BS-RNA - User Guide v1.0

Introduction

Cytosine methylation is one of the most important RNA epigenetic modifications. With the development of experimental technology, scientists attach more importance to RNA cytosine methylation and find bisulfite sequencing is an effective experimental method for RNA cytosine methylation study. However, there is no specialized tool that can deal with RNA bisulfite sequencing data at present. Hence, we have developed BS-RNA, which can analyze RNA methylation for bisulfite sequencing data. Evaluation results of the rates of uniquely mapped reads, rates of correctly mapped reads, and running time for simulated data and actual data indicate that BS-RNA can provide fast and accurate mapping of RNA bisulfite sequencing reads. Comparison between annotated cytosine methylation used BS-RNA and published research also shows that BS-RNA is an effective annotation tool for RNA bisulfite sequencing data. Its annotation result is in BED (.bed) format, including locations, sequence context types (CG/CHG/CHH, H = A, T, or C), reference sequencing depths, cytosine sequencing depths, and methylation levels of covered cytosine sites both Watson and Crick strands BS-RNA supports on both paired-end and single-end sequencing short reads.

Principle

The BS-RNA process includes three main steps: pretreatment, mapping, and annotation. The first step is the pretreatment of reference genome sequences, sequencing data, and gene annotation file: (1) the reference genome sequence is converted twice as follows: (A) cytosines are replaced by

thymines and (B) guanines are replaced by adenines. If the user needs to map multiple sequencing data to the same reference genome sequence, it could be converted only once. (2) cytosines in reads of the T-rich sequencing read file are replaced by thymines, while guanines in reads of the A-rich sequencing read file are replaced by adenines. BS-RNA only supports RNA bisulfite sequencing data from the directional library. And (3) the gene annotation file in gtf format is revised to fit the converted reference genome sequences. Each annotation line is converted twice simultaneously: “C-T” and “G-A” are appended to the chromosome label in each gene annotation file, respectivelyNext, the HISAT program is invoked by BS-RNA to build alternative splicing according to the modified annotation gene file and map preprocessed reads to the converted reference genome sequence. BS-RNA can process both single-end and paired-end bisulfite sequencing RNA data. BS-RNA filters out two types of reads that are mapped to the reference genome sequence as follows: (1) reads mapped to multiple positions and (2) reads mapped to the wrong strands (T-rich reads mapped to reverse-complement of reference sequence converted C to T or to reference sequence converted G to A, A-rich reads mapped to reference sequence converted C to T or to reverse complements of reference sequence converted G to A). The original mapping file (SAM format), filtered mapping file (SAM format), and mapping report file will be provided by BS-RNA after the mapping step is finished. Further, BS-RNA splits the filtered mapping file into two: a Watson-strand SAM format file (mapped to reference converted C to T) and a Crick-strand SAM format file (mapped to reference converted G to A). SAMtools (Li et al., 2009) is embedded into BS-RNA to convert the Watson-strand and Crick-strand SAM format files to BAM format and then sort these two BAM files according to the coordination. Single-base coverage information is extracted using the mpileup command of SAMtools from the sorted BAM

files. Next, for each cytosine position in the genome, the read base (SAMtools mpileup output information) is deemed to support methylated cytosine sequencing if it matches a dot or a comma, otherwise it is deemed to support an unmethylated cytosine sequencing (where "<" or ">" is not counted). BS-RNA provides the annotation result for each covered cytosine in a BED (.bed) file with information related to the methylation character: location of the covered cytosine in the reference, sequence context type (CG/CHG/CHH, H = A, T, or C), reference sequencing depth (total number of reads mapped to the cytosine site), cytosine sequencing depth (total number of reads that supported a methylated cytosine at this site), and methylation level (the ratio of cytosine sequencing depth to sequencing depth at the cytosine site) on the Watson strand and Crick strand for each chromosome to the users. Annotation files in the BED format make it easy to observe the methylation distributions using an IGV or UCSC browser.

Downloads

Source code of BS-RNA and a test data (simulated RNA bisulfite sequencing data set of human) is provided and could be downloaded from http://bs-rna.big.ac.cn. It is a paired-end dataset in FastQ format (Phred33). The length of read is 100bp.

The command for test the demo data could be like this:

```
BS-RNA_v0.1 --perlDir script --reads1 test_T-rich.fq --reads2 test_A-rich.fq --gene Homo_sapiens.GRCh37.75.gtf --rawRef hg19_ref --specBuild spec1.txt --specHisat spec2.txt --pathToPython /…/python2.7.8/bin/ --pathToHISAT /.../hisat-0.1.6 --pathToSamtools /…/samtools-0.1.16/ --outDir /…/demo_result
```

It will take about 4.5 hours to complete all the analysis for this test data (including indexing the reference genome sequences).

Please refer to the Manual for any questions.

