Briefings in Bioinformatics, 22(6), 2021, 1–10

https://doi.org/10.1093/bib/bbab201
Problem Solving Protocol

A new approach to decode DNA methylome and genomic variants simultaneously from double strand bisulfite sequencing

Jialong Liang†, Kun Zhang†, Jie Yang, Xianfeng Li, Qinglan Li, Yan Wang, Wanshi Cai, Huajing Teng and Zhongsheng Sun

Corresponding author: Zhongsheng Sun, Beijing Institutes of Life Science, Chinese Academy of Sciences, Beichen West Road, Chao Yang District, Beijing 100101, China. Tel.: +86 10 64864959; Fax: +86 10 84504120. Email: sunzs@biols.ac.cn; Huajing Teng, Department of Radiation Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Fucheng Road, Haidian District, Beijing 100142, China. Tel.: +86 10 88196505. Email: hjteng@bjmu.edu.cn
†These authors contributed equally to this work.

Abstract

Genetic and epigenetic contributions to various diseases and biological processes have been well-recognized. However, simultaneous identification of single-nucleotide variants (SNVs) and DNA methylation levels from traditional bisulfite sequencing data is still challenging. Here, we develop double strand bisulfite sequencing (DSBS) for genome-wide accurate identification of SNVs and DNA methylation simultaneously at a single-base resolution by using one dataset. Locking Watson and Crick strand together by hairpin adapter followed by bisulfite treatment and massive parallel sequencing, DSBS simultaneously sequences the bisulfite-converted Watson and Crick strand in one paired-end read, eliminating the strand bias of bisulfite sequencing data. Mutual correction of read1 and read2 can estimate the amplification and sequencing errors, and enables our developed computational pipeline, *DSBS Analyzer* (https://github.com/tianguolangzi/DSBS), to accurately identify SNV and DNA methylation. Additionally, using DSBS, we provide a genome-wide hemimethylation landscape in the human cells, and reveal that the density of DNA hemimethylation sites in promoter region and CpG island is lower than that in other genomic regions. The cost-effective new approach, which decodes DNA methylome and genomic

Jialong Liang is a PhD candidate of Beijing Institutes of Life Science, Chinese Academy of Sciences. His research interests include genomics and bioinformatics.
Kun Zhang is a master student of Institute of Genomic Medicine, Wenzhou Medical University. His research interests include bioinformatics.
Jie Yang is a master student of Institute of Genomic Medicine, Wenzhou Medical University. Her research interests include genomics.
Xianfeng Li received his PhD training at Beijing Institutes of Life Science, Chinese Academy of Sciences. His research interests include genomics and bioinformatics.
Qinglan Li is a PhD candidate of Beijing Institutes of Life Science, Chinese Academy of Sciences. Her research interests include genomics.
Yan Wang is an associate professor of Beijing Institutes of Life Science, Chinese Academy of Sciences. Her research interests include genomics.
Wanshi Cai is a PhD candidate of Beijing Institutes of Life Science, Chinese Academy of Sciences. His research interests include genomics and bioinformatics.
Huajing Teng is an associate professor of Department of Radiation Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education) at Peking University Cancer Hospital and Institute. His research is in the areas of bioinformatics and genomics.
Zhongsheng Sun is a professor of Beijing Institutes of Life Science, Chinese Academy of Sciences, CAS Center for Excellence in Biotic Interactions and State Key Laboratory of Integrated Management of Pest Insects and Rodents, University of Chinese Academy of Sciences. His research is in the areas of bioinformatics and genomics.
Submitted: 10 March 2021; **Received (in revised form):** 23 April 2021

variants simultaneously, will facilitate more comprehensive studies on numerous diseases and biological processes driven by both genetic and epigenetic variations.

Key words: genomic mutation; cytosine modification; CpG context; epigenomic alteration; population genomics

Introduction

Cytosine methylation of DNA is one of the extensively studied epigenetic modifications involved in the regulation of various diseases and biological processes, including the initiation and progression of cancers, embryonic development, X chromosome inactivation, genomic imprinting and silencing of transposable elements [1–6]. Alteration of DNA methylation was also unequivocally associated with population genetic variation [7, 8], subclonal evolution of tumor tissues [9–11], evolutionary divergence of duplicate genes, exon usage and local adaptation or rapid phenotypic changes of specific populations [12–16]. Simultaneous identification of genetic variation and DNA methylation was crucial for exploring the genetic and epigenetic contributions to disease and various evolutionary processes [7, 17–19]. In addition, the identification of single-nucleotide variants (SNVs) is critical for identification of allele-specific epigenetic events [20, 21], which is a driving force that led to genomic imprinting [22, 23].

Genome-wide base resolution analysis of DNA methylation is enabled by bisulfite treatment and subsequent massive parallel sequencing [24]. During bisulfite treatment, unmethylated cytosine is converted to uracil, whereas methylated cytosine and the guanine on the opposite strand is not affected. Based on this principle, several softwares, such as MethylExtract [25], Bis-SNP [26], BS-SNPer [27] and CGmapTools [20], have been developed to calling SNVs in bisulfite sequencing data. However, the sensitivity and accuracy of these tools in calling SNVs were limited by the inevitable defects of traditional bisulfite sequencing methods. Since bisulfite treatment could lead to separation of the two DNA strands, DNA degradation and reduced genome sequence complexity, which subsequently resulted in the high strand bias and alignment errors of sequencing data [28–31]. Additionally, C > T is the most frequent substitution in the population (for example it accounts for 35% of all SNPs in human dbSNP database), it is difficult to distinguish the real C > T mutations from the bisulfite induced C > T conversions. Thus, an experimental innovation and improvement on calling SNVs and methylation level from bisulfite sequencing is urgently required.

Here, we developed double strand bisulfite sequencing (DSBS) through locking Watson and Crick strand together followed by bisulfite treatment, massive parallel sequencing and computational calling for simultaneous and precise identification of DNA methylation, hemimethylation and SNVs, which offers a cost-effective and useful approach for exploring the contributions of genetic and epigenetic variations to numerous diseases and biological processes.

Materials and methods

Samples and DNA extraction

Human ovarian epithelial cell line T29 was provided by Dr Jinsong Liu (MD Anderson Cancer Center, University of Texas, TX, USA) and the cell line was cultured as previously described [32–34]. Genomic DNA was extracted from T29 cell line by using Qiagen DNeasy Blood & Tissue Kit (Qiagen, Germany), and 20 mg/ml RNase (Qiagen, Germany) was added to avoid contamination of RNA in the DNA samples. DNA was quantified by using a Qubit 2.0 fluorometer (Life Technologies, Carlsbad, CA, USA).

Adapter synthesis

Hairpin linker adapter oligonucleotides (Sequence 5′-pho-CG^{m}C^{m}CAGGTGGCAAGTGAAGCCACCTGGCGT-3′) were synthesized by Invitrogen Company (Shanghai, China). The synthesized oligonucleotide was diluted to a concentration of 1 mM, and denatured in a 95°C water bath, and then annealed by adding cold water to the final concentration of 100 μM.

