Documentation

1. Introduction

Workflow consists of running read2cpg.exe to create the tables with read start position and then the function estimate_covfunction.r as part of an R script.

Below we describe the input and output for the two components. We recommend downloading the example (files can be extracted with tar -xjvf example.tar.bz2 on Linux or other software such as WinZip www.winzip.com). This example also has a shell script (run_example.sh) to illustrate how to run the two programs with a single command.

Although the code has some more options (which would require modifying the identified setting at the beginning of the source code), we limited the number of arguments by focusing on the most relevant scenarios..

All input and output files with multiple columns are comma delimited files with headers and extension .csv. File with a single column have no headers or extension.

2. read2cpg.exe

Table 1. Arguments for read2cpg.exe

Arguments	
	INPUT
1	*Label of the sample that will be processed, eg. Sample1*
2	*Root directory where the read files are: assumes sba and multi subdirectories*
3	*Isolated CpG coordinate file prefix to which program will add the string "chr#"*
	OUTPUT
4	*Directory where the tables with counts of read start positions will be written*
5	*Directory where the reads that were QC'ed out will be written as well as a summary file.*
	SETTINGS
6	*Number of chromosomes*
7	*Include multi reads (yes/no)*
8	*Hard threshold: duplicate reads with counts higher than this will be eliminated*
9	*Soft threshold: for duplicate reads with counts higher than this but lower than the hard threshold, the program will look in the neighborhood to see if other reads are being aligned there*
10	*Neighborhood: specifies in bp the neighborhood to screen for other reads in either direction. If other reads are found the duplicate read will not be QC'ed out*
11	*Constant "d", which is the value to define the [C-d,C+d] interval to define the isolated CpGs*

ad argument 2)

The program assumes that the specified directory has the following subdirectories for uniquely mapped/sba reads (single best alignment) and multireads (reads aligning to multiple locations):

- ./sba
- ./mulit

Both directories have further subdirectories chr<i>, where <i> is the chromosome number. In each subdirectory chr<i>, every subject has its own data stored in comma delimited files with headers. The filenames have the structure: <sampleID>.read.

The columns in the .read files are:

coord: an integer that represents the coordinate start position where a sequenced read was aligned to the reference genome

pos: the number of reads aligning to the positive strand of the reference genome at this location

neg: the number of reads aligning to the positive strand of the reference genome at this location

Although not required, in the case of multireads we can have additional columns
'readnum': the number of the aligned read in the alignment match file.
'Totalalignments': The number of times a read was aligned on the reference
'Bestalignments': The number of alignments among the total number of alignments.

The files are assumed to be sorted and if a coordinate is encountered multiple times the reads at these will be automatically added

ad argument 3)

Isolated CpG coordinate file prefix to which program will add the string "chr#". The program assumes the files have no headers and just the coordinates of the isolated CpGs on the forward strand of the genome. All QC related output will involve these isolated CpGs only

ad argument 4)

Output consists of individual files with QC'ed reads and summary files with counts of included and excluded reads based on criteria specified in arguments 8-10. Output files are labeled with the sample ID, QC thresholds, and have added label "summary" for the summary file. For example:

./Sample1_ isoCpG_hard1_soft1_neigh25.csv
./summary_Sample1_ isoCpG_hard1_soft1_neigh25.csv

The files with QC'ed reads have the following columns

1) Chr: chromosome
2) coord: an integer that represents the coordinate start position of the QC;ed read
3) pos: the number of QC'ed reads aligning to the positive strand of the reference genome
4) neg: the number of QC'ed reads aligning to the negative strand of the reference
5) readtype: this is the type of read with 0 for uniquelymapped/sba reads (single best alignment) and 1 for multireads (reads aligning to multiple locations):

The summary file has 8 records and two columns. The first column gibes the label and the second column a count. The labels are:

1) iso_sba_records: a count of the number of unique location for sba reads
2) iso_sba_reads: a count of the number of sba reads before any QC
3) iso_sba_dupl_peaks: a count of the number of unique location of sba reads that were QC'ed
4) iso_sba_dupl_reads: a count of the number of sba reads that were QC'ed
5) iso_multi_records: a count of the number of unique location for multi reads
6) iso_multi_reads: a count of the number of multi reads before any QC
7) iso_multi_dupl_peaks: a count of the number of unique location of multi reads that were QC'ed
8) iso_multi_dupl_reads: a count of the number of multi reads that were QC'ed

ad argument 5)

Output consists of individual files with counts of number of reads starting at position 0 to “d”, which is the value to define the [C-d,C+d] interval to define the isolated CpGs. The two columns in the file indicate the postion (“pos” label) and the count (count label).

Output files are labeled with the sample ID and QC thresholds. For example,

./Sample1_hard3_soft1_neigh25.csv

3. estimate_covfunction.r

Table 2. Arguments estimate_covfunction.r

Arguments	
1	Vector of sample IDs
2	Directory where output will be written
3	Names of the files with the read count data
4	Analysis label (used to label the output files), For example hard1_soft1_neigh25.
4	Minimum possible read length
5	Estimate of maximum fragment size
6	MonoProc library installed (yes/no). If "no" then will look for package in subdirectory ./monoProc

ad argument 2)

1. Pdf file with figures of raw data plus curve fitting results, estimated coverage functions, and number of observations

Example: hard1_soft1_neigh25_figures.pdf

2. File with number of observations (=reads)

Example: hard1_soft1_neigh25_nobs.csv

3. Three files with coverage functions: 1) Kernel based estimates, 2) cubic spline based estimates, 3) files will aggregates which may be the mean across all individual estimates or obtained by estimating the coverage function after adding the read start data across all samples.

Example: hard1_soft1_neigh25_kernel_cov_func.csv

4. Three files with coverage standardization factors: 1) Kernel based estimates, 2) cubic spline based estimates, 3) files with aggregates (the mean across all individual estimates or obtained by estimating the coverage function after adding the read start data across all samples).

Example: hard1_soft1_neigh25_aggregate_stand_factors.csv

ad argument 3)

Input data is a table with read start counts after QC (e.g. as produced by read2cpg.exe)