*Bioinformatics*, 31(12), 2015, 2040–2042
doi: 10.1093/bioinformatics/btv089
Advance Access Publication Date: 13 February 2015
Applications Note

OXFORD

Sequence analysis

# A trimming-and-retrieving alignment scheme for reduced representation bisulfite sequencing

**Xuefeng Wang[1,2,3,*], Xiaoqing Yu[4], Wei Zhu[3], W. Richard McCombie[5], Eric Antoniou[5], R. Scott Powers[5,6], Nicholas O. Davidson[7], Ellen Li[8] and Jennie Williams[1,*]**

[1]Department of Preventive Medicine, [2]Department of Biomedical Informatics, and [3]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794, [4]Department of Biostatistics, Yale University, New Haven, CT 06520, [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, [6]Department of Pathology, Stony Brook University, Stony Brook, NY 11794, [7]Department of Medicine, Washington University St Louis, St Louis, MO 63110 and [8]Department of Medicine, Stony Brook University, Stony Brook, NY 11794, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on November 3, 2014; revised on January 22, 2015; accepted on February 6, 2015

## Abstract

**Summary:** Currently available bisulfite sequencing tools frequently suffer from low mapping rates and low methylation calls, especially for data generated from the Illumina sequencer, NextSeq. Here, we introduce a sequential trimming-and-retrieving alignment approach for investigating DNA methylation patterns, which significantly improves the number of mapped reads and covered CpG sites. The method is implemented in an automated analysis toolkit for processing bisulfite sequencing reads.

**Availability and implementation:** http://mysbfiles.stonybrook.edu/~xuefenwang/software.html and https://github.com/xfwang/BStools.

**Contact:** xuefeng.wang@stonybrook.edu

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

## 1 Introduction

The advancement of next-generation sequencing technologies offers the opportunity to investigate the epigenetic modification on genome-wide scale and at single-base resolution (Krueger *et al.*, 2012). Reduced representation bisulfite sequencing (RRBS) has emerged as a popular tool for profiling DNA methylation at CpG rich regions. A major advantage of the RRBS method is its reproducibility over time and sample types. Compared with traditional enrichment based methods, RRBS is more applicable to formalin-fixed and paraffin-embedded clinical samples, due to its low DNA input requirement. However, there are considerable challenges when analyzing the reads from RRBS due to the complexity induced by the bisulfite conversion. Highly specialized analysis pipelines are needed—to target different data types and different platforms.

Bisulfite treatment converts all un-methylated cytosines into uracils, which causes a C-to-T change in forward strand reads. Consequently, the processing of RRBS library reads involves two steps: the alignment of converted reads and the calling of methylation level (i.e. the methylated/unmethylated ratio) on each site. The challenge in the mapping of converted reads to reference genome has been extensively studied and multiple methylation-aware alignment algorithms and associated pipelines have been proposed, such as BRAT (Harris *et al.*, 2010, 2012; Krueger and Andrews, 2011; Xi *et al.*, 2012). In this manuscript, we specifically consider the issue of low alignment performance caused by variable sequencing quality. As shown in the Supplementary Figure S1, we often see low read quality for the first few bases of a read, a pattern consistent across all samples sequenced by the Illumina NextSeq platform. An important feature of reads from an RRBS library is that most sequences have Gs at the second and third bases. Consequently low read quality will thus be observed at positions 2 and 3 on NextSeq 500, because ‘G’ bases do not have a dye attached.

The easiest remedy appears to be trimming the first few low-quality bases from each read to improve the subsequent alignment process. The first base of every sequence, however, is particularly informative in indicating methylation status at one CpG site—which should predominantly be a ‘C’ if it was methylated and a ‘T’ if not methylated. Another possible solution is to mask the low-quality bases by replacing their reads with ‘N’s. We tested this solution but did not observe any improvement compared with the mapping based on raw reads. This is understandable because the ambiguity of base calls still exists. Such issues can be largely alleviated when libraries are sequenced on a HiSeq platform. The downside is that one would obtain even fewer reads per library per sequencing run. Although this may be compensated by decreasing multiplexing (the number of individual libraries sequenced at the same time), the sequencing costs per library increase accordingly. Furthermore, many smaller laboratories do not have easy access to HiSeqs, while a desktop NextSeq is both cheaper and easier to operate. NextSeq is also widely used in large sequencing facilities due to its fast turnaround time. Therefore, it is critical to adapt the current RRBS processing pipeline to incorporate this issue.

## 2 Methods and implementation

In this manuscript, we propose a trimming-and-retrieving scheme for accurate mapping of bisulfite converted reads from RRBS libraries. The workflow of this scheme is illustrated in Figure 1. The proposed scheme includes two major steps. An initial alignment step is first performed based on the trimmed reads. The trimmed bases are then retrieved and one alignment is kept only when the first base is a C in the reference. The methylation percentage calling is performed by counting methylated and unmethylated bases (and their ratio) on each site that is a C in the reference genome. This strategy ensures both accurate bisulfite-converted read alignment and methylation calling.

