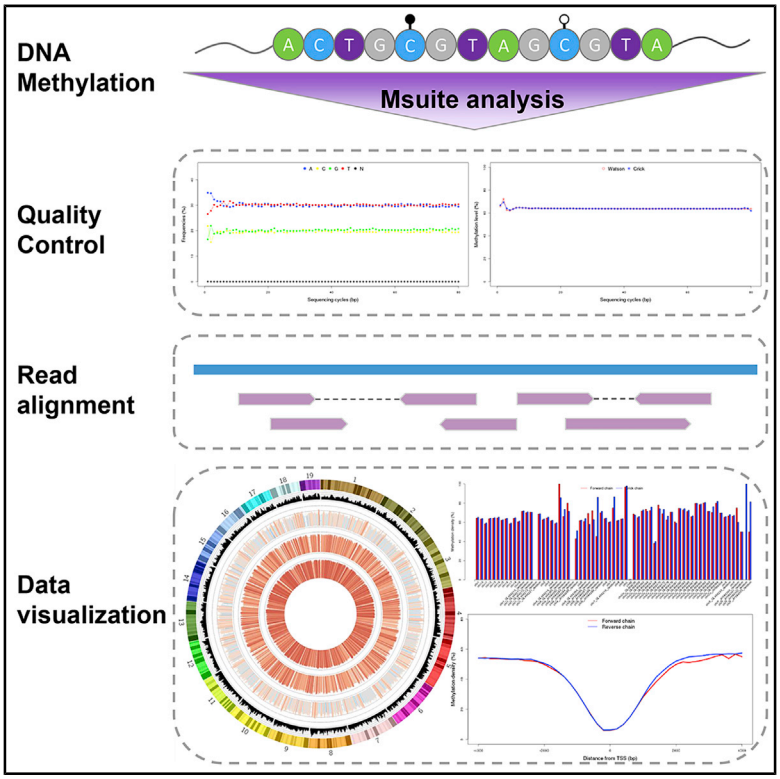
Descriptor

Patterns

Msuite: A High-Performance and Versatile DNA Methylation Data-Analysis Toolkit

Graphical Abstract



Highlights

- Msuite provides a unique 4-letter analysis mode for emerging bisulfite-free protocols
- Msuite outperforms current tools in terms of higher accuracy and lower resource usage
- Msuite has integrated quality control and fruitful data-visualization utilities
- Msuite provides an all-in-one solution for DNA methylation data analysis

Authors

Kun Sun, Lishi Li, Li Ma, Yu Zhao, Lin Deng, Huating Wang, Hao Sun

Correspondence

sunkun@szbl.ac.cn

In Brief

Emerging bisulfite-free assays for DNA methylome profiling have raised new requirements for data-analysis tools. Here, we present Msuite, an all-in-one package for DNA methylation data analysis with a unique 4-letter analysis mode for bisulfite-free protocols. Msuite has integrated quality control, read alignment, methylation call, and data visualization, and thus could serve as an optimal toolkit for DNA methylation studies.

Sun et al., 2020, Patterns *1*, 100127
November 13, 2020 © 2020 The Author(s).
https://doi.org/10.1016/j.patter.2020.100127

CellPress

Patterns

CellPress
OPEN ACCESS

Descriptor

Msuite: A High-Performance and Versatile DNA Methylation Data-Analysis Toolkit

Kun Sun,[1,7,*] Lishi Li,[1,2] Li Ma,[1] Yu Zhao,[3] Lin Deng,[1] Huating Wang,[4,5] and Hao Sun[4,6]
[1]Shenzhen Bay Laboratory, Shenzhen 518132, China
[2]Peking University Shenzhen Graduate School, Shenzhen 518055, China
[3]School of Medicine, Sun Yat-sen University, Guangzhou 510080, China
[4]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong SAR 999077, China
[5]Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong, Hong Kong SAR 999077, China
[6]Department of Chemical Pathology, The Chinese University of Hong Kong, Hong Kong SAR 999077, China
[7]Lead Contact
*Correspondence: sunkun@szbl.ac.cn
https://doi.org/10.1016/j.patter.2020.100127

