Bioinformatics, 37(21), 2021, 3934–3935
doi: 10.1093/bioinformatics/btab663
Advance Access Publication Date: 2 October 2021
Applications Note

Sequence analysis

RLM: fast and simplified extraction of read-level methylation metrics from bisulfite sequencing data

Sara Hetzel[1], **Pay Giesselmann**[1], **Knut Reinert**[2], **Alexander Meissner**[1,3,4,5,†] **and Helene Kretzmer**[1]

[1]Department of Genome Regulation, Max Planck Institute for Molecular Genetics, Berlin, Germany, [2]Department of Mathematics and Informatics, Freie Universität, Berlin, Germany, [3]Department of Biology, Chemistry and Pharmacy, Freie Universität, Berlin, Germany, [4]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA and [5]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

*To whom correspondence should be addressed.

Associate Editor: Can Alkan

Received on May 26, 2021; revised on August 11, 2021; editorial decision on September 7, 2021

Abstract

Summary: Bisulfite sequencing data provide value beyond the straightforward methylation assessment by analyzing single-read patterns. Over the past years, various metrics have been established to explore this layer of information. However, limited compatibility with alignment tools, reference genomes or the measurements they provide present a bottleneck for most groups to routinely perform read-level analysis. To address this, we developed RLM, a fast and scalable tool for the computation of several frequently used read-level methylation statistics. RLM supports standard alignment tools, works independently of the reference genome and handles most sequencing experiment designs. RLM can process large input files with a billion reads in just a few hours on common workstations.

Availability and implementation: https://github.com/sarahet/RLM.

Contact: meissner@molgen.mpg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Bisulfite sequencing experiments are the gold standard to measure DNA methylation at single-CpG resolution (Frommer *et al.*, 1992). Besides the average CpG methylation level across a cell population, each read contains information about the methylation of a single molecule present within a cell (Scherer *et al.*, 2020). This information can be used when analyzing the heterogeneity of a cell population or comparing samples of different conditions such as healthy and tumor. Different metrics have been established in order to quantify heterogeneity within and across cells based on single-read methylation patterns. Measurements of population heterogeneity include methylation entropy and epipolymorphism, which are based on so-called epialleles, assessing the patterns of methylated and unmethylated CpGs that can occur in a 4-mer of consecutive CpGs spanned by the same reads (16 epialleles are possible for a single 4-mer) (Landan *et al.*, 2012; Xie *et al.*, 2011). Additionally, the heterogeneity within a single read can be classified as concordant or discordant, which can then be aggregated across reads as the percent of discordant reads (PDR), thus again measuring the heterogeneity across different cells (Landau *et al.*, 2014). Similarly, the number of transitions from methylated to unmethylated CpGs on a read can be used and aggregated per CpG to determine the level of heterogeneity (read transition score or RTS) (Charlton *et al.*, 2018).

So far, studies that have analyzed read-level DNA methylation heterogeneity often utilized custom scripts for this purpose (Landan *et al.*, 2012; Landau *et al.*, 2014; Xie *et al.*, 2011). Only few tools are available that implement one or more of such metrics; however, they are limited in their usability by requiring a specific alignment tool, reference genome or additional input such as the exact position of CpGs of interest (He *et al.*, 2013; Scherer *et al.*, 2020; Scott *et al.*, 2020).

Here, we present RLM, a fast and scalable tool that implements established and frequently used inter- and intramolecular metrics of DNA methylation at the read level from bisulfite sequencing experiments. RLM is applicable for any reference genome, a wide range of library protocols and works with input alignment files from multiple commonly used alignment tools. Additionally, it automatically accounts for potential errors and biases caused by sequencing artifacts, mapping quality and overlapping read pairs.

2 Implementation and features

RLM is a standalone C++-based tool implemented using SeqAn (Reinert *et al.*, 2017). As input, RLM accepts BAM or SAM files from the common bisulfite alignment tools BSMAP, BISMARK, segemehl and GEM (Krueger and Andrews, 2011; Otto *et al.*, 2012; Santiago et al., 2012; Xi and Li, 2009). RLM can run with single- or paired-end sequencing data from different protocols such as whole genome bisulfite sequencing and target enrichment approaches but also reduced representation bisulfite sequencing (RRBS). To account for the artificially introduced nucleotides in RRBS experiments, CpGs at the end of the first read (and beginning of the second read) can be omitted from the analysis. Generally, reads that do not represent a primary alignment, are polymerase chain reaction duplicates, fail the quality control or do not pass a user-defined mapping quality filter are discarded.