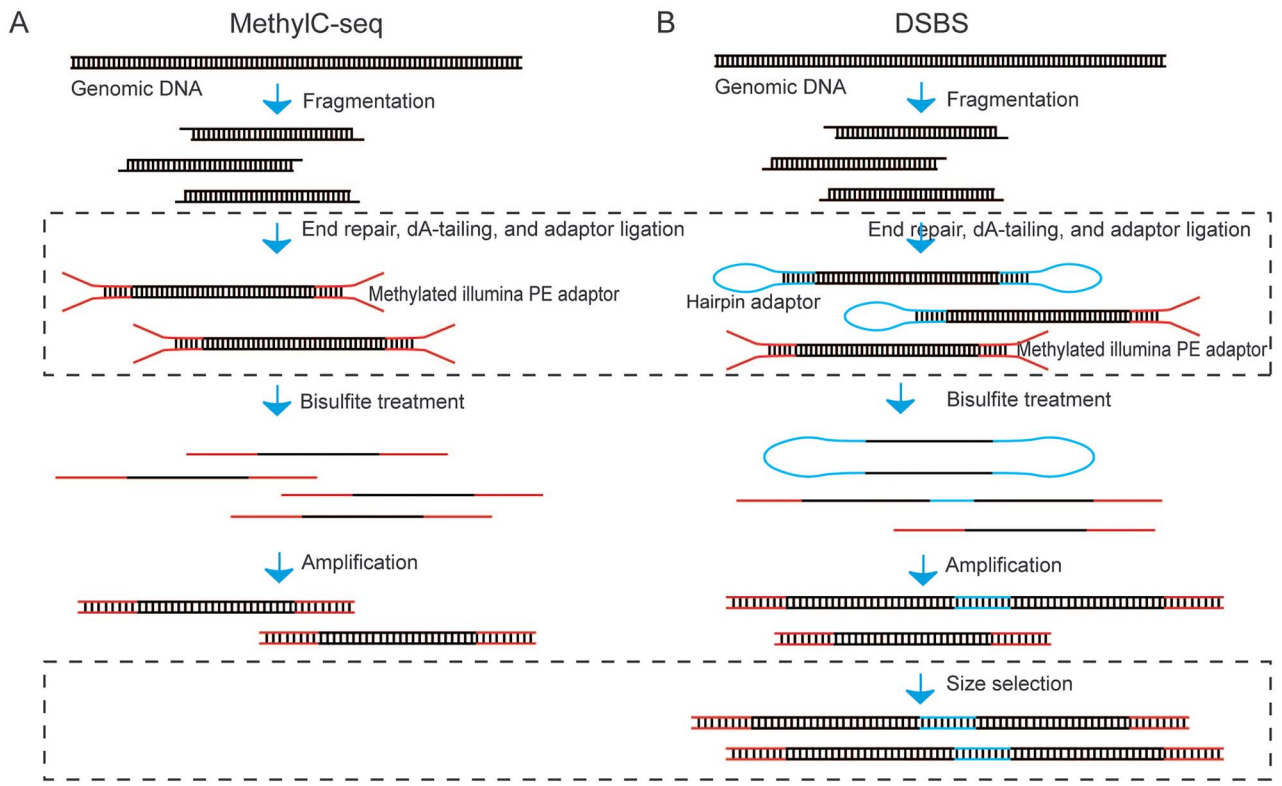
Preparation of DSBS library

Approximately 1 μg genomic DNA was fragmented to 200 bp double-stranded DNA fragments by using a Covaris E210 sonicator (Covaris Inc., MA, USA). Fragmented DNA was end-repaired and dA-tailed, and ligated with Illumina TruSeq adapter (all Cs methylated) and the Hairpin linker adapter with a ratio of 1:2:60 using KAPA Hyper Prep Kit (KAPA Biosystems, USA). The adapter-ligated DNA was treated by bisulfite using EpiTect Fast DNA Bisulfite Kit (Qiagen, Germany), and then amplified using 2× KAPA HiFi Uracil+ Readymix (KAPA Biosystems, USA). After the amplification, agarose gel-based size selection was performed to generate the sequencing library with size range of 400–800 bp (Figure 1). The DSBS libraries were sequenced for 150 bp paired-end reads using an Illumina Hiseq X Ten sequencer.

Preparation of whole-genome sequencing, target sequencing and MethylC-seq libraries

For whole-genome sequencing (WGS), about 1 μg genomic DNA was fragmented to 400 bp using a Covaris E210 sonicator (Covaris Inc., MA, USA). Fragmented DNA was end-repaired and dA-tailed, and ligated to Illumina TruSeq adapter using KAPA Hyper Prep Kit (KAPA Biosystems, USA). The adapter-ligated DNA was amplified by using 2× KAPA HiFi Hotstart Readymix (KAPA Biosystems, USA) to produce the sequencing library. To evaluate the accuracy of our SNV calling, targeted enrichment of whole exome regions was performed using AIExome Enrichment Kit V2 (iGeneTech, Beijing, China) to generate whole exome library, and then the library was sequenced for 150 bp paired-end reads using Illumina NovaSeq 6000 sequencer with average depth of 389×.

For MethylC-seq, about 1 μg genomic DNA was fragmented to 400 bp using a Covaris E210 sonicator (Covaris Inc., MA, USA). Fragmented DNA was end-repaired and dA-tailed, and ligated with Illumina TruSeq adapter (all Cs methylated) using KAPA Hyper Prep Kit (KAPA Biosystems, USA). The adapter-ligated DNA was bisulfite-treated using EpiTect Fast DNA Bisulfite Kit (Qiagen, Germany), and then amplified by using 2× KAPA HiFi Uracil+ Readymix (KAPA Biosystems, USA) to produce the sequencing library. Both WGS and MethylC-seq libraries were sequenced for 150 bp paired-end reads using an Illumina Hiseq X Ten sequencer.



C

Procedure	Step1 DNA fragmentation	Step2 End-repair and dA-tailing	Step3 Ligation	Step4 DNA cleanup	Step5 Bisulfite treatment and DNA cleanup	Step6 DNA amplification	Step7 Gel extraction and DNA cleanup
Timing	30 min	1 h	30 min	30 min	2 h	1 h	2 h
Safe Stopping Point	√			√		√	

Figure 1. Experimental diagram of double strand bisulfite sequencing (DSBS) and whole-genome bisulfite sequencing (MethylC-seq). (A) MethylC-seq: Genomic DNA is fragmented to 300–400 bp and ligated to methylated Illumina adapters. The ligated fragments are treated with bisulfite and amplified by PCR using Illumina paired-end PCR primers. (B) DSBS: Genomic DNA is fragmented to 150–200 bp and ligated to methylated Illumina adapters and hairpin adapters. The ligated fragments are treated with bisulfite and amplified by PCR using Illumina paired-end PCR primers. PCR product of longer than 400 bp are size-selected on agarose gel and sequenced on the illumine platform. (C) Flowchart regarding the timeline of the preparation of DSBS library.

Processing of DSBS sequencing data

The high-quality clean reads were generated by filtering out low-quality reads and removing adapter sequences using CutadaptV1.11 (https://github.com/marcelm/cutadapt). The clean reads were aligned to the GRCH37 human reference genome by using BSMAP [35]. The local sequence realignment and recalibration was performed by using Bis-SNP as previously described [26]. SNV calling was performed by using an inhouse python program *DSBS analyzer*, which was developed based on python3, and depends on the package of pysam, pyfasta and tabix. The output files contained the genome location of SNVs, DNA methylation levels and DNA hemimethylation levels. The source code of *DSBS analyzer* was uploaded to the website: https://github.com/tianguolangzi/DSBS.