THE BIGGER PICTURE DNA methylation is an essential epigenetic modification responsible for many biological regulation pathways. Despite the fact that various high-throughput methods have been developed for base-resolution DNA methylome profiling, DNA methylation data analysis remains a complex and challenging task. Here, we present Msuite, which has integrated quality control, read alignment, methylation call, and fruitful data-visualization functionalities, aiming to offer an all-in-one package for most of the current DNA methylation profiling assays. Msuite also provides dedicated support for emerging bisulfite-free protocols and outperforms the current tools in terms of higher accuracy and lower computational resource requirement. Hence, Msuite could serve as the optimal toolkit for DNA methylation data analysis as well as facilitating the popularization of emerging bisulfite-free protocols.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

DNA methylation is a pervasive and important epigenetic regulator in mammalian genome. For DNA methylome profiling, emerging bisulfite-free methods have demonstrated desirable superiority over the conventional bisulfite-treatment-based approaches, although current analysis software could not make full use of their advantages. In this work, we present Msuite, an easy-to-use, all-in-one data-analysis toolkit. Msuite implements a unique 4-letter analysis mode specifically optimized for emerging protocols; it also integrates quality controls, methylation call, and data visualizations. Msuite demonstrates substantial performance improvements over current state-of-the-art tools as well as fruitful functionalities, thus holding the potential to serve as an optimal toolkit to facilitate DNA methylome studies. Source codes and testing datasets for Msuite are freely available at https://github.com/hellosunking/Msuite/.

INTRODUCTION

DNA methylation is an important epigenetic regulator that plays crucial roles in a broad range of biological processes. In mammalian genomes, DNA methylation mostly involves the addition of a methyl group to cytosine nucleotides and is linked to gene repression.[1,2] The cytosine methylation pattern has been found to be tissue specific and possesses high biological and translational values, for example in transcription regulation[1–5] and cancer liquid biopsy studies.[6–12] As a consequence, DNA methylation is actively and widely investigated in various research fields.

Multiple biochemical assays have been developed for high-resolution DNA methylome profiling in the past years.[13] To differentiate methylated cytosines from unmethylated ones, conventional approaches (e.g., whole-genome bisulfite sequencing [WGBS][13,14]) apply a bisulfite treatment procedure to DNA molecules, which converts all unmethylated cytosines into uracil while leaving methylated cytosines unchanged. During the subsequent PCR amplifications, uracils are recognized as