For single-end input, RLM streams across the records of the input file and extracts information about the methylation status of a CpG based on the corresponding reference genome and the BAM file tag that represents the origin and mapping orientation of each read (e.g. 'ZS' tag for BSMAP). Reads with sequencing errors at the position of a CpG, indels or reads that span less than three CpGs are discarded.

For paired-end sequencing experiments, reads are filtered analogous to single-end experiments. Both mates are processed independently except for overlapping mates, which are merged and processed as a single, contiguous read maximizing the information that can be extracted when using downstream measurements dependent on four consecutive CpGs such as entropy. Additionally, we avoid two quantifications of the same genomic fragment, which would bias the analysis of cell population heterogeneity. To enable this, reads are kept in memory until the mate has been read and are removed from memory afterwards. If one mate needs to be excluded from the analysis, the other read will be processed independently to improve coverage.

Depending on the research question of interest, RLM offers multiple outputs that can be requested separately or all at once by the user.

Single-read information: For every read with at least three CpGs, the methylation status for each CpG, the average methylation, the transition score and the discordance are reported. The transition score is defined as the number of transitions between methylated and unmethylated CpGs divided by the number of possible transitions. This file is mandatory output as the information collected here is used for the other output files.

RTS and PDR: For every CpG spanned by a user-defined minimum number of reads, the RTS and PDR across all reads spanning the CpG are reported. Additionally, the corresponding methylation rate based on the reads considered for the read-level measurements is provided.

Entropy and epipolymorphism: For every 4-mer of consecutive CpGs spanned by a user-defined minimum number of reads, the entropy and epipolymorphism are calculated and reported together with the average methylation rate of the complete 4-mer based on the reads considered for read-level metrics. Additionally, the frequencies for all epialleles leading to the respective metrics are provided.

To complement the reported scores, RLM ships a standalone R script that provides users with a report including summary statistics and figures. Additionally, the documentation contains guidelines for its interpretation and use cases. The runtime of RLM scales linearly with the input size for both paired- and single-end modes. Large datasets with up to one billion reads can be processed in few hours with modest memory requirements (detailed performance analysis and comparison with other existing tools in the Supplementary Data).

Acknowledgement

We thank R. Rahn and E. Seiler for helpful suggestions regarding the implementation.

Funding

This work has been supported by the Max Planck Society.

Conflict of Interest: none declared.

References

Charlton,J. *et al.* (2018) Global delay in nascent strand DNA methylation. *Nat. Struct. Mol. Biol.*, **25**, 327–332.

Frommer,M. *et al.* (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. USA*, **89**, 1827–1831.

He,J. *et al.* (2013) DMEAS: DNA methylation entropy analysis software. *Bioinformatics*, **29**, 2044–2045.

Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.

Landan,G. *et al.* (2012) Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.*, **44**, 1207–1214.

Landau,D.A. *et al.* (2014) Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell*, **26**, 813–825.

Otto,C. *et al.* (2012) Fast and sensitive mapping of bisulfite-treated sequencing data. *Bioinformatics*, **28**, 1698–1704.

Reinert,K. *et al.* (2017) The SeqAn C++ template library for efficient sequence analysis: a resource for programmers. *J. Biotechnol.*, **261**, 157–168.

Santiago,M.-S. *et al.* (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods*, **9**, 1185–1188.

Scherer,M. *et al.* (2020) Quantitative comparison of within-sample heterogeneity scores for DNA methylation data. *Nucleic Acids Res.*, **48**, e46.

Scott,C.A. *et al.* (2020) Identification of cell type-specific methylation signals in bulk whole genome bisulfite sequencing data. *Genome Biol.*, **21**, 156.

Xi,Y. and Li,W. (2009) BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, **10**, 232.

Xie,H. *et al.* (2011) Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Res.*, **39**, 4099–4108.