In details, due to the low sequence complexity and high repetitive rates of bisulfite-converted genome, the alignment accuracy of bisulfite-converted sequence data was relatively lower compared with genome sequence data. To realize the high accuracy of SNV calling, a series of parameters were set to filter the invalidly aligned bisulfite-sequencing reads before the identification of SNVs and DNA methylation levels using *DSBS analyzer*. The detailed parameters include: (i) considering that the read 1 and read 2 of paired-reads were derived from one DNA fragment, the paired-end reads mapped to different locations in the reference genome were filtered, and paired-reads with overlap size shorter than 50 bp were discarded; (ii) considering that the alignment accuracy of short reads was lower, reads with length shorter than 50 bp were discarded; (iii) sequencing reads with the N bases exceeding 5 would be discarded, and sequencing reads with the ratio of bases exceeding 0.6 with sequencing quality score <20 would be discarded; (iv) sequencing reads with the amount of small insertion and deletion exceeding 1 would be discarded; (v) the paired-reads in which the number of mismatches is greater than 5 would be discarded. In addition, when identifying the SNVs and DNA methylation levels, (i) the sequencing quality score of the bases should exceed 20 in both paired-reads; (ii) there should less than two mismatches in the continuous 10 bp; (iii) the SNVs within 5 bp range of an indel

would be discarded; (iv) the allele frequency of SNVs should be higher than 0.15. According to the paired bisulfite-converted sequencing reads, *DSBS analyzer* could determine the original sequence and DNA methylation status of the double-strand DNA according to Figure 2.

To facilitate the use of DSBS *analyzer*, we have developed a pipeline based on python3, which included six steps: quality control, cleaning, genome alignment, realignment, SNV calling and evaluation of the DNA methylation and hemimethylation levels, and annotation. FastQCV0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/FastQC) and BAMQC (https://github.com/s-andrews/BamQC) were used in the quality control step. Cutadapt was used to cut the hairpin adapter and PE adapter sequence. The alignment of sequencing reads to the reference genome sequence was performed by using BSMAP [35]. Bis-SNP [26], which based on GATK, was used to realign the sorted BAM file. As the duplication of sequencing data dramatically influence the SNV calling, Fastuniq (https://github.com/dcjones/fastq-tools) was used to remove the duplicates before cutting the adapters, and the function of removing duplication in Samtools [36] and PicardV2.9 (http://broadinstitute.github.io/picard/) were used after the realignment of sorted BAM files. Finally, *DSBS analyzer* was used to call SNVs and assess the DNA methylation and hemimethylation levels. As for the computational resources and cost of *DSBS analyzer*, in this study performed on a single computing node with the CPU of 32 cores, memory of 48G and 64 threads, the total time for data cleaning, mapping, local sequence realignment and recalibration of 150 Gb raw DSBS data is about 116.76 h using 40 threads, and the time for calling SNVs, the DNA methylation and hemimethylation levels is about 4.09 h using 40 threads.

Results

Experimental and SNV calling strategy for DSBS

Fragmented genomic DNA was ligated with the hairpin adaptor and methylated Illumina adaptor simultaneously, and then treated with bisulfite and amplified by ligation-mediated polymerase chain reaction (PCR). The amplified library was size-selected to enrich library linked with hairpin adaptor for high throughput paired-end sequencing (Figure 1). Although both sequencing reads of DSBS and whole-genome bisulfite sequencing (MethylC-seq) are directional, only the sequence information of one bisulfite-converted strand of double-strand DNA was generated from one pair of paired-end reads in the MethylC-seq, whereas in the DSBS, bisulfite-converted Watson and reverse complement of bisulfite-converted Crick or bisulfite-converted Crick and reverse complement of bisulfite-converted Watson strand derived from the same DNA fragment were simultaneously sequenced in reads 1 and reads 2, respectively. Thus, the sequence of two bisulfite-converted strands derived from one double-strand DNA fragment could be obtained from one paired-end reads of DSBS.

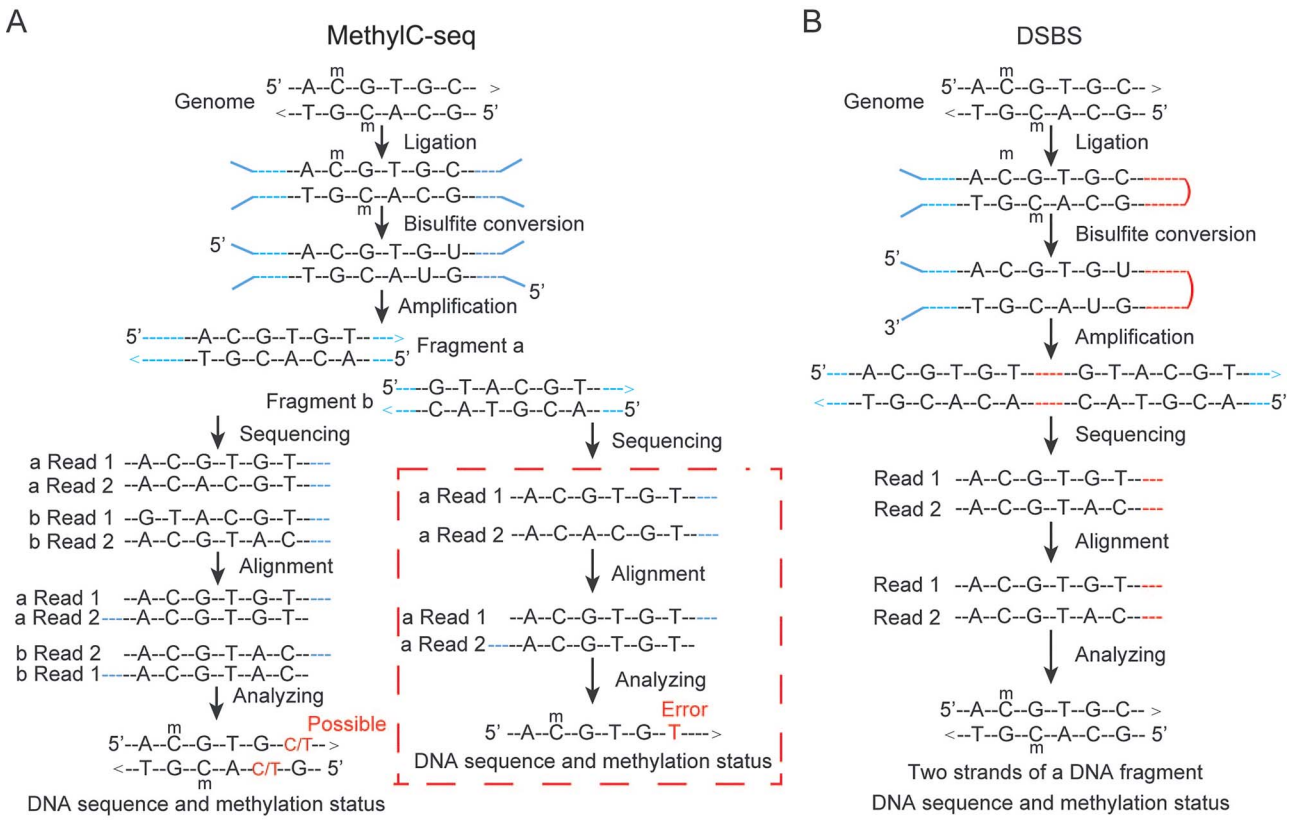
During the bisulfite treatment, unmethylated cytosine is converted to uracil, whereas the guanine on the opposite strand is not affected. This property was exploited to distinguish between bisulfite conversion of C > T and single nucleotide variation from C to T. Because in the MethylC-seq, bisulfite-converted Watson and Crick strands were separated during the bisulfite treatment, and only one strand was sequenced in large fraction of genomic regions. Therefore, C > T SNVs were difficult to be distinguished from C > T conversion induced by bisulfite treatment in MethylC-seq data. In the DSBS, bisulfite-converted Watson and Crick strand were locked together by hairpin adaptor, and simultaneously sequenced in the paired-end reads, the original sequence and modification status could be accurately deduced from the paired-end sequencing reads.

