Journal of Biotechnology 157 (2012) 1–6

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Systematic assessment of reduced representation bisulfite sequencing to human blood samples: A promising method for large-sample-scale epigenomic studies

Wang Li[a,1], Sun Jihua[a,1], Wu Honglong[b,a,1], Liu Siyang[c,a], Wang Junwen[a], Wu Boxin[a], Huang Shujia[d,a], Li Ning[a], Wang Jun[a], Zhang Xiuqing[a,*]

[a] BGI-Shenzhen, Shenzhen, Guangdong, China
[b] Genome Research Institute, Shenzhen University Medical School, Shenzhen, China
[c] School of Life Science, Sun-Yat-Sen University, Guangzhou, Guangdong, China
[d] School of Bioscience and Biotechnology, South China University of Technology, Guangzhou, Guangdong, China

ARTICLE INFO

Article history:
Received 7 March 2011
Received in revised form 23 June 2011
Accepted 29 June 2011
Available online 6 July 2011

Keywords:
Reduced representation bisulfite sequencing
Human DNA methylation
Epigenomics

ABSTRACT

Complementary to the time- and cost-intensive direct bisulfite sequencing, we applied reduced representation bisulfite sequencing (RRBS) to the human peripheral blood mononuclear cells (PBMC) from YH, the Asian individual whose genome and epigenome has been deciphered in the YH project and systematically assessed the genomic coverage, coverage depth and reproducibility of this technology as well as the concordance of DNA methylation levels measured by RRBS and direct bisulfite sequencing for the detected CpG sites. Our result suggests that RRBS can cover more than half of CpG islands and promoter regions with a good coverage depth and the proportion of the CpG sites covered by the biological replicates reaches 80–90%, indicating good reproducibility. Given a smaller data quantity, RRBS enjoys much better coverage depth than direct bisulfite sequencing and the concordance of DNA methylation levels between the two methods is high. It can be concluded that RRBS is a time and cost-effective sequencing method for unbiased DNA methylation profiling of CpG islands and promoter regions in a genome-wide scale and it is the method of choice to assay certain genomic regions for multiple samples in a rapid way.

1. Introduction

By mapping the DNA methylome, researchers can look deeply into the key epigenetic mechanism for controlling gene expression. DNA methylation determines the fate of developing cells (Lister et al., 2009; Meissner et al., 2008; Straussman et al., 2009), and consequently, aberrant methylation patterns can lead to pathological states such as cancer (Esteller, 2007). DNA methylation profiling provides valuable information on a wide variety of research fronts, including differentiation of stem cells (Lister et al., 2009; Meissner et al., 2008; Straussman et al., 2009), the mechanism of environmental influences on the development of complex disease (Lambert et al., 2011), and cellular aging (Fraga et al., 2005). Accurate DNA methylation profiling can be a key step in identifying biomarkers for therapeutic interventions (Bock, 2009).

Nowadays, analysis of DNA methylation patterns relies increasingly on sequencing-based profiling methods because it promises a sensitive and high-resolution analysis of DNA methylation patterns in quantitative terms. The gold-standard technology for detection of 5mC is direct bisulfite sequencing (BS-seq or MethylC-seq), which maps sites at single base resolution and provides unbiased DNA methylation level estimations (Harris et al., 2010; Laird, 2010; Suzuki and Bird, 2008). However, one drawback of direct bisulfite sequencing is the excessive cost necessary to achieve the depth required for accurate DNA methylation mapping and this largely limits its application for DNA methylation profiling for multiple samples in a quick manner. Since a huge proportion of the sonication generated DNA reads for direct bisulfite sequencing contain no CpG sites at all, providing no information for DNA methylation studies, one good strategy to reduce cost and save time is to enrich the effective DNA fragments before sequencing.

Reduced representation bisulfite sequencing (RRBS) involves restriction enzyme digestion and size selection of the digested DNA fragments to generate a representative sampling before bisulfite conversion and sequencing. Because every digested read are required to contain at least one CpG, the intended information can be obtained with less cost and time, thus allowing large-scale comparative studies of DNA methylation being practicable and feasible.

* Corresponding author at: Beijing Genomics Institute at Shenzhen, Shenzhen 518000, China. Fax: +86 755 25273796.
E-mail address: zhangxq@genomics.org.cn (X. Zhang).
[1] These authors contributed equally to this work.

0168-1656/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2011.06.034

RRBS was first applied using a mouse sample (Meissner et al., 2005; Smith et al., 2009) and later it was developed and applied for genome-scale DNA methylation maps of mammalian pluripotent and differentiated cells (Lister et al., 2009; Meissner et al., 2008; Straussman et al., 2009) as well as clinical samples (Gu et al., 2010). However, to our knowledge, there was not a systematic assessment of RRBS for DNA methylation profiling, which should involve some critical parameters in sequencing-based DNA methylation profiling technology such as genomic coverage, coverage depth, reproducibility and accuracy. Also, no evaluation had been carried out for the data quantity needed for reliable estimations of DNA methylation level in human samples.

