Sequence analysis

RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing

Yuanxin Xi1,†, Christoph Bock2,3,4,†, Fabian Müller2,3,4, Deqiang Sun1, Alexander Meissner4,5, and Wei Li4

1Division of Biostatistics, Dan L Duncan Cancer Center and Department of Molecular and Cellular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, 2Broad Institute, Cambridge, MA 02142, 3Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA and
4Max Planck Institute for Informatics, Saarbrücken, Germany

ABSTRACT

Summary: Reduced representation bisulfite sequencing (RRBS) is a powerful yet cost-efficient method for studying DNA methylation on a genomic scale. RRBS involves restriction-enzyme digestion, bisulfite conversion and size selection, resulting in DNA sequencing data that require special bioinformatic handling. Here, we describe RRBSMAP, a short-read alignment tool that is designed for handling RRBS data in a user-friendly and scalable way. RRBSMAP uses wildcard alignment and avoids the need for any preprocessing or post-processing steps. We benchmarked RRBSMAP against a well-validated MAQ-based pipeline for RRBS read alignment and observed similar accuracy but much improved runtime performance, easier handling and better scaling to large sample sets. In summary, RRBSMAP removes bioinformatic hurdles and reduces the computational burden of large-scale epigenome association studies performed with RRBS.

Availability: http://rrbsmap.computational-epigenetics.org/
http://code.google.com/p/bsmap/

Contact: wli@bcm.tmc.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on September 2, 2011; revised on November 18, 2011; accepted on November 29, 2011

1 INTRODUCTION

DNA methylation is an important mechanism of epigenetic regulation in development and disease. Many methods for DNA methylation profiling have been developed, but only bisulfite sequencing gives rise to comprehensive DNA methylation maps at single-base pair resolution ( Laird, 2010). Bisulfite treatment converts unmethylated cytosines (C) into uracils, which gives rise to C-to-T polymorphisms after subsequent Polymerase Chain Reaction amplification, while leaving methylated cytosines unchanged. By aligning and comparing bisulfite sequencing reads to the genomic DNA sequence, it is thus possible to infer base pair-resolution DNA methylation patterns.

Bisulfite-converted DNA can be subjected to whole-genome resequencing, giving rise to comprehensive methylomes (Laurent et al., 2010; Lister et al., 2009). However, this procedure requires deep sequencing of entire genomes, rendering it expensive in terms of sequencing cost as well as the required amount of input DNA. Reduced representation bisulfite sequencing (RRBS) addresses these limitations (Gu et al., 2010; Meissner 2005, 2008) and provides a complementary technology to whole-genome bisulfite sequencing. In RRBS, DNA is first digested with a restriction enzyme that includes a CpG in its recognition site and cuts DNA independent of the methylation status of this CpG. Subsequently, small DNA fragments are size selected to enrich for CpG-rich genomic regions (which are often associated with epigenetic regulation) and subjected to high-throughput bisulfite sequencing. By concentrating on a small but informative portion of the genome, RRBS provides high-sequencing depth at affordable cost, making RRBS well suited for detecting subtle differences in large patient cohorts. Furthermore, RRBS is readily applicable to small amounts of DNA and formal-fixed, paraffin-embedded samples (Gu et al., 2010).

RRBS is actively used by a number of groups (Baranzini et al., 2010; Bock et al., 2011; Gertz et al., 2011; Smallwood et al., 2011; Stein et al., 2011). To our knowledge, no alignment tools have been published that are tailored to RRBS data. Although some existing tools for bisulfite sequencing data (Bock et al., 2005; Chen et al., 2010; Krueger and Andrews, 2011; Li et al., 2008; Lutsik et al., 2011; Smith et al., 2009; Xi and Li, 2009) can also be applied to RRBS, doing so requires custom pre- and post-processing, thus limiting the accessibility of RRBS for researchers with limited bioinformatic support. Here we describe RRBSMAP, a short-read alignment tool that is specifically tailored to RRBS. We show that RRBSMAP provides a major advance in terms of runtime performance and usability compared to a well-validated MAQ-based pipeline, while maintaining high mapping accuracy.

2 IMPLEMENTATION

In RRBS, only those DNA fragments that start and end with a restriction digestion site and fall within the range of experimental size selection are subjected to bisulfite sequencing. For example,MspI, the restriction enzyme most commonly used in RRBS, cuts the DNA at CGG sites in a way that produces DNA fragments starting with CGG and ending with CCG after reestablishment
of double-stranded DNA (Meissner et al., 2008). Depending on the methylation status of the Cs, several types of reads can be observed after bisulfite sequencing. RRBSMAP utilizes prior knowledge about the distribution of restriction digestion sites in the genome to significantly improve runtime performance and memory efficiency compared to whole-genome bisulfite alignment. Rather than building an alignment seed table for the entire genome, RRBSMAP indexes only those genomic regions that are compatible with the RRBS protocol parameters (Supplementary Figure S1A).

