Computational Biology and Chemistry 65 (2016) 173–177

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research article

BS-RNA: An efficient mapping and annotation tool for RNA bisulfite sequencing data

Fang Liang[a], Lili Hao[a], Jinyue Wang[b,c], Shuo Shi[b,c], Jingfa Xiao[a,b,*], Rujiao Li[a,*]

[a] BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[b] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[c] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 31 August 2016
Accepted 7 September 2016
Available online 9 September 2016

Keywords:
RNA cytosine methylation
Bisulfite sequencing
Mapping software
Annotation tool

ABSTRACT

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 are only a few tools can directly deal with RNA bisulfite sequencing data efficiently. Herein, we developed a specialized tool BS-RNA, which can analyze cytosine methylation of RNA based on bisulfite sequencing data and support both paired-end and single-end sequencing reads from directional bisulfite libraries. For paired-end reads, simply removing the biased positions from the 5′ end may result in "dovetailing" reads, where one or both reads seem to extend past the start of the mate read. BS-RNA could map "dovetailing" reads successfully. The annotation result of BS-RNA is exported 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 on both Watson and Crick strands. BS-RNA is an efficient, specialized and highly automated mapping and annotation tool for RNA bisulfite sequencing data. It performs better than the existing program in terms of accuracy and efficiency. BS-RNA is developed by Perl language and the source code of this tool is freely available from the website: http://bs-rna.big.ac.cn.

1. Introduction

RNA methylation is a post-transcriptional modification, which plays a significant role as one of the epigenetic regulations. There are several types of methylation modifications identified in eukaryotes, such as N^7-methylguanosine (m^7G), N^6-methyl-2′-*O*-methyladenosine (m^6A_m), 2′-*O*-methylation (N_m), N^6-methyladenosine (m^6A), 5-methylcytosine (m^5C) and 5-hydroxymethylcytosine (hm^5C) (Liu and Jia 2014; Miao et al., 2016). m^5C of RNA plays a critical role in eukaryotes. A recent study suggests that m^5C of tRNA and rRNA is generally conserved across plantae, while animals may have independent evolution of organelle tRNA methylation which have function in regulating translation of protein (Burgess et al., 2015). NSUN2 encodes tRNA methyltransferase and also affects the function of mRNA (Squires et al., 2012; Khoddami and Cairns 2013). Mutation in NSUN2 could cause autosomal recessive intellectual disability and Dubowitz-like syndrome (Abbasi-Moheb et al., 2012; Khan et al., 2012; Martinez et al., 2012). Enrichment of m^5C on mRNA translation start site implied that there is the potential relationship between RNA cytosine methylation and protein translation (Zhang et al., 2012). m^5C bisulfite sequencing of the whole transcriptome of HeLa showed that m^5C is widely modified in both mRNA and ncRNA (Squires et al., 2012). This study indicated that previous reports about the tiny amount of m^5C in mRNA are mainly due to the limitation of the detection methods. Presently, bisulfite sequencing technology has become an effective method for studying RNA cytosine methylation in both eukaryotes and prokaryotes (Edelheit et al., 2013). However, there are only a few specialized tools could map RNA bisulfite sequencing data to reference sequences efficiently. Due to the lack of efficient mapping tools, most researchers map RNA bisulfite sequencing reads to either confirmed transcripts or self-built junctions using DNA alignment tools. The methods of mapping sequencing reads to confirmed transcript positions can only focus on known transcripts, while mapping reads to self-built junctions is difficult for paired-end sequencing reads with internal insert fragment.

Abbreviations: BS, bisulfite sequencing; mRNA, messenger RNA; BED, browser extensible data; SAM, sequence alignment/map; RAM, random access memory.

* Corresponding authors.
E-mail addresses: xiaojf@big.ac.cn (J. Xiao), lirj@big.ac.cn (R. Li).

http://dx.doi.org/10.1016/j.compbiolchem.2016.09.003

meRanTK (Rieder et al., 2015) is the only one specialized tool aiming to address the demands of RNA bisulfite sequencing data analysis currently. It aligns bisulfite sequencing reads to the reference genome with the aid of software TopHat (Kim et al., 2013) or STAR (Dobin et al., 2013) after base conversion. But meRanTK has limited ability to map "dovetailing" reads and its analysis process is much time-consuming. The biological impact of m^5C RNA remains largely unknown, partly owing to a shortage of efficient analysis tools. Hence, we developed a highly efficient and accurate tool, BS-RNA, which can map RNA bisulfite sequencing data to reference genome and simultaneously annotate coordinates, sequence types, reference sequencing depths, cytosine sequencing depths, and methylation levels of cytosines covered by sequencing reads. BS-RNA only supports bisulfite sequencing RNA data from the directional libraries, however it can process both single-end and paired-end data. BS-RNA can identify methylation positions and annotate their methylation levels based on traditional RNA mapping considering both accuracy and efficiency.