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

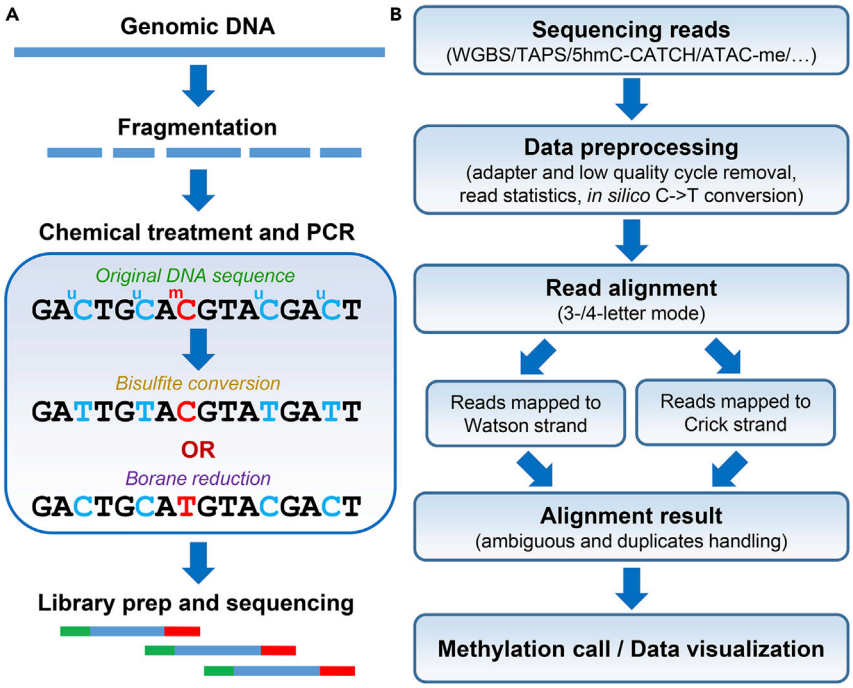


Figure 1. Schematic Workflow DNA Methylome Profiling

(A) Experimental procedures of current assays. (B) Msuite data analysis pipeline. The key step in (A) is to use chemical treatment to differentiate methylated and unmethylated cytosines (colored in red and blue, respectively), where various protocols are available that lead to different modifications to the genomic DNA.

thymines, resulting in a cytosine-to-thymine transition to the original DNA. Emerging bisulfite-free techniques (e.g., TET-assisted pyridine borane sequencing [TAPS][15–17]), however, introduce opposite modifications to DNA molecules whereby only the methylated cytosines are converted into thymines while the unmethylated cytosines are left untouched (Figure 1A). In mammalian genomes, methylated cytosines mostly appear in CpG dinucleotides, which account for a very limited proportion (e.g., ~5% in human) of all cytosines. As a result, DNA libraries generated by bisulfite-free approaches show a significantly higher nucleotide complexity (Figure 1A) because only a small proportion of the cytosines are converted after chemical treatment, which characteristically benefits in lower GC-bias and even more coverage of the genome in the sequencing data.[15,18,19] However, most of the current mainstream data-analysis tools only support 3-letter read alignment (i.e., they convert all the cytosines in both the reference genome and sequencing reads into thymines);[13,20–22] the fundamental change introduced by the emerging assays thus renders the current tools outdated due to their disability to make full use of such an advantage. Other analysis tools utilize a wild-card mapping strategy; however, they usually suffer from low speed and unsatisfactory mapping efficiency,[22,23] or are also optimized for 3-letter alignment.[24] In addition, most of the current software focuses on sequencing read alignment and requires the users to perform quality control, downstream analysis (e.g., methylation call), and data visualization. Hence, a user-friendly, multi-functional toolkit with better support for current bisulfite-free assays is of urgent demand.

In this study, we present Msuite, a package that supports data analysis of all the current mainstream DNA methylome assays. As a versatile toolkit, Msuite provides various utilitarian functions, including quality control, a novel 4-letter sequencing read alignment algorithm specifically optimized for bisulfite-free protocols, methylation call, as well as plentiful data-visualization utilities. Msuite also outperforms current state-of-the-art software in terms of higher accuracy, faster speed, and lower computational resource usage. Msuite thus provides an easy-to-use, all-in-one solution for DNA methylation data analysis. Msuite is freely available at https://github.com/hellosunking/Msuite/.

RESULTS

Overall Design of Msuite

Figure 1 shows the schematic workflow of current DNA methylation profiling assays and the Msuite data-analysis toolkit. Typically, the genomic DNA of interest is first fragmented into small pieces of several hundred base pairs long (sometimes the DNA molecules are inherently fragmented such as plasma cell-free DNA[25]), then the short DNA molecules are treated by various chemistries and several rounds of PCR cycles to differentiate methylated and unmethylated cytosines. The biochemically treated DNA molecules are then subjected to library preparation and sequencing. Raw sequencing data directly serve as the input for Msuite. Msuite adapts our previous sequencing data preprocessing tool, Ktrim,[26] to perform extra-fast, accurate adapter-/quality-trimming, and *in silico* cytosine-to-thymine conversion of the sequencing reads. Notably, Msuite supports sequencing data generated from various library preparation kits and is able to directly handle conventional WGBS and emerging sequencing protocols such as TAPS,[15] 5hmC-CATCH,[16] and ACE-seq,[17] as well as ATAC-me,[27] methyl-ATAC-seq,[28] and EpiMethylTag[29] (integrative methods that measure DNA methylation at regulatory elements, such as accessible chromatin or transcription factor binding domains). Msuite then aligns the pre-processed reads to the reference genome in either 3- or 4-letter mode based on the assay and users' settings (see Experimental Procedures). Notably, the 4-letter mode is specifically designed for TAPS-like protocols that are optimized for detecting CpG methylations, while 3-letter mode is generic and works for most kinds of current DNA methylation assays as well as scenarios with gross non-CpG methylation. The sequencing reads are aligned to Watson and Crick strands of the reference genome separately, since cytosine-to-thymine conversion has disrupted their reverse-complementary relationship. After this initial alignment, Msuite recognizes and handles the ambiguously aligned reads as well as PCR duplicates to generate the final alignment result, based on which Msuite further performs a methylation call (i.e., it reports methylation status of all cytosines in both CpG and non-CpG contexts) and data visualizations.