To achieve the above information and to determine how well RRBS may fit for the studies on human methylomes, we decided to apply RRBS to the human peripheral blood mononuclear cells (PBMC) from the Asian individual whose genome had been deciphered in the YH project (Wang et al., 2008) and whose methylome had been revealed recently by BS-seq (Li et al., 2010). The availability of the YH genome and methylome facilitated our assessment on the performance of RRBS while in turn, the first constructed MspI digested RRBS libraries for YH genome would complement the BS-seq data and provide further information upon DNA methylation status within multiple genomic features based upon intrinsic CpG context.

2. Materials and methods

2.1. Data availability

The RRBS data have been deposited into http://yh.genomics.org.cn/.

2.2. Public data used

The YH reference and the BS-seq data of PBMC were downloaded from YH database (http://yh.genomics.org.cn/). Annotations of CpG islands, Genes, repeats and non-coding RNAs were downloaded from the UCSC database (http://genome.ucsc.edu).

2.3. In silico simulation of enzyme digestion and size selection

Before our wet-lab manipulations, in silico simulation of enzyme digestion and size selection were conducted, which provided the theoretical values for RRBS to the YH methylome.

For mainly three reasons, MspI has been chosen for digestion of a human genome: (1) DNA methylation tends to occur in the CpG sites within genomes of mammals; (2) unlike its isoschizomer HpaII, MspI can cleave the 'CCGG' sequence whenever the internal C residue is methylated; (3) 'CCGG' occurs in a rather high frequency within the human genome (data not shown).

Various sizes of the digestion fragments can be chosen and the basic principle is to achieve maximum effective information with minimal cost. 40–220 bp band has been declared as being able to achieve fairy good coverage over several regions of interest in previous studies (Meissner et al., 2008). We plotted the length distribution of the MspI digested DNA fragments and found peaks occurred at 68 bp, 134 bp, and 205 bp (Sup Figure 1). To evaluate whether a narrower band rather than the 40–220 bp band could sufficiently provide good coverage and coverage depth of our regions of interests, we first analyzed bands ranging from 40 to 110 bp and 110 to 220 bp respectively. However, neither of the narrower bands could provide satisfactory results and thus we combined data from 40 to 110 bp and 110 to 220 bp bands for further assessments.

2.4. RRBS library preparation and bisulfite sequencing

The overall workflow was summarized in Supplemenatl Figure 2.

The mononuclear cells were separated through Ficoll-Paque (GE Healthcare) gradient centrifugation. The total DNA was prepared by proteinase K/phenol extraction.

Before library construction, 100 U MspI (NEB R0106L) was used to digest the 5 μg genomic DNA of PBMC from YH (37 °C, 16 h). And then MspI was inactivated by warming (80 °C, 20 min).

After digestion by MspI, the Ilumina Paired-End protocol was made use of to construct the libraries, including the procedures of end repair, <A> base addition and methylated-adaptor ligation. The concentration of primer is 10 μmol/l and 1 μl of primer was added in 50 μl PCR system.

DNA libraries of 40–110 and 110–220 base pairs were excised from the 2% TAE agarose gel afterwards. On account of the adaptor size of 115 bp, the 155–225 bp and 225–335 bp bands were actually excised respectively.

Then the excised DNA was recovered by columns, purified by MiniElute PCR Purification Kit (QIAGEN 28006), eluted in 20 μl of EB and bisulfite was converted using the EZ DNA Methylation-Gold kit (ZYMO). All the bisulfite converted products were amplified by PCR in a final reaction volume of 50 μl consisting of 10 μl purified DNA, 4 μl 2.5 mM dNTP, 1 μl PCR primer 1.0, 1 μl PCR primer 2.0, 5 μl 10× buffer, 0.5 μl JumpStart™ TaqDNA polymerase and 28.5 μl UltraPure™ Water. The following PCR Thermal cycling program was 94 °C 60 s, 11 cycles (11 cycles for 155–225 bp, 13 cycles for 225–335 bp) of 94 °C 30 s, 58 °C 30 s, 72 °C 30 s, 72 °C 5 min and products were hold at 4 °C. PCR products were purified and recovered, followed by sequencing with an IlluminaGAII.

2.5. Bioinformatic processing of the data

After removal of the adaptor sequences, the 49 bp reads from each of the biological replicates were aligned to genome reference as well as the size-selected MspI fragments generated by our in silico simulation. Because of the strand specificity of DNA methylation, two rounds of alignments were carried out, i.e. the bisulfite converted reads were aligned to the genome sequences termed the "T genome" with each cytosine converted to thymine and in the meanwhile the reads were also aligned to the genome sequences termed the "A genome" with each guanine converted to adenosine. The alignments were carried out with BGI SOAPaligner verison2.01 (Li et al., 2009), allowing up to two mismatches for successful mapping. Summary of the data quantity after each step of filtration is shown in Supplemental Table 1 and both replicates showed near complete (>99%) bisulphite conversion of non-CpG cytosines (data not shown).

2.6. Assessment of RRBS

We assessed the coverage, mean coverage depth, the genomic distribution of the detected CpG sites and reproducibility of RRBS with an increasing quantity of dataset. We also carried out comparisons between RRBS and BS-seq considering their coverage and coverage depth as well as the concordance of DNA methylation level between RRBS and BS-seq to YH genome. The statistical process was performed as followed applying Perl and R programming:

1) *Coverage of the genome*: the number of cytosines that were detected by RRBS and that met certain requirements (for example, covered with at least 4 reads) divided by the total number of cytosines within the YH genome.*Coverage of the ROIs*: the