BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 16 2012, pages 2180–2181
doi:10.1093/bioinformatics/bts337

Sequence analysis

Advance Access publication June 10, 2012

SAAP-RRBS: streamlined analysis and annotation pipeline for reduced representation bisulfite sequencing

Zhifu Sun[1], Saurabh Baheti[1], Sumit Middha[1], Rahul Kanwar[2], Yuji Zhang[1], Xing Li[1], Andreas S. Beutler[2], Eric Klee[1], Yan W. Asmann[1], E. Aubrey Thompson[3] and Jean-Pierre A. Kocher[1,*]

[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, [2]Department of Medical Oncology, Mayo Clinic College of Medicine, Rochester, MN 55905, USA and [3]Department of Cancer Biology, Mayo Clinic Comprehensive Cancer Center, Jacksonville, FL 32224, USA

Associate Editor: Alex Bateman

ABSTRACT

Summary: Reduced representation bisulfite sequencing (RRBS) is a cost-effective approach for genome-wide methylation pattern profiling. Analyzing RRBS sequencing data is challenging and specialized alignment/mapping programs are needed. Although such programs have been developed, a comprehensive solution that provides researchers with good quality and analyzable data is still lacking. To address this need, we have developed a Streamlined Analysis and Annotation Pipeline for RRBS data (SAAP-RRBS) that integrates read quality assessment/clean-up, alignment, methylation data extraction, annotation, reporting and visualization. This package facilitates a rapid transition from sequencing reads to a fully annotated CpG methylation report to biological interpretation. Availability and implementation: SAAP-RRBS is freely available to non-commercial users at the web site http://ndc.mayo.edu/mayo/research/biostat/stand-alone-packages.cfm.

Contact: baheti.saurabh@mayo.edu or sun.zhifu@mayo.edu

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

Received on February 26, 2012; revised on May 27, 2012; accepted on June 6, 2012

1 INTRODUCTION

DNA methylation is one of the major epigenetic control mechanisms for gene regulation, cellular differentiation, embryogenesis, X chromosome inactivation, genomic imprinting and tumorigenesis (Laird, 2003). Reduced representation bisulfite sequencing (RRBS) has been demonstrated to be a cost-effective approach to analyze DNA methylation pattern in CpG dense regions of the genome (Bock *et al*., 2010; Harris *et al*., 2010; Wang *et al*., 2012). RRBS data analysis is challenging due to the bisulfite-mediated C to T conversion. Several specific RRBS sequence read alignment programs have been developed, which include BSMAP/RRBSMAP (Xi and Li, 2009; Xi *et al*., 2012), BISMARK (Krueger and Andrews, 2011), BSSEEKER (Chen *et al*., 2010) and Illumina's bisulfite sequencing software (http://www.illumina.com/applications/epigenetics.ilmn). Each aligner uses a different strategy for bisulfite-converted read mapping. BISMARK and BSSEEKER convert reference genome (C to T for + and G to A for − strand) and reads (C to T) for alignment. Illumina's pipeline does not align to the whole genome; rather, digitally digested *MspI* fragments are generated as multi-fasta files, which then undergo C/T and G/A conversion. BSMAP/RRBSMAP does not need the reference genome preparation step but everything is handled internally using bitwise masking. In addition to alignment, other steps are equally important to get high-quality and analyzable data for downstream analyses. The *MspI* digestion, bisulfite treatment and subsequent size selection tend to generate a higher portion of low-quality or adapter-contaminated reads; accurate extraction of CpG methylation status from alignment is not implemented in alignment programs or with limited function and more importantly, CpG methylation is highly context specific in the genome and it would be impossible to interpret the data without annotation. Herein, we present a comprehensive workflow that combines all these steps into a single application that can be run on a high-performance computing cluster or on an individual workstation. The workflow accommodates aligned BAM files from different aligners or conduct annotation alone. Moreover, the workflow can be easily extended to analyze whole genome bisulfite sequencing.

2 METHODS

2.1 Architecture and implementation

The SAAP-RRBS consists of four main modules: (1) sequence read assessment and clean-up; (2) alignment to reference genome; (3) methylation status extraction and (4) CpG reporting and annotation (Supplementary Fig. S1). These modules use a suite of public and in-house developed tools using Perl, Python, Java and C.

2.2 Sequence read quality assessment and trimming

SAAP-RRBS first removes low-quality bases and adapter sequences from either end of a read (Martin, 2011). Resulting reads that are shorter than a specified length are discarded to reduce non-unique mapping. The FastQC tool is incorporated to report summarized metrics on the quality of the reads.