Table 1. Comparison of Major Features between Msuite and Current Software

	Msuite	Bismark	BWA-meth	Methy-Pipe
Underlying aligner	bowtie2	bowtie2/Hisat2	BWA	SOAP2
Output format	SAM/BAM	BAM	SAM	SOAP-like
Align mode	3-/4-letter	3-letter only	3-letter only	3-letter only
Indel support	yes	Yes	yes	no
Quality control	yes	No	no	yes
Methylation call	yes	manually	no	yes
Data visualization	yes	No	no	yes
Multiple-file support	yes	Yes	yes	yes
Sequencing mode	paired-/single-end	paired-/single-end	paired-/single-end	paired-/single-end
Parallelization	yes	Yes	yes	yes

Feature Comparison of Msuite and Current Software

To demonstrate the usability of Msuite, we compared the most valuable features for DNA methylation data analysis between Msuite and current state-of-the-art software: Bismark,[30] BWA-meth,[20] and our previously developed Methy-Pipe[21] (Table 1). Msuite employs bowtie2[31] as the bottom aligner, which is the same as Bismark (which also supports Hisat2[32]) while different from BWA-meth and Methy-Pipe (which use BWA[33] and SOAP2,[34] respectively). As a result, Msuite, Bismark, and BWA-meth output the alignment results in standardized SAM/BAM[35] format while Methy-Pipe records the data in an alternative format similar to SOAP2. Moreover, Msuite tolerates insertion and deletions in the sequencing data, a feature also supported by Bismark and BWA-meth while not in Methy-Pipe. Msuite supports both 4- and 3-letter alignment while the others only provide 3-letter alignment. In addition, both Msuite and Methy-Pipe automatically perform DNA methylation calls in the data. In contrast, Bismark provides a script but requires the users to run it manually; BWA-meth only performs read alignment without any downstream analysis support. Lastly, Msuite has integrated built-in quality controls, including adaptor-/quality-trimming and removal of PCR duplicates, as well as various data-visualization functions. Msuite thus provides an easy-to-use, all-in-one solution for DNA methylation data analysis and can be readily integrated with other software for comprehensive data mining.

Benchmark Performance Evaluation of Msuite

Sequencing read alignment is the most challenging and imperative step in DNA methylation data analysis. We benchmarked and compared the performance of read alignment algorithms between Msuite and current software. For a fair comparison, Methy-Pipe is excluded from this analysis because both Methy-Pipe and its underline aligner have not been updated for more than 5 years. The latest versions of Bismark (v0.22.3) and BWA-meth (v0.2.2) as well as their underline aligners (bowtie2 v2.3.5.1 and BWA v0.7.17) were downloaded from the literature[20,30,31,33] and installed on a computing server equipped with Intel Xeon CPU, 192 Gb memory, and standard CentOS 64-bit Linux system. A total of 900 *in silico* experiments following the BS-seq or TAPS protocol were performed (see Experimental Procedures). The averaged alignment statistics, running time, and peak memory usage on 1 million *in silico* simulated paired-end reads following the TAPS protocol are shown in Table 2, and the results for paired-end reads following BS-seq protocol as well as single-end data are included in Table S1 (notably, BWA-meth fails in processing single-end 36-bp reads). We measured mapping efficiency as the proportion of reads that could be mapped by the aligner, and accuracy as the proportion of correct alignments (i.e., aligned loci are exactly the same as that in simulation) in the mapped reads. In brief, both mapping efficiency and accuracy are high and comparable among the software benchmarked, while Msuite in 4-letter mode is slightly better. Intriguingly, even though Msuite performs an additional adapter- and quality-trimming step before alignment (whose time is counted in Table 2), it runs faster and uses much less memory than Bismark and BWA-meth, particularly in 4-letter mode.