Installation

BS-RNA is written in Perl and is executed from the command line in LINUX system. To install BS-RNA simply copy the BS-RNA_v0.1.tar.gz file into a BS-RNA installation folder and extract all the files by typing:

```
tar xzf BS-RNA_v0.1.tar.gz
```

BS-RNA requires a working of Perl, Python (at least Python2.7.8), HISAT (at least HISAT-0.1.6) and SAMtools (at least SAMtools-1.0). Therefore it is a requirement that they are installed on your machine. BS-RNA will assume that these software are all in the working path unless their paths are specified manually. Furthermore bowtie2 should also be in the working path as HISAT uses the bowtie2 implementation to handle most of the operations on the FM index.

Manual

First you need download the reference genome sequences files of your concerned species and place them in a folder. Only single-entry files are supported. BS-RNA supports reference genome sequences in FastA format. The name begin with "chr" and the only allowed file extension is .fa. Secondly a gene model annotation file also need to be downloaded, which should be in GTF format.

Furthermore, two configure files are also needed for indexing the reference genome sequences and mapping the RNA sequencing data to the reference genome sequences if the user want to custom the corresponding parameters. An instruction on how to generate the configure file for hisat-build indexer or hisat could be found in the downloaded package. Each option should be specified in

one single line.

Usage: BS-RNA_v1.0 [Options]

Options:

--perlDir	String	Full path of the perl scripts
--reads1	String	Input T-rich reads file
--reads2	String	Input A-rich reads file
--gene	String	Supply BS-RNA with a set of gene model annotations, a GTF format file
--rawRef	String	Directory of raw reference genome sequences
--convertRef	String	Directory of converted reference genome sequences
--pathToPython	String	Full path </.../.../> to the Python installation on your system If not specified it is assumed that Python is in the PATH
--pathToHISAT	String	Full path </.../.../> to the HISAT installation on your system If not specified it is assumed that HISAT is in the PATH
--pathToSAMtools	String	Full path </…/.../> to the SAMtools installation on your system If not specified it is assumed that SAMtools is in the PATH
--phred64	String	Qualities are ASCII chars equal to the Phred quality plus 64

		"off" means Qualities are ASCII chars equal to the Phred quality plus 33. Default: off
--specBuild	File	Configure file for hisat-build indexer
--specHisat	File	Configure file for hisat
--outDir	String	Result output directory
--h or help		Display this message

A typical command for analyzing paired-end RBS-seq data is as follows:

```
BS-RNA_v0.1 --perlDir script --reads1 test_T-rich.fq --reads2 test_A-rich.fq --gene Homo_sapiens.GRCh37.75.gtf --rawRef hg19_ref --specBuild spec1.txt --specHisat spec2.txt --pathToPython /…/python2.7.8/bin/ --pathToHISAT /.../hisat-0.1.6 --pathToSamtools /…/samtools-0.1.16/ --outDir /…/demo_result
```

While for a single-end T-rich reads file is like this:

```
BS-RNA_v0.1 --perlDir script --reads1 test_T-rich.fq --gene Homo_sapiens.GRCh37.75.gtf --rawRef hg19_ref --specBuild spec1.txt --specHisat spec2.txt --pathToPython /…/python2.7.8/bin/ --pathToHISAT /.../hisat-0.1.6 --pathToSamtools /…/samtools-0.1.16/ --outDir /…/demo_result
```

Or for a single-end A-rich reads file:

```
BS-RNA_v0.1 --perlDir script --reads2 test_A-rich.fq --gene Homo_sapiens.GRCh37.75.gtf --rawRef hg19_ref --specBuild spec1.txt --specHisat spec2.txt --pathToPython /…/python2.7.8/bin/ --pathToHISAT /.../hisat-0.1.6 --pathToSamtools /…/samtools-0.1.16/ --outDir /…/demo_result
```

If the reference genome sequences have been converted in the previous analysis, please skip this step by adding this option to save time: “--convertRef path_of_converted_reference_genome”. In this situation, BS-RNA generates three folders in the specified output directory:

Map: contains mapping result file in SAM format and another file with spliced sites.

Filter: contains filtered mapping result file in SAM format and a statistic file called "filter_mapping.sam.maprate" containing the following information:

total: total reads number

map: mapped reads number

uniq: uniq mapped reads number

cor: correctly mapping on a corresponding strand reads number

used%: percent of correctly mapping on a corresponding strand reads

ps. The reads are mapped to the converted reference genome sequences, therefore the chromosome present in the SAM file contain "C-T" (represent the chromosome which convert all cytosines to thymines) or "G-A" (represent the chromosome which convert all guanines to adenines).

Level: contains bed files, which presents the following information for each covered cytosine site:

chr: name of the chromosome

start: cytosine chromosomal coordinates (0-based)

end: cytosine chromosomal coordinates (1-based)

strand: "+" means forward strand and "-" means crick strand

mCtype: methylation site type, one of the following[CG, CHG, CHH]

depth: total number of reads mapped to the cytosine site

mCdep: total number of reads that supported a methylated cytosine at this position

level: methylation level at the cytosine position

If the "--convertRef" option is not specified, an extra folder named "ref_all_C-T_G-A" will also be created in the output directory. This folder contains the concatenated raw genomce sequences and converted genome sequences in FastA format as well as the corresponding bowtie2 indexed files.