2. Methods

2.1. Software dependencies and system requirements

The widely used RNA-seq sequencing data alignment programs such as TopHat2 (Kim et al., 2013) and GSNAP (Wu and Nacu, 2010) may take several days to process a single RNA-seq experiment. Another program STAR (Dobin et al., 2013) processes faster than most other tools at the expense of large memory requirements. HISAT (Kim et al., 2015) (hierarchical indexing for spliced alignment of transcripts) employs two types of indexes for alignment and solves the challenging spliced-alignment problems using strategies specially designed for different read types. It is a fast system with modest memory cost currently. HISAT2 is an enhanced version of HISAT and more new features are achieved in this version, such as improving of alignment accuracy by preventing reads from being mapped to pseudogenes, updating python script to extract SNPs and haplotypes from VCF files, etc. Also some bugs are fixed, such as a bug caused reads to map beyond reference sequences and a deadlock issue that happened very rarely. Here we implemented HISAT2 in BS-RNA to improve the mapping speed and quality of RNA bisulfite sequencing data.

BS-RNA was developed by Perl programming language and is executed from the command line in LINUX system. It requires a working of Perl, Python, HISAT2, SAMtools (Li et al., 2009) and Bowtie2 (Langmead and Salzberg, 2012). The source code of this tool is freely available from the website: http://bs-rna.big.ac.cn A detailed user manual is also provided.

2.2. Workflow of BS-RNA

The BS-RNA process (Figs. 1 and 2) includes three main steps: pre-treatment, mapping, and annotation. The first step is the pre-treatment of reference genome sequences, sequencing data, and gene annotation file: (1) the reference genome sequence is converted twice in parallel as follows: (A) cytosines are replaced by thymines and (B) guanines are replaced by adenines. The conversion of the genome sequence is only need to be performed once, which means it could be reused for all the following analysis, which depends on the same reference genome sequence. (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. 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 the gene annotation file, respectively.

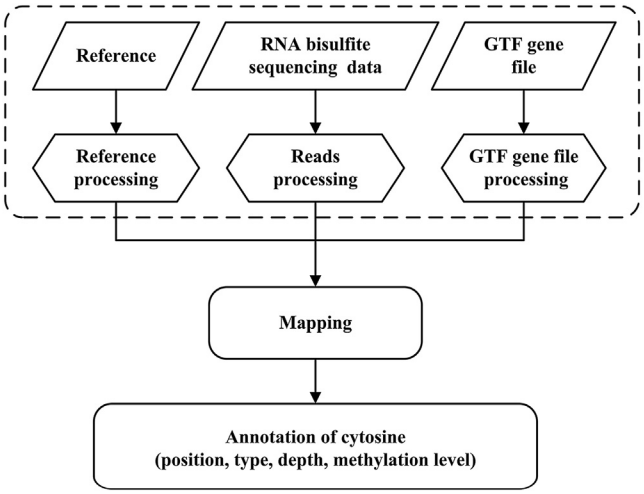

Fig. 1. Flowchart of mapping and annotation of BS-RNA. Three main steps: pre-treatment, mapping and annotation of cytosine.

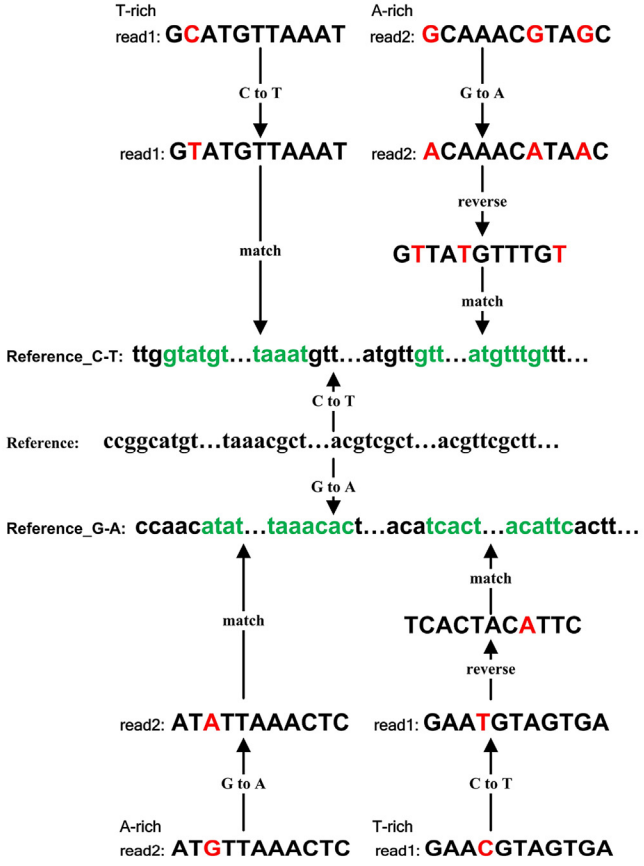

Fig. 2. Mapping principle of BS-RNA. T-rich reads mapped to reference sequence with Cs converted to Ts or to reverse-complement of reference sequence with Gs converted to As; A-rich reads mapped to reference sequence with Gs converted to As or to reverse complements of reference sequence with Cs converted to Ts.

Next, the HISAT2 program is invoked by BS-RNA to build alternative splicing according to the modified annotation gene file and align pre-processed reads to the converted reference genome sequence. BS-RNA filters out two types of reads which 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