To further explore the advantage of Msuite's unique 4-letter alignment mode, we generated *in silico* data originating from CT- or GA-rich regions in human genome (see Experimental Procedures). For reads generated from regions with CT/GA proportion higher than 70%, 4-letter alignment mode shows apparently better mapping efficiency and accuracy; for regions where CT/GA proportion is higher than 80%, the advantage of 4-letter alignment mode becomes highly remarkable in terms of superior mapping efficiency, higher accuracy, and less alignment time (Table 2). In fact, CT/GA-rich regions are ~9.2 Mbp long in total, which accounts for ~0.31% of the human genome; however, ~4.5% of them locate in promoter regions, a proportion much higher than the genomic background (~2.9%); i.e., these regions are enriched in regulators and thus possess biological relevance. Hence, prominently improved performance in aligning reads in CT/GA-rich regions justifies the merit of 4-letter mode in DNA methylation data analysis.

We further profiled the accuracy of the methylation call function of Msuite on the benchmark dataset. The overall methylation densities deduced by Msuite are in close approximation to the preset methylation densities during simulation; the differences are on a similar level to the preset sequencing error rate and show no relationship with the preset methylation densities in the data (Figure S1).

Analysis Results in Real Datasets

To illustrate the usage of Msuite, we applied it to a real dataset generated from both WGBS and TAPS protocols on murine

Table 2. Benchmark Evaluation of Msuite and Current Software

	Msuite (4-Letter Mode)	Msuite (3-Letter Mode)	Bismark	BWA-meth[a]
1 Million paired-end 100-bp reads				
Mapping efficiency (%)	96.19	95.75	95.79	95.58
Accuracy (%)	99.94	99.95	99.93	99.96
Running time (s)	119.45	195.80	236.37	155.14
Peak memory (Gb)	3.74	3.74	40.21	12.52
1 Million paired-end 36-bp reads				
Mapping efficiency (%)	92.46	91.62	91.71	91.46
Accuracy (%)	99.79	99.63	99.65	99.67
Running time (s)	126.60	173.84	164.80	283.55
Peak memory (Gb)	3.77	3.78	40.16	22.52
1 Million paired-end 36-bp reads originating from CT/GA $\geq$ 70% regions				
Mapping efficiency (%)	79.67	67.14	68.39	76.31
Accuracy (%)	99.78	99.03	98.94	92.06
Running time (s)	120.30	307.00	294.80	360.80
1 Million paired-end 36-bp reads originating from CT/GA $\geq$ 80% regions				
Mapping efficiency (%)	73.53	53.14	54.60	65.85
Accuracy (%)	99.81	98.38	98.19	86.91
Running time (s)	112.60	375.00	366.00	469.70

Eight threads were used for benchmark testing, and the data were simulated following the TAPS protocol.
[a]For BWA-meth, alignments with a score of 0 were discarded due to abnormally high error rate.

