BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 2 2013, pages 268–270
doi:10.1093/bioinformatics/bts675

Genome analysis

Advance Access publication November 17, 2012

PASS-bis: a bisulfite aligner suitable for whole methylome analysis of Illumina and SOLiD reads

Davide Campagna[1], Andrea Telatin[1], Claudio Forcato[2], Nicola Vitulo[1] and Giorgio Valle[1,2,*]

[1]CRIBI Biotechnology Centre and [2]Department of Biology, Università di Padova, 35131 Padova, Italy

Associate Editor: Martin Bishop

ABSTRACT

Summary: The sequencing of bisulfite-treated DNA (Bi-Seq) is becoming a gold standard for methylation studies. The mapping of Bi-Seq reads is complex and requires special alignment algorithms. This problem is particularly relevant for SOLiD color space, where the bisulfite conversion C/T changes two adjacent colors into 16 possible combinations. Here, we present an algorithm that efficiently aligns Bi-Seq reads obtained either from SOLiD or Illumina. An accompanying methylation-caller program creates a genomic view of methylated and unmethylated Cs on both DNA strands.

Availability and implementation: The algorithm has been implemented as an option of the program PASS, freely available at http://pass.cribi.unipd.it.

Contact: pass@cribi.unipd.it

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

Received on August 24, 2012; revised on November 12, 2012; accepted on November 13, 2012

1 INTRODUCTION

Bisulfite sequencing (Bi-Seq) is an established method to provide information on methylation status of DNA (Grunau *et al.*, 2001). Unmethylated Cs are deaminated by bisulfite treatment, and after PCR, they turn into Ts, whereas methylated Cs remain unchanged. The Bi-Seq reads are mapped onto the reference genome by taking into account that Cs may have become Ts. The resulting T/C discrepancies are then used to infer the methylation status. Although the general rationale of Bi-Seq is easily understood, the analysis of the data is difficult. The two DNA strands change asymmetrically, as the Cs occur in different positions. As a result, each strand produces a different reference genome for Bi-Seq mapping, with the consequent increase of complexity and a possible loss of mapping specificity. A further specificity loss may be due to the diminished linguistic complexity, as we essentially switch from a four-base to the three-base alphabet (AGT).

Some early methods for Bi-Seq mapping required long execution times, being based on probability vectors (Cokus *et al.*, 2008) or on iterations of all possible C/T conversions (Xi and Li, 2009). More recently, some efficient algorithms have been developed such as BS Seeker (Chen *et al.*, 2010) and Bismark (Krueger and Andrews, 2011). These latter algorithms are based on the full C/T conversion on both plus and minus strands of the reference genome as well as a full C/T conversion of the reads, so that the mapping is performed using only the three letters AGT. A post-mapping procedure will then re-consider the original four-base sequences and compare the occurrences of Cs in the genome and in the reads.

The AGT conversion approach is effective, but it cannot be directly applied to SOLiD reads because they are encoded in color space, where each color corresponds to four possible words of two bases. As a consequence, the available software to align Bi-Seq SOLiD reads is still slow and inefficient. In this respect some of the best programs are SOCS-B (Ondov *et al.*, 2010), based on an iterative algorithm, and B-SOLANA (Kreck *et al.*, 2012)

The difficulty of mapping Bi-Seq SOLiD reads is unfortunate, as color space has some interesting potential owing to the double encoding of each base and to the fact that after bisulfite treatment, the full set of four colors would still be present.

Here we introduce a new implementation of PASS (Campagna *et al.*, 2009) that allows the mapping of Bi-Seq reads on an AGT simplified genome, managing both base space as well as native color space. The program is able to perform alignments with gaps, a feature that is not available in most programs for bisulfite mapping. Furthermore, when possible, it discriminates the genomic strand from which the read originated.

2 ALGORITHM AND IMPLEMENTATION

The PASS algorithm (Campagna *et al.*, 2009) is based on seed-word search and extension. Seed words are typically 12–14 bases long with gapped patterns. To speed up the extension process, PASS uses pre-computed scoring tables (PST) containing the results of the alignment of all the possible words of length w (typically $w = 8$) aligned against each other. Therefore, the PASS process works at three levels: (i) the general level of PSTs that needs to be calculated only once; (ii) the genomic level, to create the hash table for seed words; and (iii) the mapping of the reads.

We were able to implement the bisulfite mapping algorithm as an option of the main PASS program, thus taking full advantage of the ample set of alignment parameters available in PASS. This implementation required actions at each of the three aforementioned levels. Furthermore, two different pipelines had to be established for base- and color space. Nevertheless, the execution of the program is relatively simple and fast, requiring only the specification of a few parameters. However, more optional parameters may be used to define color space, gaps and other

*To whom correspondence should be addressed.

alignment arguments. A schematic diagram of the alignment process is shown in Figure 1, both for base space and color space.

3 STRATEGY FOR COLOR SPACE MAPPING

SOLiD reads are much better aligned in their native color space than in base space. In a normal SOLiD mapping procedure, PASS translates the genomic sequence into color space, thus making possible the mapping of the reads in color space. For Bi-Seq, PASS first performs a C/T substitution of the reference genome, and then it converts the AGT genome to color space (see Figure 1).