Bisulfite conversion only occurs on unmethylated Cs, and have no effect on Gs on the complementary strand, therefore, bisulfite-converted Watson and Crick strand are not fully reverse complementary to each other. In the paired-end sequencing of DSBS, bisulfite-converted Watson and reverse complement of bisulfite-converted Crick; or bisulfite-converted Crick and reverse complement of bisulfite-converted Watson derived from one double-strand DNA fragment, reflected by read 1 and read 2, should be aligned to the same position of the reference genome. These paired-end reads aligned to different positions were considered as alignment error, and not used for the calling of SNVs. Therefore, the false positive SNVs caused by alignment errors could be eliminated. In the paired-end reads of DSBS, read 1 and read 2 were different in Cs and Gs, whereas same in As and Ts (Figure 2). Due to the sequencing and amplification errors, positions which were not in accordance with this principle, were excluded for SNV calling. Therefore, when calling SNVs in DSBS sequencing data, the mutual correction of read1 and read2 could eliminate the effect of sequencing, amplification and alignment errors. We provided the bioinformatic pipeline, *DSBS Analyzer*, for SNV calling, evaluation of DNA methylation and hemimethylation level in DSBS sequencing data, which is freely available at https://github.com/tianguolangzi/DSBS.

Evaluation of SNV calling by DSBS

To evaluate the minimum sequencing data required by DSBS, we plotted the saturation curve of sequencing data. The 10× genome-wide coverage was saturated in about 150 Gb sequencing data. We then compared the coverage of genome fraction of DSBS and WGS and observed DSBS and WGS covered 91.2% and 91.3% of genome fractions in depth ≥ 1, and covered 88.7% and 90.2% of genome fractions in depth ≥ 10, respectively (Figure 3). By analyzing the coverage of DSBS and WGS in different genomic elements, we revealed that the coverage of DSBS had no bias in different genomic elements.

To further evaluate the sensitivity and accuracy of DSBS in SNV calling, we compared the performance of SNV calling of DSBS with two published high-performance computational pipelines in calling SNVs from methylation sequencing data, Bis-SNP [26] and BS-SNPer [27], and WGS. With the average sequencing depth of 30×, we identified SNVs in WGS data of human ovarian epithelial cell line, T29, by using four SNV calling softwares, GATK, VarScan, Bcftools, Freebayes, respectively. Of 3 071 837 SNVs simultaneously identified by these four softwares, 2 839 016 (92.42%) SNVs could be identified by DSBS, while only 1 758 715 (57.25%) and 2 308 667 (75.16%) SNVs could be identified by Bis-SNP and BS-SNPer, respectively (Figure 3). Regarding a total of 3 398 625 SNVs identified by DSBS, 3 148 862 (92.65%) SNVs could be validated by WGS, which revealed a low false positive rate (7.35%) of SNV calling by DSBS. As for 3 579 175 SNVs called by Bis-SNP, only 2 069 360 (57.82%) could be validated by WGS, and 3 317 957 SNVs called by BS-SNPer, only 2 577 798 (77.69%) could be validated by WGS (Figure 3). As the vast majority of adaptation or disease-related SNVs were located at nonrepeat regions, we further evaluated the accuracy of SNV calling by DSBS in nonrepeat regions, and observed the accuracy of SNV calling by DSBS in nonrepeat regions was 95.4%, which was higher than that in the whole-genome regions. Considering a proportion of these so called false positive SNVs in DSBS could be the truly occurred SNVs with lower allelic frequency, we then



C

Status of original base	Status of original base on opposite strand	The detected base on reads 1	The detected base on reads 2
A	T	A	A
T	A	T	T
^{m}C	G	C	C
G	^{m}C	G	G
C	G	T	C
G	C	G	A

Figure 2. Strategy of genetic variants and methylation status calling in double strand bisulfite sequencing. **(A)** In MethylC-seq, bisulfite-converted Watson strand and reverse complement of bisulfite-converted Watson strand were sequenced in read 1 and read 2 of paired-reads. Two strands of a DNA fragment were separately sequenced, or only one strand of a DNA fragment was sequenced due to the sequencing bias. (B) Bisulfite-converted Watson strand and reverse complement of bisulfite-converted Crick strand derived from the same double-strand DNA fragment were sequenced in read 1 and read 2, and aligned to the same position on reference genome. By simultaneous analyzing the sequence of read 1 and read 2, the sequence and DNA methylation state of DNA fragment could be deduced. (C) Base status of the double-strand bisulfite-sequenced DNA.

evaluated the SNVs called by DSBS but not identified by any of the four SNV calling softwares based on WGS data. Of the 1 984 exonic SNVs called by DSBS but not by WGS, we observed that 526 (26.51%) exonic SNVs could be validated by deep sequencing (with average depth of 389×) of whole exome regions. Thus, we inferred the actual false positive rate of DSBS was lower than that mentioned previously.

Evaluation of the DNA methylation levels at CpG sites by DSBS

To evaluate the performance of the DNA methylation level identified by DSBS, we compared the coverage of DSBS sequencing data with the data of MethylC-seq, the golden standard of DNA methylation analysis. Analyzing 150 GB clean data generated by DSBS and MethylC-seq, respectively, we observed DSBS covered 92% whereas MethylC-seq covered 90% of the total CpGs in the human genome with sequencing depth ≥ 2, and DSBS and MethylC-seq both covered 89% of the total CpGs in the human genome with sequencing depth ≥ 10 (Figure 4). By analyzing the coverage of DSBS and MethylC-seq in different genomic elements, we also revealed that the coverage of DSBS was similar in different genomic elements.

To evaluate the reproducibility of DSBS in evaluating the DNA methylation levels, we count the DNA methylation levels of CpGs with at least 10-fold sequencing depth, and determined the Pearson correlation coefficient value between two DSBS experimental replicates. The Pearson correlation coefficient value between two DSBS replicates was 0.939, indicating an excellent reproducibility of DSBS in evaluating the DNA methylation levels (Figure 4). We then compared the DNA methylation levels of DSBS with the one from MethylC-seq by evaluating

