Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw226
Advance Access Publication Date: 23 February 2016
Applications Note

OXFORD

Sequence analysis

BRAT-nova: fast and accurate mapping of bisulfite-treated reads

Elena Y. Harris[1,†,*], Rachid Ounit[2,†] and Stefano Lonardi[2]

[1]Department of Computer Science, California State University, Chico, CA 95929, USA and [2]Department of Computer Science and Eng, University of California, Riverside, CA 92521, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
Associate Editor: Inanc Birol

Received on January 5, 2016; revised on March 28, 2016; accepted on April 15, 2016

Abstract

Summary: In response to increasing amounts of sequencing data, faster and faster aligners need to become available. Here, we introduce BRAT-nova, a completely rewritten and improved implementation of the mapping tool BRAT-BW for bisulfite-treated reads (BS-Seq). BRAT-nova is very fast and accurate. On the human genome, BRAT-nova is 2–7 times faster than state-of-the-art aligners, while maintaining the same percentage of uniquely mapped reads and space usage. On synthetic reads, BRAT-nova is 2–8 times faster than state-of-the-art aligners while maintaining similar mapping accuracy, methylation call accuracy, methylation level accuracy and space efficiency.
Availability and implementation: The software is available in the public domain at http://compbio.cs.ucr.edu/brat/
Contact: elenah@cs.ucr.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Whole genome bisulfite sequencing (BS-Seq), allows genome-wide studies of DNA methylation at single base pair resolution. Sodium bisulfite treatment of DNA (Frommer *et al.*, 1992) followed by PCR and sequencing enables the detection of the methylation status of individual cytosines. The first step in BS-Seq analysis is to map bisulfite-treated reads to the reference genome. Since sodium bisulfite converts unmethylated cytosines to thymines, the criteria for mapping BS-Seq reads requires one to allow a T in a read to match a C in the reference genome.

Various indexes have been used to accelerate the mapping of BS-Seq reads to a reference genome. In practice, tools that employ the FM-index (Ferragina and Manzini, 2000) achieve the best balance between space requirements, mapping accuracy and speed. Among these, Bismark (Krueger and Andrews, 2011) and BS-Seeker/BS-Seeker2 (Chen *et al.*, 2010; Guo *et al.*, 2013) are arguably the most commonly used tools for BS-Seq data.

Here, we introduce BRAT-nova, a completely rewritten and improved implementation of BRAT-BW (Harris *et al.*, 2012) for aligning BS-Seq reads. BRAT-nova employs a novel space-efficient representation of the genome and supports local alignment by allowing one indel per read. BRAT-nova is 2–11 times faster on the human genome than Bismark, BS-Seeker2 and BSMAP (Xi and Li, 2009) while demonstrating comparable results in terms of the proportion of mapped unique reads, mapping accuracy, methylation level and methylation call accuracy and RAM usage.

2 Methods

For a directional library (i.e. when the chosen BS-Seq protocol produces reads only from the two original strands), in order to allow a T in a read to map to a C in a genome, both Bismark and BS-Seeker2 use two FM-indexes built from positive strand of the reference genome: in the first, Cs are converted to Ts, and in the second, Gs are converted to As. During the mapping phase, reads with Cs converted to Ts are mapped to the first index, while the reverse-complement of the reads with Gs replaced by As are mapped to the second index. Thus, in Bismark and BS-Seeker2 each read requires two alignments to two distinct indexes. BRAT-nova instead uses a single FM-index built on the concatenation of the positive and

