BS Seeker: precise mapping for bisulfite sequencing

Pao-Yang Chen, Shawn J. Cokus and Matteo Pellegrini*

Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles.

Abstract

Background

Bisulfite sequencing using next generation sequencers yields genome-wide measurements of DNA methylation at single nucleotide resolution. Traditional aligners are not designed for mapping bisulfite-treated reads, where the unmethylated Cs are converted to Ts. We have developed BS Seeker, an approach that converts the genome to a three-letter alphabet and uses Bowtie to align bisulfite-treated reads to a reference genome. It uses sequence tags to reduce mapping ambiguity. Post-processing of the alignments removes non-unique and low-quality mappings.

Results

We tested our aligner on synthetic data, a bisulfite-converted *Arabidopsis* library, and human libraries generated from two different experimental protocols. We evaluated the performance of our approach and compared it to other bisulfite aligners. The results demonstrate that among the aligners tested, BS Seeker is more versatile. It offers comparable performance to RMAP when tags are not available, and improves mapping using tags. In the case of mapping human genome, BS Seeker is far more efficient than RMAP.

*To whom correspondence should be addressed.

Conclusions

BS Seeker provides fast and accurate mapping of bisulfite-connverted reads. It can work with BS reads from the two major experimental protocols, and is able to map large genomes such as human. The Python program is freely available at http://paoyang.bol.ucla.edu/Site/BS_Seeker.html .[1]

Background

Epigenetic regulation, such as cytosine (C) DNA methylation, is important in gene regulation and transposon silencing. The gold standard technique for studying DNA methylation is genomic bisulfite sequencing [1]. Sodium bisulfite converts unmethylated Cs to uracils, but 5-methylcytosines remain unconverted. Hence, after PCR amplification, unmethylated Cs are converted to thymines (T) while methylated Cs are unchanged. Recently, [2] developed a protocol, BS-Seq, which couples bisulfite sequencing with next generation sequencing and completed a first single nucleotide resolution map of methylation in *Arabidopis*. While BS-Seq opens up new avenues for genome-wide measurements of DNA methylation [3-6], aligning millions of bisulfite-treated short reads (BS reads) onto the reference genome remains a challenge. Mapping bisulfite-converted reads leads to ambiguity, since Ts in the read can map to both genomic Cs or Ts. Traditional aligners such as BLAT [7], SOAP [8], and Bowtie [9] do not explicitly enable bisulfite mapping.

Currently, there are only a few aligners explicitly designed for mapping BS reads. CokusAlignment [2] treats each cycle in a read as probabilities of A, C, T, G and uses a suffix tree searching algorithm. However, to date only the *Arabidopsis* suffix tree has been published. Other newly-developed bisulfite mapping software includes BSMAP [10] and RMAP [11], which, unlike CokusAlignment, model reads as discrete base calls instead of probability vectors. BSMAP enumerates all possible combinations of C/T conversion in the BS read to find the uniquely mapping position with the least mismatches on the refer-

[1] This is a temporary URL before BS Seeker is officially released.

ence genome. It is reported to have a similar sensitivity as CokusAlignment and outperformed the methods described in [3] and [4]. The bisulfite mapping in RMAP uses Wildcard matching for mapping Ts.

Two library protocols have been developed for constructing bisulfite converted libraries (see Figure 1). Cokus *et al*'s protocol [2] uses two amplification steps: the first amplification generates both forward and reverse bisulfite-converted sequences ligated with DNA adapters of DpnI restriction sites. These sequences are then digested by DpnI restriction enzymes that result in the 5-bp sequence tags on the bisulfite-converted sequences. There are two patterns of tags based on the forward (+FW) and reverse (-FW) directions of the bisulfite-converted sequences. After the ligation with standard Solexa adapters and the second amplification step, four types of bisulfite-converted reads are generated. They are forward and reverse reads from Watson (+FW, +RC) and from Crick stands (-FW, -RC), respectively (see Fig. 2A). These tags are essential to reduce the ambiguity of certain classes of reads. Unlike BSMAP or RMAP, BS Seeker is able to explicitly use the tags to improve mapping. The second experimental protocol (i.e. Lister et al) generates bisulfite libraries using premethylated adapters, and in this case no tags are present and all reads are +FW or -FW.

BS Seeker uses Bowtie for mapping BS reads generated from either experimental protocol. It maps C/T converted FW reads to the C/T converted reference strands, and G/A converted RC reads to G/A converted, reverse complements of the strands. Post-processing removes low-quality mappings based on the number of mismatches. Although the similar mapping strategy of treating all Cs as Ts (Gs as As) has been used in [3, 6, 12], their implementations are not available, and the current aligners are still to be improved in terms of the efficiency and accuracy.

For the evaluation of our aligner, we use synthetic BS reads to assess the accuracy of mapping. We further examine the mapping results by comparing the methylation statistics in the synthetic reads and in

the mapped sequences. We then use BS Seeker to align an *Arabidopsis* library and human libraries from the two experimental protocols, which provides us with an evaluation using real data.