We have developed a robust and automated analysis toolkit BStools to incorporate the proposed scheme. BStools runs on Unix platform and is implemented with a combination of Perl, R and Unix shell scripts. This toolkit is built on top of a previously proposed pipeline (Sun *et al.*, 2013) for bisulfite-treated methylation sequencing quality assessment, in which BRAT (Harris *et al.*, 2010) was used as the upstream aligner. A computationally efficient script is provided for fast retrieval of trimmed reads. Note that the new toolkit not only incorporates quality control and alignment programs but also includes functions to perform methylation calls and to analyze the subsequent differentially methylated regions (DMRs). Some useful graphical functions are also provided for convenient visualization of methylation calls and DMRs. This package provides a seamless integration of multiple functions of bisulfite sequencing analysis by fine-tuning the input and output file formats and objects. Here, we provide an example of the potential of the proposed scheme. We analyzed 45 samples (including 25 tumor and 23 normal samples) with three mapping methods based on: raw reads, trimmed reads and retrieved reads, respectively. As shown in Figure 1B, the alignment performance of the trimming-and-retrieving scheme is greatly improved compared with the mapping results based on the raw reads. Compared with the raw read mapping, the proportion of uniquely mapped reads increased from <20% to around 60% on the sample tested. Note that there is a slight decrease in the alignment rate compared with the trimmed method, because the false-called reads (i.e. the first base is not C in the reference) are removed in the retrieving step. These false-called reads may arise for many reasons, for example, incorrect enzyme digestion, sequencing and alignment errors. However, the number of covered CpG site is significantly improved as compared with the simple trimming method (around 30% increase) when using the trimming-and-retrieving method. This demonstrates that our method can boost not only sensitivity but also specificity of methylation calls.

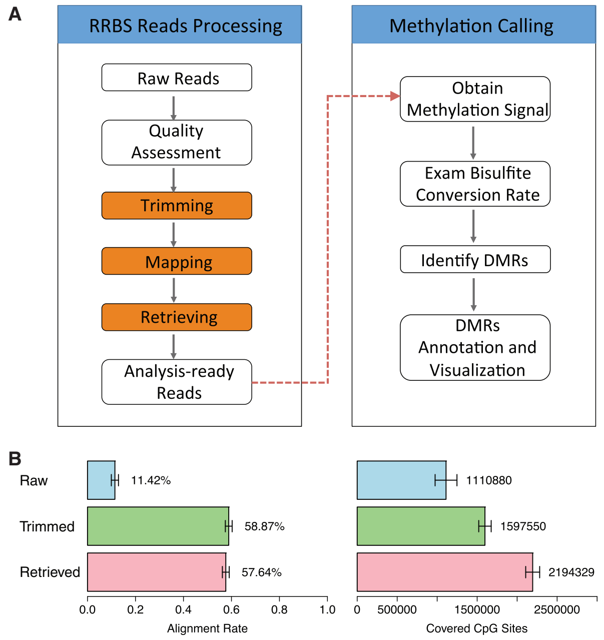

**Fig. 1.** RRBS analysis pipeline and results. (**A**) Schematic view of the RRBS sequencing analysis workflow. Please refer to the Supplementary Information for a detailed description of the trimming-and-retrieving method. (**B**) The mapping and methylation calling results from analyzing 45 samples sequenced by the NextSeq platform based on the proposed scheme

Although the processing pipeline introduced is mainly used to address issues with the NextSeq platform, the same scheme can be adapted to fit other sequencing platforms with low read quality bases at the 5′ end. One limitation of this pipeline is that it relies on directional sequencing, which is widespread way of RRBS. Users need to take extra care when analyzing the non-directional or paired-end RRBS libraries.

## 3 Conclusions

The new alignment method proposed and the associated RRBS anlaysis toolkit provides a timely tool for the accurate mapping and calling of RRBS, especially with NextSeq data.

## Acknowledgements

The authors would like to thank the Alvin J. Siteman Cancer Center at Washington University School of Medicine and the WU DDRCC Biobank Core, which provided colon cancer tissues.

## Funding

The work was supported in part by the Simons Foundation to E.L., a Cancer Center Grant (CA045508) from the NCI to CSHL, NCI U01 Grant (CA168409) to R.S.P., an NCI Cancer Center Support Grant #P30 CA91842 to Siteman Cancer Center, #P30 DK 52574 to N.O.D. and R01CA140487 to J.W. X.Y. was supported by Yale Comprehensive Cancer Center (P30 CA16359 from NIH).

*Conflict of Interest*: none declared.

## References

Harris,E.Y. *et al.* (2010) BRAT: bisulfite-treated reads analysis tool. *Bioinformatics,* **26**, 572–573.

Harris,E.Y. *et al.* (2012) BRAT-BW: efficient and accurate mapping of bisulfite-treated reads. *Bioinformatics*, **28**, 1795–1796.

Krueger,F. *et al.* (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat. Methods*, **9**, 145–151.

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

Sun,S. *et al.* (2013) MethyQA: a pipeline for bisulfite-treated methylation sequencing quality assessment. *BMC Bioinformatics*, **14**, 259.

Xi,Y. *et al.* (2012) RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. *Bioinformatics*, **28**, 430–432.