2.3 Alignment to reference genome

BSMAP version 2.43 is integrated into the pipeline as the default aligner. For RRBS, -D C-CGG option is turned on for RRBSMAP alignment mode. Note that RRBSMAP was integrated into BSMAP since version 2.0. Early versions of BSMAP are very slow although alignment accuracy is comparable to

*To whom correspondence should be addressed.

other tools (Chatterjee *et al*., 2012; Chen *et al*., 2010; Krueger *et al*., 2012). The introduction of RRBSMAP has significantly increased performance as the program restricts alignment to genomic locations with *MspI* cut sites (CCGG) (Xi *et al*., 2012). We have compared RRBSMAP with three other alignment programs using simulated and real data, and the results are presented in Supplementary Tables S1 and S2 and Figure S2. RRBSMAP shows the highest mapping rate and accuracy, with the similar speed as BSSEEKER. The estimated methylation ratio from RRBSMAP is highly correlated with that from other aligners ($R^2 > 0.99$). In addition, RRBSMAP does not need to prepare reference genome and can be used with both single- and pair-end sequencing data.

2.4 Methylation status extraction

The BAM file is parsed to extract methylation status of C in CpG context. For reads mapped to the forward strand, the total numbers of C and T at the C position of a CpG in the reference are counted and a methylation ratio at that position is calculated by dividing the total number of C by the total number of C plus T (Cs/(Cs + Ts)). For reads mapped to the reverse strand, the numbers of G and A at the G position (complementary C at the reverse strand) of the CpG are counted to obtain the complementary methylation ratio (Gs/(Gs + As)).

2.5 CpG annotation

The annotation module provides each CpG with rich genomic contextual information, including whether it is in exons, within a gene, in a CpG island and distance to transcription start site (TSS). The annotation module also searches the dbSNP database and marks the C position with known single nucleotide polymorphism (SNP) and alternative allele. To speed up the process, annotations for all CpGs in the genome are pre-computed in the initial setup.

2.6 Final report and data visualization

The pipeline generates two main reports for end users: (1) a summary report for all samples in a run, including QC metrics, summary statistics for each sample and links to more detailed reports and (2) a methylation data report with annotations for each sample. This report contains methylation data for CpGs with coverage $\geq 10\times$ and base quality score ≥ 20. Each reported CpG site is dynamically linked to a local instance of the Integrative Genomics Viewer (IGV) for read- and base-level data visualization.

3 RESULTS AND CONCLUSIONS

3.1 Performance and benchmark measurements

SAAP-RRBS is most suitable for a cluster environment where multiple samples can be run in parallel. It takes ~4–6 h to complete the pipeline for a sample (MCF7 breast cancer cell line) with ~50 millions of reads in a Linux platform with CPU at 2.67 GHz, 8 cores and >100 GB RAM. The resulting methylation ratio is highly correlated with the results from BISMARK, BSSEEKER and the Illumina pipeline ($R^2 > 0.99$) as well as from Illumina methylation27k microarray ($R^2 > 0.9$; Supplementary Fig. S2). Some well-known methylation patterns in the cell line are also observed from the data such as promoter hypermethylation of *WT1*, *HOXA5*, *IRX1 and PAX7* (Supplementary Fig. S3).

3.2 Summary report

The summary report contains a link to the FastQC report and summary statistics such as total reads passing filters, mapped reads, mapping rate, bisulfite conversion rate, captured Cs in CpG context, Cs with strand-specific coverage $\geq 10\times$ and base quality ≥ 20 and CpGs that overlap with dbSNP, CpGs within CpG islands and CpGs with genes (Supplementary Fig. S4).

3.3 Annotated report

The final report contains strand-specific CpG methylation and detailed annotations for each C position. The number of methylated Cs, total number of Cs and methylation ratio are reported, along with information of the CpG relative to genomic features such as nearby gene, distance to TSS, overlapping SNP, alternative allele and a CpG island. In the report, users can click a hyperlink for a CpG site to open IGV and view the aligned reads and base level information (Supplementary Fig. S5).

RRBS has become a popular and efficient way to profile genome-wide methylation patterns for novel discoveries in biomedical research. However, the high-throughput and complex data need a fast and convenient bioinformatics tool to process the data into interpretable formats. We have developed a comprehensive pipeline that integrates necessary analyses into a single package. The pipeline is modular, which allows users to use their own alignment tool. Users can also provide pre-computed methylation data with chromosome coordinates and run the workflow in an annotation-only module. The pipeline is highly automated and can be run in a single machine or within a cluster environment where many samples can be processed in parallel for increased performance.

ACKNOWLEDGEMENT

We are very grateful to Dr Yuanxin Xi at Baylor College of Medicine for his assistance to BSMAP/RRBSMAP.

Funding: Center for Individualized Medicine at Mayo Clinic Rochester, MN, and the 26.2 with Donna Foundation.

Conflict of Interest: none declared.

REFERENCES

Bock,C. *et al.* (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.*, **28**, 1106–1114.

Chatterjee,A. *et al.* (2012) Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic acids Res.*

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

Harris,R.A. *et al.* (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.*, **28**, 1097–1105.

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

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

Laird,P.W. (2003) The power and the promise of DNA methylation markers. *Nat. Rev. Cancer*, **3**, 253–266.

Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, **17**, 3.

Wang,L. *et al.* (2012) Systematic assessment of reduced representation bisulfite sequencing to human blood samples: a promising method for large-sample-scale epigenomic studies. *J. Biotechnol.*, **157**, 1–6.

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

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