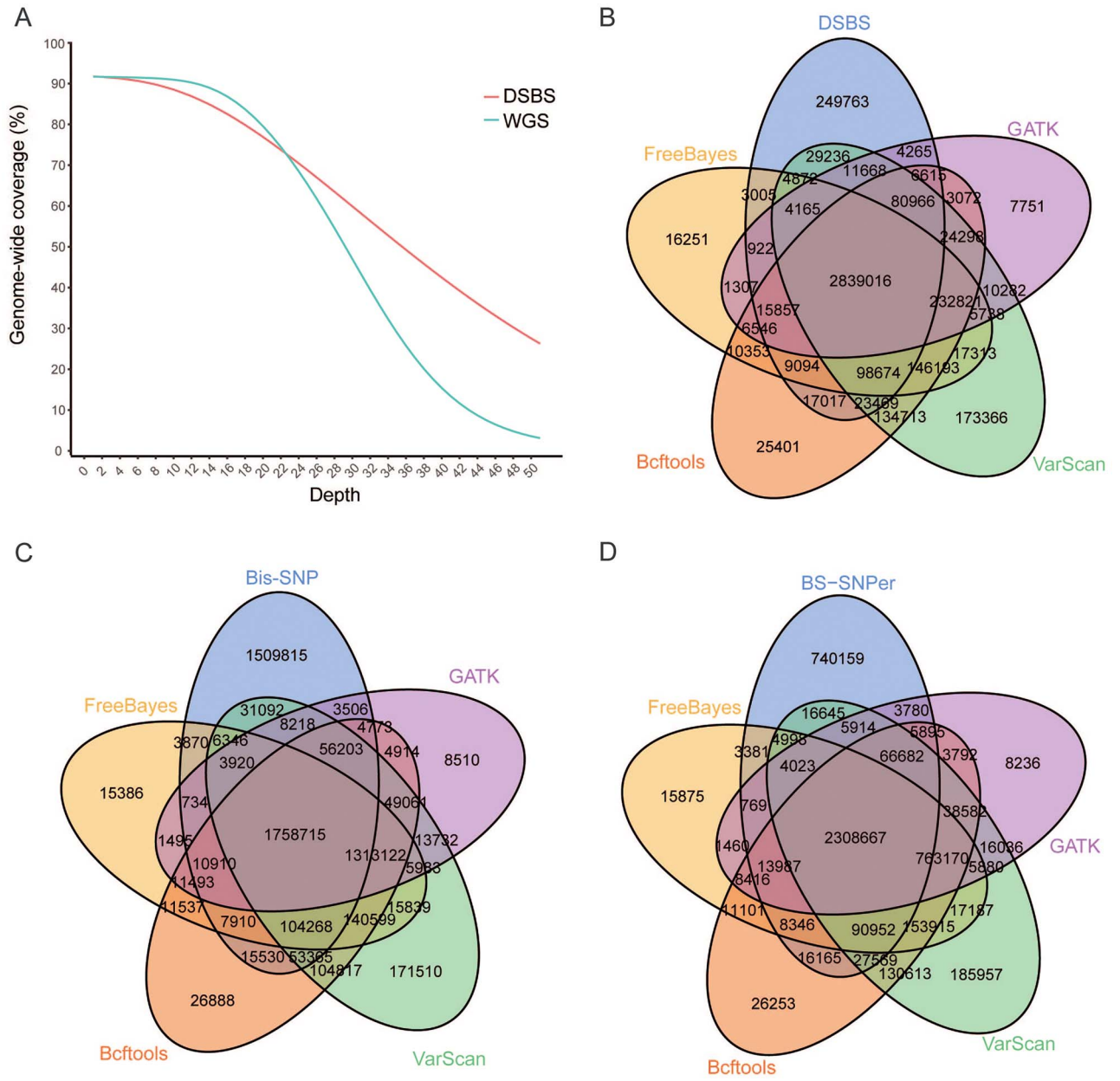


Figure 3. The comparison of double strand bisulfite sequencing (DSBS) and whole-genome sequencing (WGS) in identifying SNVs. (A) Genome fraction coverage of DSBS and WGS. X-axis denotes sequencing depth and y-axis denotes the fraction of genome that is at or above a given sequencing depth. Venn diagram shows the overlap of SNVs identified by DSBS (B), Bis-SNP (C), BS-SNPer (D) and WGS using different tools.

all CpGs identified by DSBS from MethylC-seq with sequencing depth ≥ 10 in T29 cell line. Our result demonstrated that the DNA methylation level for DSBS versus MethylC-seq were correlated at Pearson correlation coefficient of 0.947 (Figure 4).

Signature of DNA hemimethylation across genomic regions

DNA hemimethylation, with only one of the two complementary DNA strands methylated, was intermediate product in DNA methylation maintenance during DNA replication, and it was regarded to regulate DNA methylation inheritance [37–39]. It was reported that the majority of intermediately (40–60%) methylated CpG dinucleotides were hemimethylated [40], and 10% of CpGs in embryonic stem cells (ESCs) and trophoblast stem cells remain hemimethylated, which could regulate chromatin interaction and transcription [40, 41]. DSBS simultaneously sequenced bisulfite-converted Watson and Crick strand of bisulfite-treated DNA, which enables it to precisely identify hemimethylation sites. Although previous study has performed to investigate the DNA methylation fidelity during mouse ESC self-renewal and differentiation [40, 41], it only analyzed the CpG-rich regions, and the distribution of DNA hemimethylation in the whole genome and different genomic regions have not been evaluated. Here, we depicted the genome-wide hemimethylation signature of T29 cell line by using DSBS, and observed the average hemimethylation level of CpGs (depth ≥ 10) was 0.1304. We identified 14 547 891 hemimethylated CpGs (depth ≥ 10), including 6 812 824 sites with hemimethylation level ≥ 0.2 in the whole genome. To evaluate the distribution of

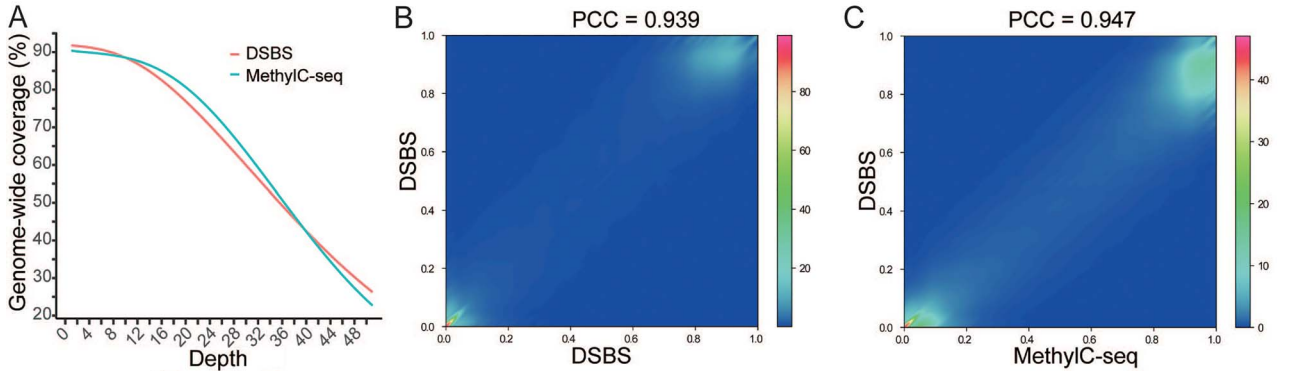


Figure 4. The comparison of double strand bisulfite sequencing (DSBS) and whole-genome bisulfite sequencing (MethylC-seq) in evaluating DNA methylation levels. (A) Genome fraction coverage of DSBS and MethylC-seq. *X*-axis denotes sequencing depth and *y*-axis denotes the fraction of genome that is at or above a given sequencing depth. (B) Scatter plots show Pearson correlation coefficient (PCC) of CpG methylation levels identified by two DSBS technical replicates ($n = 16\,509\,627$). (C) Scatter plots show PCC of CpG methylation levels identified by DSBS and MethylC-seq ($n = 15\,141\,119$).

