BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 24 2013, pages 3227–3229
doi:10.1093/bioinformatics/btt548

Sequence analysis

Advance Access publication September 23, 2013

BSeQC: quality control of bisulfite sequencing experiments

Xueqiu Lin[1], Deqiang Sun[2,3], Benjamin Rodriguez[2,3], Qian Zhao[1], Hanfei Sun[1], Yong Zhang[1] and Wei Li[1,2,3,*]

[1]Department of Bioinformatics, School of Life sciences and Technology, Tongji University, Shanghai 20092, China, [2]Dan L. Duncan Cancer Center and [3]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX 77030, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Bisulfite sequencing (BS-seq) has emerged as the gold standard to study genome-wide DNA methylation at single-nucleotide resolution. Quality control (QC) is a critical step in the analysis pipeline to ensure that BS-seq data are of high quality and suitable for subsequent analysis. Although several QC tools are available for next-generation sequencing data, most of them were not designed to handle QC issues specific to BS-seq protocols. Therefore, there is a strong need for a dedicated QC tool to evaluate and remove potential technical biases in BS-seq experiments.

Results: We developed a package named BSeQC to comprehensively evaluate the quality of BS-seq experiments and automatically trim nucleotides with potential technical biases that may result in inaccurate methylation estimation. BSeQC takes standard SAM/BAM files as input and generates bias-free SAM/BAM files for downstream analysis. Evaluation based on real BS-seq data indicates that the use of the bias-free SAM/BAM file substantially improves the quantification of methylation level.

Availability and implementation: BSeQC is freely available at: http://code.google.com/p/bseqc/.

Contact: wl1@bcm.edu

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

Received on April 23, 2013; revised on September 2, 2013; accepted on September 18, 2013

1 INTRODUCTION

DNA methylation, an epigenetic modification affecting the organization and function of the genome, plays an important role in several key biological processes, including transcriptional regulation, X-chromosome inactivation, genomic imprinting, silencing of repetitive elements and chromatin structure (Robertson, 2005). Bisulfite conversion followed by next-generation sequencing (BS-seq) has emerged as a powerful technique for detecting genome-wide DNA methylation at single-nucleotide resolution (Laird, 2010). Ideally, BS-seq experiment should be able to directly and exactly quantify the methylation level of every cytosine in the genome. However, current BS-seq protocols possess several intrinsic technical biases, which may impact methylation level estimation. These biases include overhang end-repair, 5′ bisulfite conversion failure, sequencing into the adaptor and 3′ low sequencing quality (Bock, 2012; Liu *et al.*, 2012). Although the last two biases are also present in other next-generation sequencing data, the first two biases, i.e. end-repair and 5′ bisulfite conversion failure, are highly specific to BS-seq protocols. BS-seq reads with the first two biases can be perfectly mapped to the reference genome; however, they will result in biased methylation estimation. First, after sonication, the overhangs of DNA fragments are normally end-repaired with unmethylated cytosines to restore the double stranded DNA (Illumina WGBS for Methylation Analysis Guide). This may introduce artificially low methylation rates at both ends of the DNA fragments. Second, bisulfite conversion failure is known to be enriched at the 5′ end of reads, most likely caused by the re-annealing of sequences adjacent to the methylated adapters in bisulfite conversion (Berman *et al.*, 2012). This may bring in artificially high methylation rates at the 5′ end of reads. Most quality control (QC) tools, such as htSeqTools (Planet *et al.*, 2012) and FastQC (Andrews, 2010), only focus on the general sequence quality and adaptor trimming, or other next-generation sequencing applications such as RNA-seq (Wang *et al.*, 2012). To the best of our knowledge, BSmooth and Bis-SNP are the only two tools that support QC specific to BS-seq (Hansen *et al.*, 2012; Liu *et al.*, 2012). However, their QC functions are highly limited. The QC model in BSmooth was designed to work with its own alignment and analysis pipeline, and thus cannot take standard BAM/SAM files as input. Furthermore, BSmooth can detect but not correct the biases. Bis-SNP only supports the correction of 5′ bisulfite conversion failure. Because BS-seq experiments are widely used and the resulting data will continue to grow exponentially in the near future, there is a strong need for a dedicated QC tool to assess and improve the quality of BS-seq experiments. BSeQC, the tool we developed, is focused on the technical biases specific to BS-seq experiments. BSeQC can automatically evaluate and remove the biases based on a user defined statistical cutoff. Evaluation based on real BS-seq data indicates that, after bias removal, the quantification of methylation level is significantly improved.

2 METHODS

BSeQC uses M-bias plot, i.e. average DNA methylation level for each read position, to assess and visualize bisulfite-specific biases (Hansen *et al.*, 2011). For each CpG in uniquely aligned read, BSeQC records its relative position in the read and methylation state (methylated: C or unmethylated: T). For a given input SAM/BAM file, BSeQC piles up all the records and plots the mean methylation level at each read position.

*To whom correspondence should be addressed.