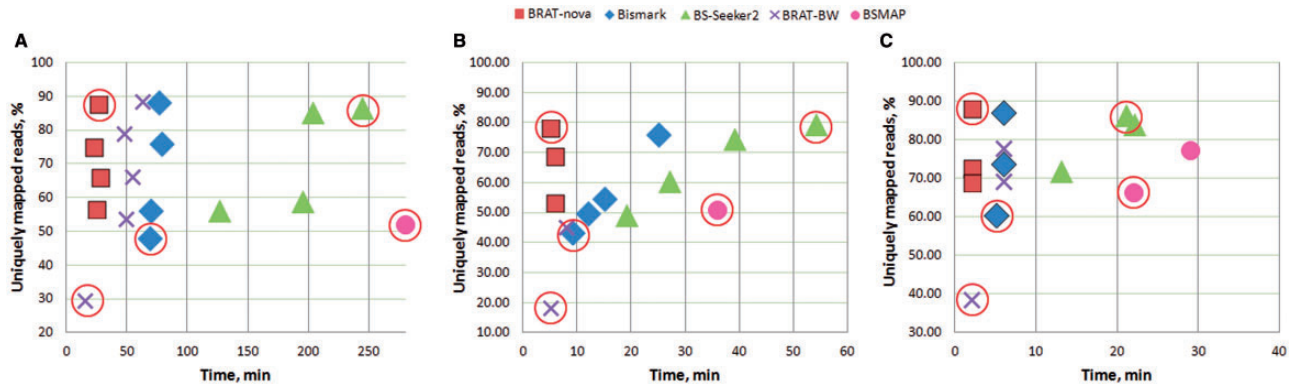

Fig. 1. Percentage of uniquely mapped reads as a function of running time, for several choices of the parameters on three human datasets. A description of the parameters used in these experiments is provided in Supplemental Tables 1-3. (A) ~10.6M single-end 101bp real *SRR306435* reads, (B) 1M paired-end 101bp real *SRR306435* reads, (C) 1M real single-end 76bp real *SRR306421* reads; circles indicate default parameters. BRAT-nova shows similar results in terms of the range of the percentage of uniquely mapped reads as other tools, and is faster. Bismark shows speed comparable to BRAT-nova on default parameters, but at the cost of mapping fewer unique reads

negative strands with Cs converted to Ts, and each read with Cs converted to Ts is aligned only once to a single index, which speeds up the mapping step. For large genomes such as the human genome, the concatenation of positive and negative strands is longer than 2^{32} bases, which is the current genome size limit for BRAT-BW. To handle this, BRAT-nova uses an additional bit per base to track the strand identity.

After mapping bisulfite-treated reads, BRAT-nova determines the methylation level of each cytosine. The methylation level of a cytosine is defined as the fraction of mapped reads that have a C (i.e. methylated, thus not converted to a T) in that genomic location, to the total number of reads mapped to that location (i.e. with a C or a T).

BRAT-nova supports a single variable-length indel in the middle of a read by using two exactly matched seeds surrounding the indel followed by the linear time dynamic programming algorithm described in the Supplemental Notes.

3 Experimental results and discussion

To assess the performance of BRAT-nova, we benchmarked it against Bismark (that supports end-to-end alignment with indels), BS-Seeker-2 (that supports local alignment with indels), BSMAP (supports end-to-end alignment with one gap) and BRAT-BW (that supports end-to-end alignment with no indels). We evaluated all tools using real human genome reads from dataset SRR306435 (Molaro *et al.*, 2011) and dataset SRR306421 (Hodges *et al.*, 2011), as well as synthetic reads generated from human genome GRCh38. Our experiments measured mapping efficiency, mapping accuracy, methylation call accuracy, methylation level accuracy, running time and space usage. Below, we report the results of our benchmarking; see Supplemental Notes for a full description of the benchmarking methods.

On real reads, we ran each tool using various parameter settings and measured the percentage of uniquely mapped reads (i.e. reads mapped with the highest score to a single location), running time and RAM usage. Experimental results are reported in Figure 1 (see Supplemental Tools 1–3 for more details). The percentage of uniquely mapped reads and running time can vary significantly depending on the parameter settings. In terms of uniquely mapped reads, BRAT-nova showed a comparable range of performance as other tools, but it was 2–11 times faster (Supplemental Fig. S1).

On synthetic reads, we measured two types of mapping accuracy defined as the ratio of uniquely mapped reads aligned within 50bp and 0bp of the original positions (same chromosome, same strand) to the total number of uniquely mapped reads. Supplemental Figure S2 and Supplemental Tables S4–S6 report the results of the mapping accuracy tests. Again, the mapping accuracy can vary significantly depending on the parameter settings. All tools showed a higher mapping accuracy with stricter parameters at the expense of a smaller percentage of reads mapped. In these experiments, BRAT-nova showed comparable results with all other tools. Next, we measured the performance of the tools in terms of methylation call accuracy and methylation level accuracy. Methylation call accuracy was measured as the ratio of the cytosines whose methylation status was correctly identified (methylated or not methylated) to the total number of the cytosines covered by at least ten reads. A cytosine was considered to be methylated if it had a methylation level of at least 0.5, and unmethylated otherwise. To calculate the methylation level accuracy, we used a randomized analysis (see Supplemental Notes). Supplemental Figures S3 and S4 and Supplemental Tables S7 and S8 show the results for methylation call and methylation level accuracy tests; BRAT-nova showed comparable results to the other tools. With strict parameters, all tools mapped fewer reads with higher mapping accuracy. However, despite decreased mapping accuracy with loose parameters, methylation level accuracy was 2–4% higher for all tools compared to strict parameters. In these experiments, BRAT-nova was 2–8 times faster than Bismark, BS-Seeker2, BSMAP and BRAT-BW.

Funding

This project was supported in part by NSF IIS-1302134.

Conflict of Interest: none declared.

References

Chen,P.Y. *et al.* (2010) BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics*, **11**, 203.

Ferragina,P. and Manzini,G. (2000) Opportunistic data structures with applications. In: *Proceedings of IEEE Foundation of Computer Science*.

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.

Guo,W. *et al.* (2013) BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics*, **14**, 774.

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

Hodges,E. *et al.* (2011) Directional DNA changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol. Cell*, **44**, 1–12.

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

Molaro,A. *et al.* (2011) Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell*, **146**, 1029–1041.

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