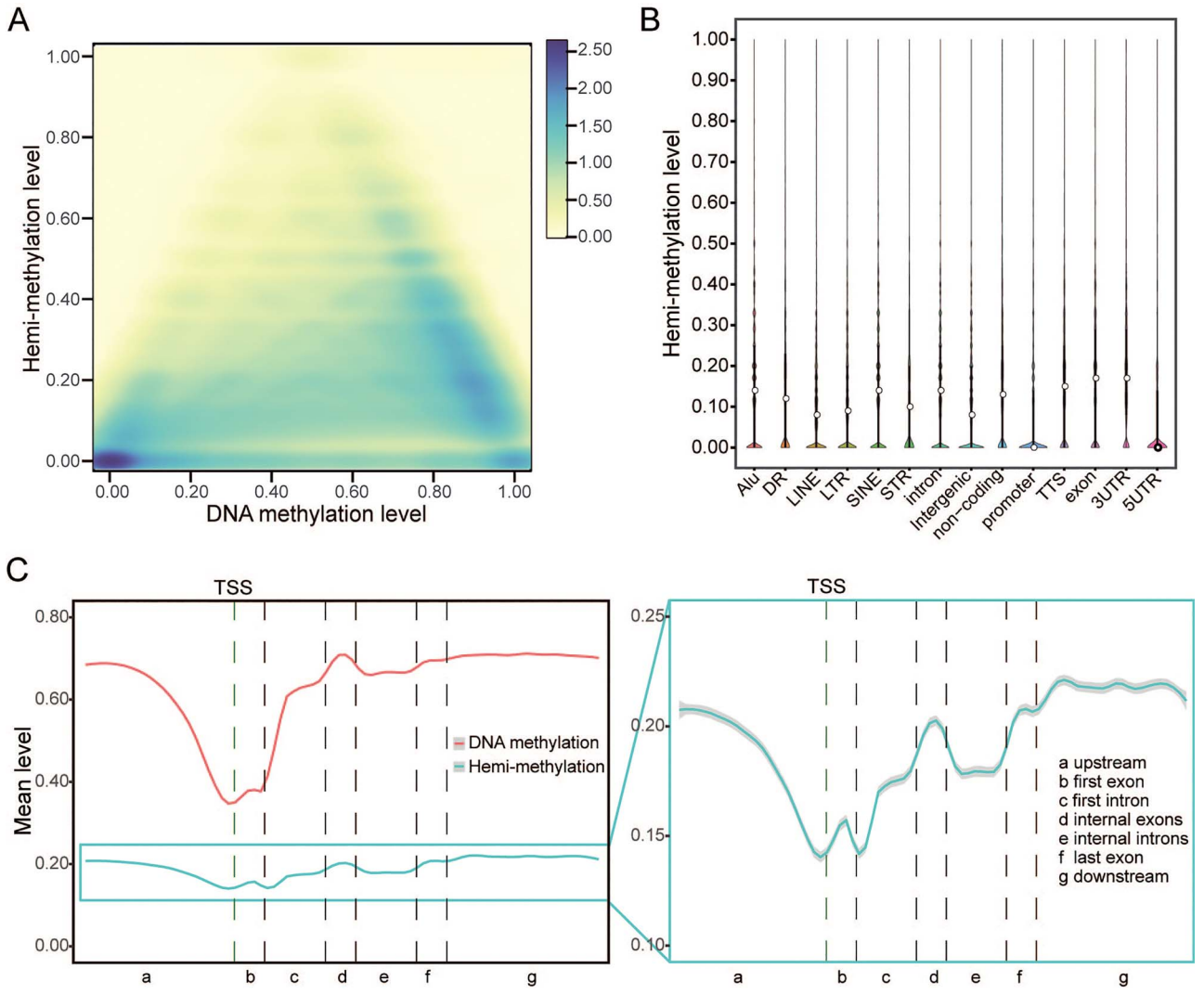
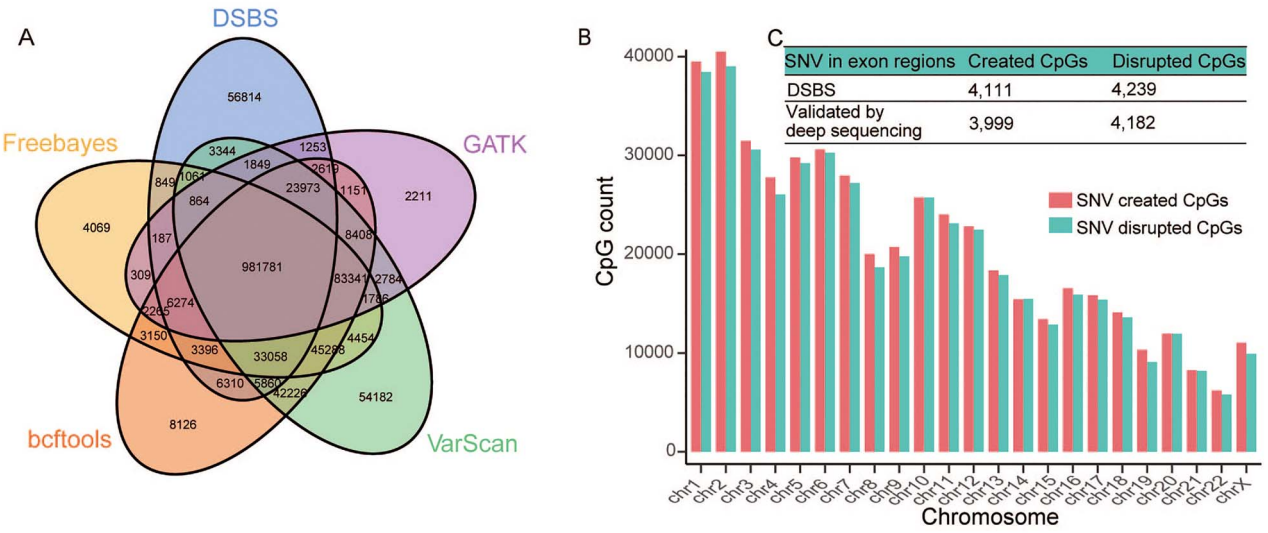


Figure 5. DNA hemimethylation levels across genomic regions. (A) Scatter plots showing the correlations between methylation level and hemimethylation level ($n = 21\,752\,893$). (B) The distribution of hemimethylation level in different genomic elements. (C) Metaplot of DNA methylation and hemimethylation levels across gene bodies. DR, direct repeat; LINE, long interspersed nuclear elements; LTR, long terminal repeat; SINE, short interspersed nuclear elements; STR, short tandem repeats; UTR, untranslated region; TSS, transcription start site; TTS, transcription termination site.



SNV in exon regions	Created CpGs	Disrupted CpGs
DSBS	4,111	4,239
Validated by deep sequencing	3,999	4,182

Figure 6. The comparison of double strand bisulfite sequencing (DSBS) and whole-genome sequencing (WGS) in identifying C > T SNVs. (A) Venn diagram shows the overlap of C > T SNVs identified by DSBS and WGS using different tools. Identification (B) and validation (C) of single-nucleotide variants (SNVs) which could disrupt the existed CpGs or create new CpGs relative to the reference genome in the human ovarian epithelial cell line.

DNA hemimethylation sites and the levels across the human genome, we analyzed the density and the average level of DNA hemimethylation in different genomic element regions. The results revealed that the density and the average level of DNA hemimethylation varied across different genomic elements (Figure 5), and the density of DNA hemimethylation sites in promoter region and CpG island were lower than that in other genomic regions.

To further investigate the relationship between DNA hemimethylation and methylation levels, we evaluated the hemimethylation levels and DNA methylation levels on each identified CpG (Figure 5). As for CpGs with the low DNA methylation levels (<0.5), the DNA hemimethylation level was positively associated with DNA methylation level (PCC = 0.277). While regarding the CpGs with high DNA methylation level (≥0.5), the DNA hemimethylation level was negatively associated with DNA methylation level (PCC = −0.392). We further evaluated the relationship between the hemimethylation levels and the DNA methylation levels on CpGs in different genomic element regions, and unveiled that both the DNA hemimethylation and DNA methylation levels were lower near the transcriptional start site than other genomic regions.