Implementation

Depending on the bisulfite experimental protocol that was used to generate the library [2, 4], BS reads may be observed in either of four or two forms. Cokus et al's protocol generates a forward read (+FW) from the Watson strand, the reverse complement (+RC) of +FW, a forward read (–FW) from the Crick strand, and the reverse complement (–RC) of –FW, see Fig. 2A. Lister et al's protocol generates only +FW and –FW reads. We first describe how BS Seeker handles data generated from Cokus et al's protocol. It first converts all Cs to Ts on FW reads and both strands of the reference genome, so that the subsequent mapping is performed using only 3 letters, A, T, G. Similarly, G/A conversion is performed on RC reads and both strands of the reverse complement of the reference genome. Then it uses Bowtie to map the C/T converted FW reads to the C/T converted Watson and Crick strands, and the G/A converted RC reads to the two G/A converted reverse complements of the Watson and Crick strands. Reads that do not have a tag are treated as if they could be both FW and RC reads. During each of the four runs of Bowtie, the mapped positions for each read are recorded. After all the runs of Bowtie are complete, only unique alignments are retained. Here, we define unique alignments as those that have no other hits with the same or fewer mismatches in the 3-letter alignment (between the converted read and the converted genomic sequence). Finally, we calculate the number of mismatches, except that the read Ts aligned to genomic Cs, between the original BS reads and the unconverted genome (see Fig 2B). The low-quality alignments with the number of mismatches larger than the user-defined value are discarded. Aligning

reads generated from Lister et al's protocol is simpler, since there are no RC reads, and consequently Bowtie is only run twice.

Results and Discussion

We tested BS Seeker against the other two publicly released bisulfite aligners, BSMAP and RAMP, by mapping synthetic reads. The sensitivity and the specificity of the aligner's output are assessed by calculating the percentage of reads it mapped uniquely and their accuracy, which is the ratio of the number of correctly mapped reads over the total uniquely mapped reads. We also calculate the inferred average methylation rate in order to discern possible mapping bias from the aligners. We then show our mapability on a lane of experimental data from *Arabidopsis*. Finally, we mapped the BS reads of human libraries generated from the two experimental protocols, in order to demonstrate our mapability on different protocols. We further compare the two protocols by calculating their error rates per cycle. We also compare to RMAP on the efficiency of mapping human genome.

Mapping synthetic reads

We simulated 36-mer BS reads from human chromosome 21. One million contained no base calling errors and another one million had base calling errors that follow the error distribution from [13], see Supplementary Information for details. These were mapped to chromosome 21 using all three aligners. The simulated data were generated using both protocols. As shown in Table 1 and Supplementary Table S1, we found that in all cases BSMAP was significantly slower than BS Seeker and RMAP. When the simulated reads are based on Lister et al's protocol with no tags, RMAP has a slight speed advantage over BS Seeker, but the two methods are otherwise quite similar. However, when the Cokus et al library protocol is used, BS Seeker had a higher accuracy and shorter run time than RMAP. Furthermore, the methyla-

tion rate inferred from the mapped reads using BS Seeker is very close to that of the initial synthetic data, indicating that the alignments are less biased than those of RMAP. These results suggest that BS Seeker is able to work with data from both protocols. By explicitly using tag information, it further leads to superior mapping results when data is generated by the Cokus et al. protocol.

Mapping BS reads from *Arabidopsis and Human*

We mapped 2,946,339 BS reads of a single lane of an *Arabidopsis* library from Cokus, et al., (2008), which is the same library tested in Xi and Li, (2009). BS Seeker is able to uniquely map 56.3% of the reads, which compares favorably to the coverage reported in Xi et al.. The methylation rates inferred from the mapped reads for CG, CHG and CHH (H stands for A or C or T) are 25.5%, 8%, and 2.2%, which are comparable to the published results [2].

Comparison of libraries from Cokus *et al*'s and Lister *et al*'s protocols

In order to compare the two library protocols, BS Seeker is used to map BS reads from human libraries generated from the two protocols [6]. The mapabilities of reads from Cokus et al's protocol and from Lister et al's protocol are very close (38.3% and 38.6% respectively). The CG methylation rate we obtained from the mapping of the reads from Lister et al's protocol is 82.3%, which coincides with the published result of 82.7% [6]. We further examine the reads quality by calculating the percentage of an A or G on bisulfite-converted read being aligned to a genomic C. The higher the percentage suggests the more errors due to the library. As shown in Figure 3, the reads from Lister et al's protocol show a higher error rate on all cycles against those from Cokus et al's protocol. Although the Cokus et al's protocol is more complicated in the preparation as well as the mapping procedure, from our comparison it produces are reads more accurately than those from Lister et al's library protocol.

We also used RMAP to align the same human libraries. The running time for RMAP to map one lane of reads is between 11.9-13.7 hours, while for BS Seeker it is between 20-50 minutes. Both aligners showed close mapability, see Supplementary Information. The significant advantage on the mapping efficiency suggests that BS Seeker is the only bisulfite aligner feasible for mapping large genomes.