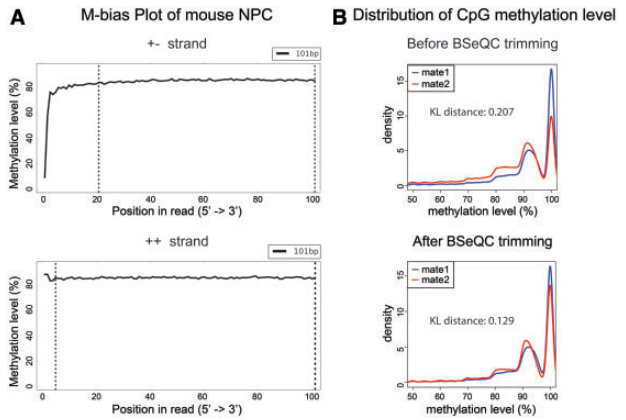

Fig. 1. BSeQC analyses of paired-end mNPC data. (**A**) M-bias plots for two complementary strands of chr1. Horizontal line: M-bias plot; dash vertical line: trimming positions. (**B**) Distributions of methylation levels. Blue line: mate1 (++ and −+ strands); red line: mate2 (+− and −− strands)

As the methylation levels are expected to be independent of the read positions, the M-bias plot ought to exhibit a horizontal line. If there are biases, the M-bias plot will show position-specific deviation from the horizontal line, which are often observed at both ends of reads. Because different DNA strands and read lengths can have distinct biases, BSeQC generates separate M-bias plots for different strand and read length configurations. BSeQC also produces non-CpG cytosine M-bias plots to detect bisulfite conversion failure because non-CpG cytosines should be almost completely converted.

To remove biases detected in the M-bias plot, BSeQC uses a statistical cutoff to make trimming decision automatically. Given the observation that read center positions (30–70% of read length) usually have high quality, whereas both read ends are susceptible to technical biases, the average methylation levels in read center positions can be used to fit a normal NULL distribution $\sim N(\mu, \sigma^2)$. The methylation level of each read end position is then evaluated by a *P*-value, indicating the statistical significance of the deviation from the NULL. The positions of all the reads will be trimmed if they exhibit significant biases, i.e. $P \leq 0.01$. BSeQC automatically uses the most stringent trimming decision made by either CpG or non-CpG cytosines M-bias plots. After bias removal, BSeQC generates a corresponding bias-free SAM/BAM files for the downstream analysis.

3 RESULTS

We demonstrated the capabilities of BSeQC using two publicly available datasets: (i) paired-end WGBS from mouse ES-derived neuronal progenitor cells (mNPC) (Stadler *et al.*, 2011), and (ii) single-end WGBS from H1 human ES cells (Lister *et al.*, 2009). After reads mapping using BSMAP (Xi and Li, 2009), we observed the end-repair bias in the 5′ of '±' strand (Supplementary Table S1) and the 5′ bisulfite conversion failure in the '++' strand in the M-bias plots of paired-end mNPC data (Fig. 1A). This observation indicates that different DNA strands may exhibit different technical biases in BS-seq experiments (Supplementary Note S1). As expected, the 5′ bisulfite conversion failure is more obvious in non-CpG cytosine M-bias plots (Supplementary Fig. S1). In the single-end H1 data, we used BSMAP to remove adaptor and low-quality nucleotides during mapping. However, there are still biases in the 3′ of '++' strands of both replicates, probably indicating residual adapters and low-quality nucleotides that cannot be fully removed by conventional QC approaches (Supplementary Fig. S2A). In addition, we observed different biases in different read length in H1 data (Supplementary Fig. S3).

Because there is no ground truth for methylation estimation, we decided to use the concordance between two replicates to evaluate the performance of BSeQC bias trimming. As expected, the agreement of methylation level distributions between two read mates derived from the same paired-end mNPC data is significantly increased after BSeQC trimming, especially in high methylation levels (Fig. 1B). The Kullback–Leibler distance, a measurement of difference between two probability distributions (Kullback and Leibler, 1951), decreases from 0.207 to 0.129 after BSeQC trimming. Similar improvement can also be observed in single-end H1 data (Supplementary Fig. S2B), RRBS data (Supplementary Fig. S4) and mouse NPC data (Supplementary Fig. S5). Together, we conclude that the usage of BSeQC can significantly improve the methylation estimation.

4 CONCLUSION

BSeQC serves as a critical step between reads mapping and methylation estimation. BSeQC comprehensively evaluates the technical biases related to BS-seq protocols and automatically makes trimming decision using a statistical cutoff. This is a vast improvement to arbitrarily read trimming (e.g. the first three and last three nucleotides). In addition to M-bias plot guided bias trimming, BSeQC can also automatically remove clonal reads from over-amplification and avoid double counting of overlapped segments in paired-end reads (Supplementary Fig. S6). With its standard SAM/BAM input/output interface, BSeQC can be easily incorporated into any existing BS-seq data analysis pipeline. In summary, we believe BSeQC will greatly facilitate the analysis and understanding of DNA methylation bisulfite sequencing data.

Funding: 973 Program of China [2010CB944900], CPRIT RP110471-C3 and NIH [R01HG007538].

Conflict of Interest: none declared.

REFERENCES

Andrews,S. (2010) FastQC. A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (8 October 2013, date last accessed).

Berman,B.P. *et al.* (2012) Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.*, **44**, 40–46.

Bock,C. (2012) Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.*, **13**, 705–719.

Hansen,K.D. *et al.* (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.*, **13**, R83.

Hansen,K.D. *et al.* (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.*, **43**, 768–775.

Kullback,S. and Leibler,R.A. (1951) On information and sufficiency, information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86.

Laird,P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.

Lister,R. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

Liu,Y. *et al*. (2012) Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol*., **13**, R61.

Planet,E. *et al*. (2012) htSeqTools: high-throughput sequencing quality control, processing and visualization in R. *Bioinformatics*, **28**, 589–590.

Robertson,K.D. (2005) DNA methylation and human disease. *Nat Rev. Genet*, **6**, 597–610.

Stadler,M.B. *et al*. (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, **480**, 490–495.

Wang,L. *et al*. (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, **28**, 2184–2185.

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