Genetic background should be considered in population epigenetic studies

C > T is the most frequent substitution in the population, for example it accounts for 35% of all SNPs in human dbSNP database. Through our DSBS pipeline, we identified 1 129 492 C > T/G > A substitution in the human ovarian epithelial cell line, and 1 072 678 (94.97%) of them could be identified by WGS (Figure 6). In addition, 8 049 (94.90%) of 8 482 exonic C > T/G > A substitution could be validated by deep sequencing of whole exome regions, suggesting the high reliability of our pipeline in identifying C > T mutations from the bisulfite-treated sequencing data. The substitution of C > T usually occurs in the CpG context, which are frequently methylated in vertebrate genomes [42]. Thus, the identification of SNVs in bisulfite sequencing data is essential for accurate quantification of the methylation levels. Here, in the human ovarian epithelial cell line, we identified 466 714 SNVs which disrupted the existed CpGs in the reference genome, and 482 461 SNVs which created new CpGs relative to the reference genome (Figure 6). We further validated the accuracy of these disrupted and created CpGs using deep sequencing of target regions, and 4 182 of 4 239 (98.66%) disrupted CpGs and 3 999 of 4 111 (97.28%) created CpGs residing in the exon regions were validated by whole exome deep sequencing, respectively (Figure 6). These SNVs could have impact on the evaluation of the DNA methylation level of CpG, while not considered in the evaluation of the DNA methylation levels in the traditional MethylC-seq method. To accurately evaluate the DNA methylation levels, we further recalculated the DNA methylation levels of these CpGs in consideration of SNVs. Compared with DNA methylation levels evaluated by MethylC-seq, the DNA methylation levels recalculated by DSBS in consideration of SNVs has dramatic differences.

Discussion

Genetic and epigenetic variations may affect each other, and alteration of DNA methylation has been reported to be involved in various evolutionary and biological processes [7–16]. However, in the previous epigenetic variation studies, genetic background was frequently ignored [43, 44], and the sensitivities and accuracies of several published computational pipelines in calling SNVs from the methylation sequencing data were limited by the inevitable defects of traditional bisulfite sequencing methods [20, 25–27]. To deal with the challenge, here, we developed DSBS by locking Watson and Crick strand together followed by bisulfite treatment, massive parallel sequencing and computational calling. Although the strategy of hairpin bisulfite sequencing has been described to improve the mapping efficiency and accuracy in quantitative detection of 5-methylcytosine [28, 31, 45, 46], an integrated experimental and computational pipeline that simultaneously assesses DNA methylation, hemimethylation and genomic variants is an advantage of our work. Using the approach, we unveiled approximately 0.95 million SNVs which could break the existed CpGs or create new CpGs relative to the

reference genome. Many SNVs located at regulatory elements may shape patterns of population epigenomic variation [8, 47, 48]. Recently, interaction between genetic and epigenetic variations has begun to be dissected during the evolution of populations or tumors, such as population divergence of recent human evolution [7], tumors' clonal evolution [9, 10], local adaptation or rapid phenotypic changes of specific populations [13–15]. Thus, the cost-effective new approach we provided here, which can be used for simultaneous decoding DNA methylome and genomic variants in numerous vertebrate species with reference genome, will facilitate more comprehensive studies on various kind of disease and biological processes driven by both genetic and epigenetic variations.

We acknowledge that our approach has some limitations. The bisulfite treatment we used could not distinguish DNA methylation and DNA hydroxymethylation [46, 49]. Thus, the existence of DNA hydroxymethylation in specific loci may influence the detection of DNA methylation [50]. In addition, bisulfite conversion of cytosine to thymine may lead to selective and context-specific DNA degradation [30, 51]. Displacement of bisulfite conversion by using enzymatically conversion methods including TAPS [51] and EM-seq [52], which have mild damage to DNA, may facilitate the application of DSBS method. Furthermore, we enriched the library with hairpin adapter by using size selection in DSBS, which might lead to a small fraction of library without hairpin adapter. Modification of hairpin adapter with biotin, and binding with streptavidin beads after ligation is an alternative fragment enrichment choice. To deal with the challenge of calling somatic mutations from methylation sequencing data, we will update the computational pipeline in identifying somatic mutations from the double strand bisulfite sequencing data in the future.

Key Points

- We offered a useful approach for simultaneously deciphering DNA methylome, hemimethylome and SNVs with high sensitivity and accuracy.
- Genetic background should be considered in population epigenomic studies.
- We provided a genome-wide hemimethylation landscape using DSBS.
- The DNA hemimethylation levels near the transcription start sites were relatively low, and different repetitive elements were found with different hemimethylation levels.

Data availability

The raw sequencing datasets including DSBS, WGS, MethylC-seq and deep sequencing of target exon regions were deposited at the NCBI (https://www.ncbi.nlm.nih.gov/bioproject/722313) under accession number PRJNA722313. The source code of *DSBS analyzer* was available via the website: https://github.com/tianguolangzi/DSBS.

Authors' contributions statement

H. T. and Z. S. conceived and designed the study. J. L., Y. W., H. T. and Z. S. wrote the manuscript. J. L., K. Z., X. L. and H. T. analyzed the data. J. L., J. Y., Q. L. and W. C. prepared the samples and collected data.

Funding

The National Natural Science Foundation of China (32071156 and 31872237), Guangzhou and Guangdong Key Project (202007030002 and 2018B030335001) and the State Key Laboratory of Integrated Management of Pest Insects and Rodents (IPM1918).

References

1. Waters SA, Livernois AM, Patel H, *et al*. Landscape of DNA methylation on the marsupial X. *Mol Biol Evol* 2018;**35**(2):431–9.
2. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* 2019;**20**(10):590–607.
3. Huh I, Wu X, Park T, *et al*. Detecting differential DNA methylation from sequencing of bisulfite converted DNA of diverse species. *Brief Bioinform* 2019;**20**(1):33–46.
4. Klein HU, Hebestreit K. An evaluation of methods to test predefined genomic regions for differential methylation in bisulfite sequencing data. *Brief Bioinform* 2016;**17**(5):796–807.
5. Yang X, Gao L, Zhang S. Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns. *Brief Bioinform* 2017;**18**:761–73.
6. Daron J, Slotkin RK. EpiTEome: simultaneous detection of transposable element insertion sites and their DNA methylation levels. *Genome Biol* 2017;**18**(1):91.
7. Carja O, MacIsaac JL, Mah SM, *et al*. Worldwide patterns of human epigenetic variation. *Nat Ecol Evol* 2017;**1**(10):1577–83.
8. Taudt A, Colome-Tatche M, Johannes F. Genetic sources of population epigenomic variation. *Nat Rev Genet* 2016;**17**(6):319–32.
9. Wen Y, Wei Y, Zhang S, *et al*. Cell subpopulation deconvolution reveals breast cancer heterogeneity based on DNA methylation signature. *Brief Bioinform* 2017;**18**(3):426–40.
10. Brocks D, Assenov Y, Minner S, *et al*. Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Rep* 2014;**8**(3):798–806.
11. Teng H, Xue M, Liang J, *et al*. Inter- and intratumor DNA methylation heterogeneity associated with lymph node metastasis and prognosis of esophageal squamous cell carcinoma. *Theranostics* 2020;**10**(7):3035–48.
12. Li S, Zhang J, Huang S, *et al*. Genome-wide analysis reveals that exon methylation facilitates its selective usage in the human transcriptome. *Brief Bioinform* 2018;**19**(5):754–64.
13. Xu G, Lyu J, Li Q, *et al*. Evolutionary and functional genomics of DNA methylation in maize domestication and improvement. *Nat Commun* 2020;**11**(1):5539.
14. Artemov AV, Mugue NS, Rastorguev SM, *et al*. Genome-wide DNA methylation profiling reveals epigenetic adaptation of stickleback to marine and freshwater conditions. *Mol Biol Evol* 2017;**34**(9):2203–13.
15. Rodriguez Barreto D, Garcia de Leaniz C, Verspoor E, *et al*. DNA methylation changes in the sperm of captive-reared fish: a route to epigenetic introgression in wild populations. *Mol Biol Evol* 2019;**36**(10):2205–11.
16. Keller TE, Yi SV. DNA methylation and evolution of duplicate genes. *Proc Natl Acad Sci USA* 2014;**111**(16):5932–7.
17. Mazor T, Pankov A, Johnson BE, *et al*. DNA methylation and somatic mutations converge on the cell cycle and define similar evolutionary histories in brain tumors. *Cancer Cell* 2015;**28**(3):307–17.