embryonic stem cells (ESCs) by Liu et al.[15] The WGBS data were analyzed using the 3-letter mode, while the TAPS data were analyzed using both 3- and 4-letter modes against reference mouse genome (NCBI assembly GRCm38). The final analysis report on the TAPS data using 4-letter mode is shown in Figure 2, and the reports for the other two analyses are provided in Figure S2. Various figures are provided to help the users to assess the quality of their data, including base composition, M-bias (average methylation level for each cycle) plots,[36] and DNA methylation signals around promoters. In this dataset, the base composition plots show high cytosine proportion in read 1 as well as high guanine proportion in read 2, which directly reflects the improved sequence complexity of the TAPS protocol. In addition, the methylation level shows a decreased signal around promoters, which is consistent with the knowledge that most promoters are hypomethylated for active transcription, thus providing a preliminary assessment for the users to inspect the validity of their data. Notably, the 4-letter mode took ~290 min to complete read alignment of the TAPS data using 32 threads, while the 3-letter mode took ~440 min; therefore, the 4-letter mode was ~50% faster. In addition, despite imperfect cytosine-to-thymine conversion rate in this TAPS experiment[15] and gross non-CpG methylation in ESCs,[14,15] which means a high proportion of cytosines in CpH context are converted into thymines and thus affect the performance of 4-letter mode, we found that the 4-letter mode only shows a 2.17% deficit in final mapped reads compared with the 3-letter mode. Moreover, the final mapped reads reported by Msuite is comparable with the original report by Liu et al. using a different alignment strategy; however, we find that Msuite provides even more read coverage on CpG islands, especially for the highly methylated ones (Figure S3; see Discussion).[15] On the other hand, the deduced DNA methylation densities on CpG loci between 4- and 3-letter modes are largely the same (Pearson's $R > 0.99$, $p < 2.2 \times 10^{-16}$), with only 3.78% showing methylation differences higher than 10% (such CpG sites suffering from much lower coverage and enrichment in repeat regions, Figure S4). Moreover, the methylation densities deduced from TAPS data are also in good agreement with the WGBS data (Figure S4), which is consistent with the original report by Liu et al.[15]

Data Visualization

Besides the analysis report that allows the users to conveniently inspect the key statistics as well as quality assessments of their data (Figure 2), Msuite has also packaged various data-visualization utilities. For instance, during methylation call, Msuite records the DNA methylation densities for each CpG site in a BEDGRAPH (http://genome.ucsc.edu/goldenPath/help/bedgraph.html) format file, which can be readily visualized in the UCSC genome browser[37] or Integrative Genomics Viewer (IGV).[38] As illustrated in Figure 3A, a low methylation level and open chromatin signal are found on *Pou5f1* gene (also known as *Oct4*, a transcription factor expressed in ESCs but not in somatic tissues such as liver[39]) in murine ESCs[40] in contrast to liver tissue. Msuite also provides utilities to summarize the DNA methylation densities for easy incorporation with other data-visualization software, such as Circos.[41] An example is shown in Figure 3B, where murine placental tissue presents conspicuous global hypomethylation compared with ESCs and liver tissue.

In addition, Msuite contains a dedicated tool, Mviewer, adapted from our previous BSviewer[42] software, to provide fast, nucleotide-level, and genotype-preserved DNA methylation sequencing data visualization. An example of Mviewer's output on a WGBS

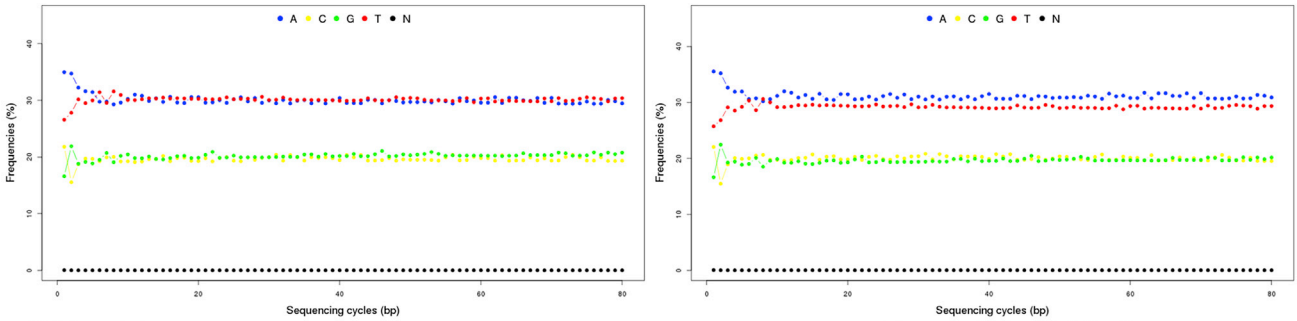
Alignment statistics

Total input reads	**455,548,210**
After preprocessing	**454,949,691 (99.87 %)**
Total aligned reads	**335,553,983 (73.76 %)**
Forward chain	167,870,243 (36.90 %)
Reverse chain	167,683,740 (36.86 %)
Unique reads	**324,098,228 (96.59 %)**
PCR duplicates	**11,455,755 (3.41 %)**

