Supplemental code file for “Enhanced Reduced Representation Bisulfite Sequencing for assessment of DNA methylation at base pair resolution”

This file contains instructions on how to perform sequence alignment and analyze data from Enhanced Reduced Representation Bisulfite Sequencing (ERRBS) results. Further information on analysis of aligned data can be found at: <https://code.google.com/p/methylkit/>, in the user guide at <http://methylkit.googlecode.com/git/inst/doc/methylKit.pdf>, and as described in Akalin, A. *et al*.1

> indicates command lines to be used directly in R

PATH represents the directory path in which data files (ERRBS data as shown in Table 2) are stored

fastq1 is an example file name for use in data alignment

filename1 and filename2 are example file names for use and sample1 (reference) and sample2 (experimental) are example sample names for use in the commands

GENOME = genome used for alignments of data (for example: hg18 or hg19 or other genome)

**Data alignment:**

The objective of the data alignment is to convert this sequenced data to a concise and precise CpG methylation report. The analysis described consists of the following steps: 1) quality filtering of the reads, 2) alignment to the reference genome using modified alignment procedure that accounts for the C->T conversion introduced during the bisulfite conversion step, and 3) generation of files for quality control review, data visualization and further data analysis. Currently, there is no single software package that performs this entire process from start to end. There are a number of packages that are available for the purpose of this analysis that have been integrated into user-friendly systems such as Galaxy1 and Gobi-web2. The analysis described here requires running command line programs and scripting, familiarity with software packages and access to a compute server (typically a 60GB RAM linux server).

1. Filter the reads for reads that pass quality filtering (protocol step 11.2):

> gunzip -c fast1.fastq.gz | grep -A 3 '^@.\* [^:]\*:N:[^:]\*:' | grep -v '^\-\-$' >> fastq\_PF.fastq

The <is filtered> field contains a Y if the reads is filtered or an N if the read not filtered (i.e. passes the filter). An example of Illumina identifiers are:

@HWI-ST963:320:C5BLMDCXX:1:1201:1459:1876 1:Y:0:

…

@HWI-ST963:320:C5BLMDCXX:1:1201:1459:1876 1:N:0:

…

1. Trim adapter sequences from pass filter reads (protocol step 11.3):

> far --source /fast1\_PF.fastq --target fastq1\_PF\_trimmed --format fastq --min-overlap 6 --adapters adapters.fa --min-readlength 21 --cut-off 2 --trim-end right\_tail -th 4 -u 2

or

> flexbar -r fastq1\_PF.fastq -t fastq1\_PF\_trimmed -f i1.8 -ao 6 -a adapters.fa -m 21 -at 2 -ae right -u 2 -n 4

1. Align the reads using Bismark (protocol step 11.4):

> bismark -l 50 -q --unmapped --ambiguous --phred33-quals --directional /genome/hg19 fastq1\_PF\_trimmed.fastq

1. Sort the aligned reads (protocol step 11.5):

> sort -k4,4 -k5,5n -k3,3 fastq1\_PF\_trimmed.fastq\_bismark.txt -o fastq1\_sortedBismarkAlignments.txt

1. Compute methylation scores (protocol step 11.6; see custom script I in Supplementary Code file 2):

> methylationCall\_fromBismark.pl --minqual 20 --mincov 0 --sample fastq1 fastq1\_sortedBismarkAlignments.txt

1. Compute bisulfite conversion rates (protocol step 11.7; see custom script II in Supplementary Code file 2):

> conversionRate.pl -mincov 10 -cpg methylcall.CpG.fastq1.mincov0.txt -chg methylcall.CHG.fastq1.mincov0.txt -chh methylcall.CHH.fastq1.mincov0.txt

1. Generate BAM format file (protocol step 11.8):

> bismark2SAM\_v5\_xm.pl -c hg19.chrom.sizes -i fastq1\_sortedBismarkAlignments.txt -o fastq1.sam

> samtools view -bS fastq1.sam > fastq1.bam

> samtools sort fastq1.bam fastq1.sorted

> samtools index fastq1.sorted.bam

1. Generate wiggle format file (protocol step 11.8):

> createWigFromCpG.pl -g /genomes/hg19 -i cpg.fastq1.mincov10.txt -o fastq1.wig

**Data analysis:**

1. Download/install the methylKit3 R package from <https://code.google.com/p/methylkit/#Installation> and load the package into R:

> library(methylKit)

2. Read in the methylation call files:

> file.list = list(“PATH/filename1”, “PATH/filename1”)

3. Read the files to a methylRawList object (myobj):

> myobj=read(file.list, sample.id=list("sample1","sample2"), assembly="GENOME", treatment=c(0, 1))

4. Determine strand specific percent methylation statistics (figure 5C):

> getMethylationStats(myobj[[2]],plot=F,both.strands=T) # determines statistics for second sample

> getMethylationStats(myobj[[2]],plot=T,both.strands=T) # determines statistics and generates plot (as seen in figure 5C for second sample)

5. Determine read coverage per CpG information (Figure 5B):

> getCoverageStats(myobj[[2]],plot=T,both.strands=T) # determines frequency of read coverage per CpG for second sample

6. merge samples into a single object for further analysis:

> meth=unite(myobj, destrand=FALSE)

7. determine correlation between the two samples (see figure 5D):

> getCorrelation(meth,plot=T)

8. Determine CpG sites shared by both samples and what is the differential methylation between them:

> myDiff=calculateDiffMeth(meth) # this object is further filtered for statistically significant differences (for example qvalue<0.01 and methylation difference >25%).

9. Annotate CpGs shared by both samples to genomic regions (figures 5E and 5F):

To perform this command, first download annotation files from UCSC as instructed at <https://code.google.com/p/methylkit/#Downloading_Annotation_Files> (genic feature file and CpG island feature file) and import annotation files into R:

> gene.obj=read.transcript.features(“("/PATH/genic features file”) # annotation file for genic features

> cpg.obj=read.feature.flank("/PATH/CpG island Annotation file name")

Annotate and plot genic annotated sites:

> diffAnn=annotate.WithGenicParts(myDiff,gene.obj)

> plotTargetAnnotation(diffAnn,precedence=TRUE, main="CpG methylation annotation")

> cpgi.ann=annotate.WithFeature.Flank(myDiff, cpg.obj[[1]], cpg.obj[[2]],feature.name="CpGi",flank.name="Shores")

> plotTargetAnnotation(cpgi.ann,col=c("green","gray","white"), main="CpG annotation")

**References:**

1 Goecks, J., Nekrutenko, A., Taylor, J., Galaxy, T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. **11** (8), R86, doi: 10.1186/gb-2010-11-8-r86 (2010).

2 Dorff, K.C. *et al.* GobyWeb: simplified management and analysis of gene expression and DNA methylation sequencing data. *PLoS One*. **8** (7), e69666, doi: 10.1371/journal.pone.0069666 (2013).

3 Akalin, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* **13**, R87, doi:10.1186/gb-2012-13-10-r87 (2012).