Whole Genome Bisulphite Sequencing (WGBS) data processing workflow

The paired end Illumina WGBS sequence data was processed using in house scripting (bash and R) and a range of purpose built bioinformatics tools on University of Edinburgh high performance computing facility (Eddie3). Briefly FASTQ files for each sample across multiple lanes were merged together.

Raw reads were trimmed to remove the low complexity sequence tag introduced during Accel-NGS Methyl-seq DNA kit library prepration and trim bases with phred score less than 20 using TrimGalore! v. 0.5.0 (trim_galore -q 20 --fastqc --paired --clip_R2 18 --three_prime_clip_R1 18 --retain_unpaired –o Trim_out INPUT_R1.fq.gz INPUT_R2.fq.gz)

A bisulphite-sequencing amenable reference genome was built using the Rambouillet Benz2616 genome available from NCBI (Oar_rambouillet_v1.0 GCA_002742125.1) with the BSSeeker2 script bs_seeker2-build.py using bowtie v2.3.4.3[1] and default parameters. Paired-end, trimmed reads were aligned to the reference genome using the BSSeeker2 script bs_seeker2-align.py and bowtie v2.3.4.3[1] allowing four mismatches (-m 4). Aligned bam files were sorted with samtools v1.6[2] and duplicate reads were removed with picard tools v2.17.11 (https://broadinstitute.github.io/picard/) MarkDuplicates function. Unpaired reads were retained and aligned with the same settings before being merged with the paired end data.

DNA methylation levels were called using the BSSeeker2 script bs_seeker2-call_methylation.py with default options to generate ATCGmap and CGmap files for each sample. Read depth and methylation level coverage statistics and visualisations were generated within CGmaptools[3].

Written by Alex Caulton

The correlation between samples based on the similarity of their methylation profiles was determined utilizing the methylKit package[4] within R. Specifically individual cytosine methylation levels were filtered for a minimum read depth of 10x coverage and CGmap files were converted to methylkit appropriate format using a custom python script. The pearson correlation coefficient between samples was determined based on similarity of methylation profiles between samples.

Hyper-methylated and hypo-methylated regions were determined for each sample using methpipe v3.4.3[5]. Specifically, CGmap files for each sample were reformatted for the methpipe v3.4.3 workflow using custom awk scripts. The methpipe symmetric-cpgs program was used to merge individual methylation levels at symmetric CpG pairs. Hypo-methylated and hyper-methylated regions were determined using the hmr program within methpipe which uses a hidden Markov model (HMM) using a Beta-Binomial distribution to describe methylation levels at individual CpG sites, accounting for the read coverage at each site.

Visualisation of the individual CpG site methylation levels with a minimum read depth cut-off of 10x coverage was done using Gviz package v.1.28.3[6].

References

1. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
2. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).
3. Guo, W. *et al.* CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. *Bioinformatics* **34**, 381–387 (2018).
4. Akalin, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, R87 (2012).

Written by Alex Caulton

5. Song, Q. *et al.* A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics. *PLOS ONE* **8**, e81148 (2013).

6. Hahne, F. & Ivanek, R. Visualizing Genomic Data Using Gviz and Bioconductor. in *Statistical Genomics: Methods and Protocols* (eds. Mathé, E. & Davis, S.) 335–351 (Springer, 2016). doi:10.1007/978-1-4939-3578-9_16.