Color space makes C/T conversion difficult, requiring two steps: first from color space to base space, and then after C/T conversion, back from base space to color space. The full C/T conversion is required for mapping purposes, whereas for the detection of methylation, the original Cs or Ts are considered. The main risk of this procedure is that an error in the color space sequence would put out of frame the conversion into base space, with the consequent production of wrong seed words. However, in practical terms, this problem does not have a relevant impact, as most reads are likely to have a good-quality stretch to build at least one correct seed word. The Exact Call Chemistry (Massingham and Goldman, 2012) that has recently become available may further reduce the importance of this problem. However, to maximize the performance of the program in color space, PASS implements an additional mapping strategy that takes the unmapped reads obtained with the aforementioned procedure and automatically runs them through a more conservative mapping procedure based on a combinatorial assortment of genomic C/T conversions. This second strategy is useful only when the quality of the reads is not optimal; therefore, it has been implemented as an option to avoid unnecessary waste of time. It must be stressed that after the seeding step, all the remaining alignment procedures, including those for filtering with pre-computed score tables, occur in color space, taking full advantage of the rules of color compatibility that apply for single base substitutions.

Table 1 compares the performance of PASS-bis with two other programs designed for color space mapping of Bi-Seq reads. The analysis was also extended to base space reads (BS), such as those produced by Illumina. It can be seen that, despite bisulfite treatment, PASS aligns color space reads practically as well as base space reads, thus allowing an efficient Bi-Seq analysis of SOLiD data.

The execution time of PASS-Bis depends on the parameters; it is faster than SOCS and generally slightly slower than B-SOLANA. A detailed analysis of execution times and memory requirements is given in the User Manual available with the PASS Package.

The output of PASS-bis is a SAM file that besides the standard mapping information includes an additional field indicating on which of the four possible bisulfite-modified strands the alignment occurred. An accompanying methylation caller is supplied with the PASS package. The methylation caller reads the SAM file and infers the methylation status of each C from all the reads covering that position. The program also discriminates the C/T instances due to bisulfite from those due to SNPs. Moreover, it identifies reads mapping with strand ambiguity

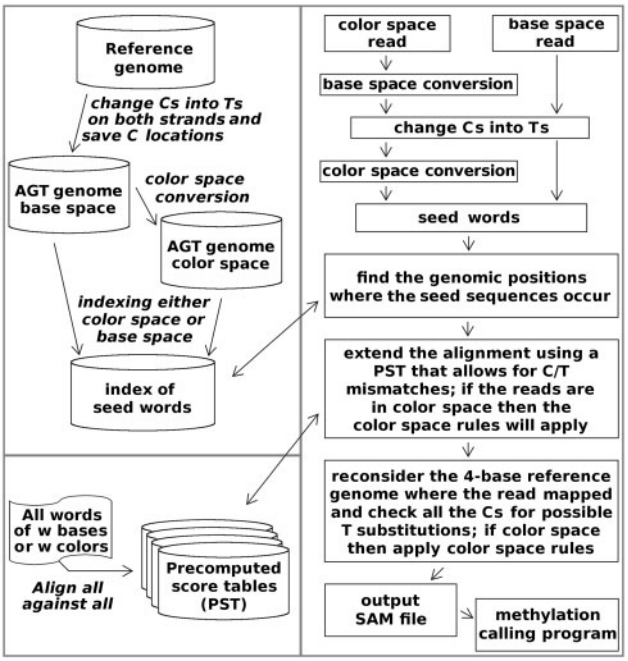

Fig. 1. Workflow of the alignment procedures. Top-left: procedure to process the reference genome. Bottom-left: universal precomputed score tables. Right: steps for the alignment of a bisulfite read

Table 1. Comparison of bisulfite mappers working in color space

Program	Me%	UC	UW	MM	TotM
B-SOLANA	0	407 157	13 635	7573	428 365
B-SOLANA	50	85 357	4002	1478	90 837
B-SOLANA	100	27 375	1646	437	29 458
SOCS	0	419 798	22 307	20 088	462 193
SOCS	50	417 717	17 147	18 289	453 153
SOCS	100	207 917	220 528	22 481	450 926
PASS(CS)	0	837 498	11 569	139 124	988 191
PASS(CS)	50	853 538	9883	122 867	986 288
PASS(CS)	100	833 285	42 065	120 194	995 544
PASS(BS)	0	862 353	1517	134 168	998 038
PASS(BS)	50	875 687	1588	120 849	998 124
PASS(BS)	100	869 016	6336	122 924	998 276

B-SOLANA 1.0, SOCS 2.2 and PASS 1.7 were tested on a simulated test set of 1 million reads with different levels of cytosine methylation (Me%). Sequencing errors and mutations were inserted in the reads, respectively, at 2% and 0.1% of the positions. All the reads were in color space, with the exception of PASS (BS), which were in base space. UC, unique correct, reads that map correctly at a unique position; UW, unique wrong position; MM, reads with multiple map; TotM, total reads mapped.

such as those occurring when a read originates from a DNA strand without Cs or from a strand fully methylated. Some examples and statistics of the methylation caller are given in the Supplementary Material.

Funding: Research supported by the Italian Epigenomics Flagship Project EPIGEN (MIUR/CNR) and by Fondazione Cariparo, Chromus Project.

Conflict of Interest: none declared.

REFERENCES

Campagna,D. *et al.* (2009) PASS: a program to align short sequences. *Bioinformatics*, **25**, 967–968.

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

Cokus,S.J. *et al.* (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, **452**, 215–219.

Grunau,C. *et al.* (2001) Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acid Res.*, **29**, e65.

Kreck,B. *et al.* (2012) B-SOLANA: an approach for the analysis of two-base encoding bisulfite sequencing data. *Bioinformatics*, **28**, 428–429.

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

Massingham,T. and Goldman,N. (2012) Error-correcting properties of the SOLiD Exact Call Chemistry. *BMC Bioinformatics*, **13**, 145.

Ondov,B.D. *et al.* (2010) An alignment algorithm for bisulfite sequencing using the Applied Biosystems SOLiD System. *Bioinformatics*, **26**, 1901–1902.

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