Availability and requirements

- **Project name:** BS Seeker
- **Project home page:** http://paoyang.bol.ucla.edu/Site/BS_Seeker.html
- **Operating system(s):** Linux, Mac OS
- **Programming language:** Python
- **Other requirements:** Python 2.5.2 or higher, Bowtie 0.10.0 or higher
- **License:** Free for all users
- **Any restrictions to use by non-academics:** None

Authors' contributions

PC wrote BS Seeker and drafted the manuscript. SC and MP participated in the design and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Krishna Chodavarapu for testing the program. PC is supported by Eli & Edythe Broad Center of Regenerative Medicine & Stem Cell Research at UCLA.

References

1. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL: **A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands**. *Proc Natl Acad Sci U S A* 1992, **89**(5):1827-1831.
2. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE: **Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning**. *Nature* 2008, **452**(7184):215-219.
3. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB *et al*: **Genome-scale DNA methylation maps of pluripotent and differentiated cells**. *Nature* 2008, **454**(7205):766-770.
4. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: **Highly integrated single-base resolution maps of the epigenome in Arabidopsis**. *Cell* 2008, **133**(3):523-536.
5. Smith ZD, Gu H, Bock C, Gnirke A, Meissner A: **High-throughput bisulfite sequencing in mammalian genomes**. *Methods* 2009, **48**(3):226-232.
6. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM *et al*: **Human DNA methylomes at base resolution show widespread epigenomic differences**. *Nature* 2009, **462**(7271):315-322.
7. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656-664.
8. Li R, Li Y, Kristiansen K, Wang J: **SOAP: short oligonucleotide alignment program**. *Bioinformatics* 2008, **24**(5):713-714.
9. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3):R25.
10. Xi Y, Li W: **BSMAP: whole genome bisulfite sequence MAPping program**. *BMC Bioinformatics* 2009, **10**:232.
11. Smith AD, Chung W, Hodges E, Kendall J, Hannon G, Hicks J, Xuan Z, Zhang MQ: **Updates to the RMAP short-read mapping software**. *Bioinformatics* 2009.
12. Hodges E, Smith AD, Kendall J, Xuan Z, Ravi K, Rooks M, Zhang MQ, Ye K, Bhattacharjee A, Brizuela L *et al*: **High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing**. *Genome Res* 2009, **19**(9):1593-1605.
13. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing**. *Nucleic Acids Res* 2008, **36**(16):e105.

Figures

Fig. 1. The two library protocols generating bisulfite-converted reads.

Cokus et al's experimental protocol uses two amplification steps for generating bisulfite-converted sequences and for high throughput sequencing. The bisulfite-converted reads are preceded by one of two tags in the first 5 nucleotides of reads. Lister et al's protocol generates bisulfite libraries using premethylated adapters, and in this case no tags are present.

Fig. 2. Schematic diagrams of the 4 forms of BS reads, mapping and post processing.

2**A**. BS reads may be in one of the 4 forms: +FW, +RC, –FW, –RC. 2**B**. Bowtie aligns C/T converted reads to the C/T converted strands. During the post processing, the number of mismatches is counted

except those between read Ts and genomic Cs. Low-quality mappings with many mismatches are removed.

Fig. 3. The fractions of error matching between reads and genomic sequences.

For each C on the genomic sequences, the relative frequency of the C being mapped to a read A or G is calculated according to its cycle number on the sequence. The test library from Cokus et al's protocol has reads length of 47 base pairs, and that from Lister et al's protocol has reads length of 87 base pairs.

Tables

Table 1. Mapping 1M synthetic human chr. 21 reads onto human chr. 21

Aligner	Experimental protocol	Uniquely mapped reads [a] (%)	Accuracy (%)	Methylation rates[b] (CG/CHG/CHH) (%)			CPU time
No base calling error							
BS Seeker	Lister et al	91.7	100	72.0	0	0	209s [c]
BSMAP	Lister et al	92.1	100	72.3	0	0	15h43m20s
RMAP	Lister et al	91.7	100	72.0	0	0	185s
BS Seeker	Cokus et al	89.6	100	72.0	0	0	263s [c]
BSMAP	Cokus et al	89.8	99.6	72.4	0	0	15h46m40s
RMAP	Cokus et al	80.2	99.0	71.3	0.02	0.1	400s
Simulated base calling errors							
BS Seeker	Lister et al	91.0	99.54	71.5	0.4	0.4	217s
BSMAP	Lister et al	91.1	99.57	72.1	0.4	0.4	15h19m51s
RMAP	Lister et al	91.0	99.52	71.6	0.4	0.4	188s

[a] Up to 2 mismatches are allowed. [b] The simulated methylation rates are set to be 72%, 0%, and 0% for CG, CHG, and CHH. [c] Preprocessing reference genome needs 2-5 cpu minutes for the first run.

Additional Files

Supplementary Information (PDF)

Supplementary materials to the BS Seeker project.

BS Seeker (compressed file; TGZ)

BS Seeker source code: Compressed file containing the source code for BS Seeker.