18. Kronholm I, Bassett A, Baulcombe D, *et al*. Epigenetic and genetic contributions to adaptation in Chlamydomonas. *Mol Biol Evol* 2017;**34**(9):2285–306.
19. Wang F, Zhang S, Wen Y, *et al*. Revealing the architecture of genetic and epigenetic regulation: a maximum likelihood model. *Brief Bioinform* 2014;**15**(6):1028–43.
20. Guo W, Zhu P, Pellegrini M, *et al*. CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. *Bioinformatics* 2018;**34**(3):381–7.
21. Fang F, Hodges E, Molaro A, *et al*. Genomic landscape of human allele-specific DNA methylation. *Proc Natl Acad Sci U S A* 2012;**109**(19):7332–7.
22. Vincenz C, Lovett JL, Wu W, *et al*. Loss of imprinting in human placentas is widespread, coordinated, and predicts birth phenotypes. *Mol Biol Evol* 2020;**37**(2):429–41.
23. Tycko B. Allele-specific DNA methylation: beyond imprinting. *Hum Mol Genet* 2010;**19**(R2):R210–20.
24. Lister R, Pelizzola M, Dowen RH, *et al*. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;**462**(7271):315–22.
25. Barturen G, Rueda A, Oliver JL, *et al*. MethylExtract: high-quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000Res* 2013;**2**:217.
26. Liu Y, Siegmund KD, Laird PW, *et al*. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol* 2012;**13**(7):R61.
27. Gao S, Zou D, Mao L, *et al*. BS-SNPer: SNP calling in bisulfite-seq data. *Bioinformatics* 2015;**31**(24):4006–8.
28. Porter J, Sun MA, Xie H, *et al*. Investigating bisulfite short-read mapping failure with hairpin bisulfite sequencing data. *BMC Genomics* 2015;**16**(Suppl 11):S2.
29. Mattox AK, Wang Y, Springer S, *et al*. Bisulfite-converted duplexes for the strand-specific detection and quantification of rare mutations. *Proc Natl Acad Sci U S A* 2017;**114**(18):4733–8.
30. Olova N, Krueger F, Andrews S, *et al*. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol* 2018;**19**(1):33.
31. Giehr P, Walter J. Hairpin bisulfite sequencing: synchronous methylation analysis on complementary DNA strands of individual chromosomes. In: Tost J (ed). *DNA Methylation Protocols*. New York, NY: Springer New York, 2018, 573–86.
32. Young T, Mei F, Liu J, *et al*. Proteomics analysis of H-RAS-mediated oncogenic transformation in a genetically defined human ovarian cancer model. *Oncogene* 2005;**24**(40):6174–84.
33. Wang Y, Li G, Mao F, *et al*. Ras-induced epigenetic inactivation of the RRAD (Ras-related associated with diabetes) gene promotes glucose uptake in a human ovarian cancer model. *J Biol Chem* 2014;**289**(20):14225–38.
34. Cai W, Mao F, Teng H, *et al*. MBRidge: an accurate and cost-effective method for profiling DNA methylome at single-base resolution. *J Mol Cell Biol* 2015;**7**(4):299–313.
35. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* 2009;**10**(1):232.
36. Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**(16):2078–9.
37. Harrison JS, Cornett EM, Goldfarb D, *et al*. Hemi-methylated DNA regulates DNA methylation inheritance through allosteric activation of H3 ubiquitylation by UHRF1. *Elife* 2016;**5**:e17101.
38. Xu C, Corces VG. Nascent DNA methylome mapping reveals inheritance of hemimethylation at CTCF/cohesin sites. *Science* 2018;**359**(6380):1166–70.
39. Patiño-Parrado I, Gómez-Jiménez Á, López-Sánchez N, *et al*. Strand-specific CpG hemimethylation, a novel epigenetic modification functional for genomic imprinting. *Nucleic Acids Res* 2017;**45**(15):8822–34.
40. Zhao L, Sun MA, Li Z, *et al*. The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. *Genome Res* 2014;**24**(8):1296–307.
41. Sharif J, Endo TA, Nakayama M, *et al*. Activation of endogenous retroviruses in Dnmt1(−/−) ESCs involves disruption of SETDB1-mediated repression by NP95 binding to hemimethylated DNA. *Cell Stem Cell* 2016;**19**(1):81–94.
42. Holliday R, Pugh JE. DNA modification mechanisms and gene activity during development. *Science* 1975;**187**(4173):226–32.
43. Zhao L, Liu D, Xu J, *et al*. The framework for population epigenetic study. *Brief Bioinform* 2018;**19**(1):89–100.
44. Gunasekara CJ, Scott CA, Laritsky E, *et al*. A genomic atlas of systemic interindividual epigenetic variation in humans. *Genome Biol* 2019;**20**(1):105.
45. Ming X, Zhu B, Zhang Z. Simultaneously measuring the methylation of parent and daughter strands of replicated DNA at the single-molecule level by hammer-seq. *Nat Protoc* 2021;**16**(4):2131–57.
46. Giehr P, Kyriakopoulos C, Lepikhov K, *et al*. Two are better than one: HPoxBS - hairpin oxidative bisulfite sequencing. *Nucleic Acids Res* 2018;**46**(15):e88.
47. Cheung WA, Shao X, Morin A, *et al*. Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome. *Genome Biol* 2017;**18**(1):50.
48. Martin-Trujillo A, Patel N, Richter F, *et al*. Rare genetic variation at transcription factor binding sites modulates local DNA methylation profiles. *PLoS Genet* 2020;**16**(11):e1009189.
49. Gao F, Xia Y, Wang J, *et al*. Integrated analyses of DNA methylation and hydroxymethylation reveal tumor suppressive roles of ECM1, ATF5, and EOMES in human hepatocellular carcinoma. *Genome Biol* 2014;**15**(12):533.
50. Gao F, Xia Y. Hydroxymethylation- and methylation-sensitive tag sequencing: how will this technology change clinical applications of DNA methylation profiling? *Epigenomics* 2013;**5**(4):355–7.
51. Liu Y, Siejka-Zielińska P, Velikova G, *et al*. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat Biotechnol* 2019;**37**(4):424–9.
52. Vaisvila R, Ponnaluri VKC, Sun Z, *et al*. EM-seq: detection of DNA methylation at single base resolution from Picograms of DNA. bioRxiv 2020;2019.12.20.884692. doi: 10.1101/2019.12.20.884692 preprint: not peer reviewed.