The bisulfite alignment is performed using a wildcard alignment algorithm originally developed for BSMAQ (Xi and Li, 2009). The algorithm implements a bitwise mask to account for C-to-T conversions introduced by bisulfite treatment (Supplementary Figure S1B). It avoids aligning Ts in the reference to Cs in the reads, which is a common source for bias when converting all Cs to Ts first and then performing the three-nucleotide alignment (Supplementary Table S1 and Supplementary Figure S2). (Note that the post-processing step by which three-nucleotide bisulfite aligners remove incorrect alignments cannot completely resolve this issue because spurious hits could already have masked correct hits during the alignment step.)

RRBSMAP implements many features that facilitate the routine handling of large RRBS datasets. First, RRBSMAP operates directly on BAM/SAM files as input and output, and it does not require any preprocessing or post-processing steps or temporary files, which makes it straightforward to run RRBSMAP as a part of large-scale sequencing pipelines. Second, RRBSMAP provides flexible options for trimming adapter sequences and low-quality nucleotides prior to alignment. Third, RRBSMAP allows the user to choose alternative restriction enzymes beyond the defaultMspI, providing flexibility to explore new bisulfite sequencing protocols. Fourth, RRBSMAP efficiently aligns both single-end and paired-end sequencing with flexible read lengths. Fifth, RRBSMAP fully supports multicore and multiprocessor hardware and lets the user specify how many threads should be run in parallel.

3 RESULTS

We compare RRBSMAP with a MAQ-based pipeline that has been extensively used in our previous studies. This pipeline involves five steps: (i) in silico digestion of the target genome and indexing of size-selected restriction fragments; (ii) conversion of BAM files into BFA files; (iii) alignment using the bisulfite alignment mode of MAQ (http://maq.sourceforge.net/); (iv) conversion of the MAP files into BAM files; and (v) coordinate mapping of the aligned reads to remove the effects of in silico digestion. This custom pipeline provides high accuracy, and we used it successfully in a number of publications (Carone et al., 2010; Gu et al., 2010; 2011; Harris et al., 2010; Smith et al., 2009; Ock et al., 2010; 2011). However, it is neither user-friendly nor easily portable and currently not publicly available.

The comparison was performed on a total of nine RRBS libraries that were prepared from human embryonic stem cell DNA and sequenced on Illumina GAIIx (Supplementary Table S2). For pairs of biological replicates sequenced with 36bp single-ended reads (rows 1-6 in Supplementary Table S2), we observed equal consistency between replicates and equal bisulfite conversion rates, but slightly higher aligned read numbers and CpG coverage for RRBSMAP than for MAQ. The central processing unit (CPU) time was 5-fold lower and the actual (wall clock) time on a multicore processor node was 30-fold lower for RRBSMAP than for MAQ. The memory consumption of RRBSMAP was 3-fold lower than for MAQ. We also tested RRBSMAP and MAQ on technical replicates sequenced with 76 bp paired-end reads (rows 7-9 in Supplementary Table S2), which is a rarely used configuration for RRBS that we had not previously tested with the MAQ-based pipeline. Compared to 36 bp single-end, RRBSMAP almost doubles the CpG coverage that is achievable with RRBS, while not sacrificing accuracy or performance. In contrast, the MAQ-based pipeline does not work well for 76bp paired-end reads, resulting in substantially lower alignment rates while maintaining high accuracy and consistency between replicates. In summary, RRBSMAP performs at least as well as an extensively validated MAQ-based pipeline on the key performance metrics of bisulfite sequencing, and it significantly outperforms the MAQ-based pipeline in terms of runtime performance and in terms of genomic coverage for 76 bp paired-end reads. The performance of RRBSMAP is sufficient to process thousands of RRBS libraries, a situation that is become increasingly realistic as DNA methylation profiling follows genome-wide association studies and tackles large samples cohorts.

Funding: National Institutes of Health Roadmap Epigenomics (grant U01DA025956 to Y.X. and W.L.); (U01ES017155 to A.M.); Cancer Prevention and Research Institute of Texas (grant RP110471-C3 to Y.X. and W.L.); 973 project 2010CB944900 of China (to W.L.), in part; Pew Charitable Trusts (to A.M.); Feodor Lynen Fellowship (Alexander von Humboldt Foundation) and a Charles A. King Trust Postdoctoral Fellowship (Charles A. King Trust, N.A., Bank of America, Co-Trustee) (to C.B.).

Conflict of Interest: none declared.

REFERENCES

Y. Xi et al.


