and the methylation state of all cytosine positions in the read is 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-coversion rate was calculated as the percentage of cytosine sequenced at cytosine reference positions in the lambda genome.

1.4 Estimating methylation level

To identify the methylation site, we modeled the sum Mc of methylated counts as a binomial (Bin) random variable with methylation rate.
In order to calculate the methylation level of the sequence, we divided the sequence into multiple bins, with bin size is 10kb. 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:

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

1.5 Differentially methylated analysis

Differentially methylated regions (DMRs) were identified using the DSS software(Feng et al., 2014; Wu et al., 2015; Park et al., 2016). The core of DSS is a new dispersion shrinkage method for estimating the dispersion parameter from Gamma-Poisson or Beta-Binomial distributions.
DSS possess three characteristics to detect DMRs. First, spatial correlation. Proper utilization of the information from neighboring Cytosine sites can help improve estimation of methylation levels at each Cytosine site, and hence improve DMR detection. Second, the read depth of the Cytosine sites provides information on precision that can be exploited to improve statistical tests for DMR detection. Finally, the variance among biological replicates provides information necessary for a valid statistical test to detect DMRs， when there is no biological replicate, DSS combining data from nearby Cytosine sites and using them as ‘pseudo-replicates’ to estimate biological variance at specific locations.
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 (upstream 2kb from the TSS) have an overlap with the DMRs.

1.6 GO and KEGG enr ichment analysis of DMR-r elated 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
KEGG (Kanehisa et al., 2008) is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high- through put experimental technologies (http://www.genome.jp/kegg/). We used KOBAS software (Mao et al., 2005) to test the statistical enrichment of DMR- related genes in KEGG pathways.

2 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

Kanehisa M, Araki M, Goto S, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008;36(Database issue):D480-D484. doi:10.1093/nar/gkm882

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

Lister R, Mukamel EA, Nery JR, et al. Global epigenomic reconfiguration during mammalian brain development. Science. 2013;341(6146):1237905. doi:10.1126/science.1237905

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

Park Y, Wu H. Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics. 2016;32(10):1446-1453. doi:10.1093/bioinformatics/btw026

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

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