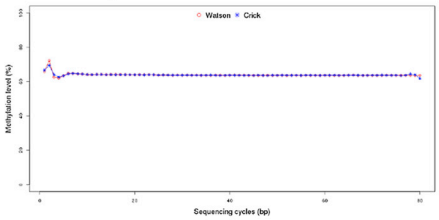
Methylation statistics

Overall methylation density	**62.36 %**
Forward chain	62.35 %
Reverse chain	62.36 %
Reads mapped to Lambda genome	**489,103**
C->T conversion rate for lambda genome	**96.98 %**

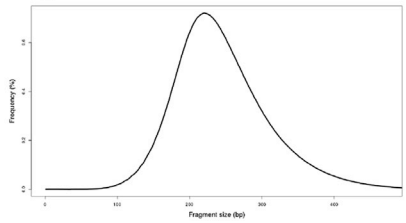
Base composition in the sequenced reads



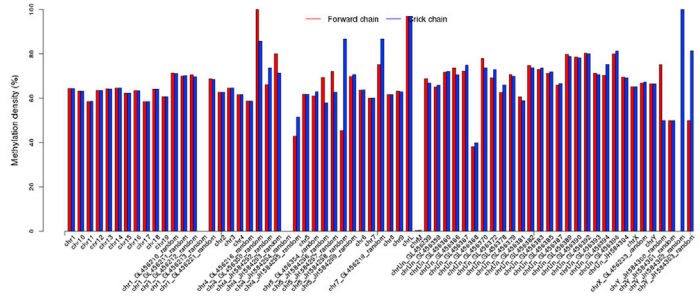
M-bias plot



Fragment size distribution



Methylation level per chromosome



Methylation level around TSS

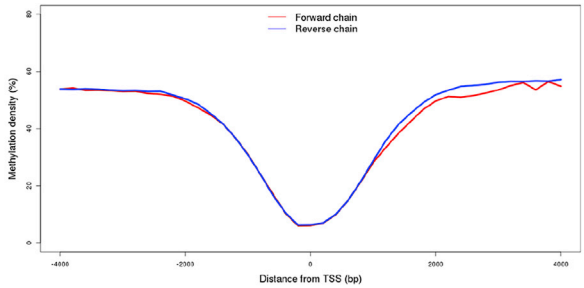


Figure 2. Example of Msuite Analysis Summary on a Real Dataset Generated Using TAPS Protocol on Murine ESCs

Msuite outputs key statistics and various figures to help the users to assess the quality as well as obtain a rough impression of their data.

dataset generated from human placental tissue[42–44] is shown in Figure 3C, which highlights the genotype information (i.e., an A/G heterozygous locus at chr7:130493085, which is recorded as rs2301335 in dbSNP[45]) as well as the allele-specific methylation pattern of the *MEST* imprinting gene in human placenta.

DISCUSSION

The development of a WGBS protocol as well as the first base-resolution human methylome was accomplished over a decade ago.[14] However, due to the non-unified, context-dependent cytosine-to-thymine conversions introduced in DNA methylation profiling assays, the bioinformatics analysis is still challenging and complex.[13,22] Tremendous changes have been made in the emerging approaches such as the TAPS protocol, which have demonstrated desirable benefits compared with conventional bisulfite treatment protocols, including higher sequence complexity and lower DNA degradation.[15,46] It is therefore of value to design and implement dedicated analysis tools to meet the requirements raised by these novel protocols as well as facilitate their applications and advances. To this end we present Msuite, a modern data-analysis toolkit that supports almost

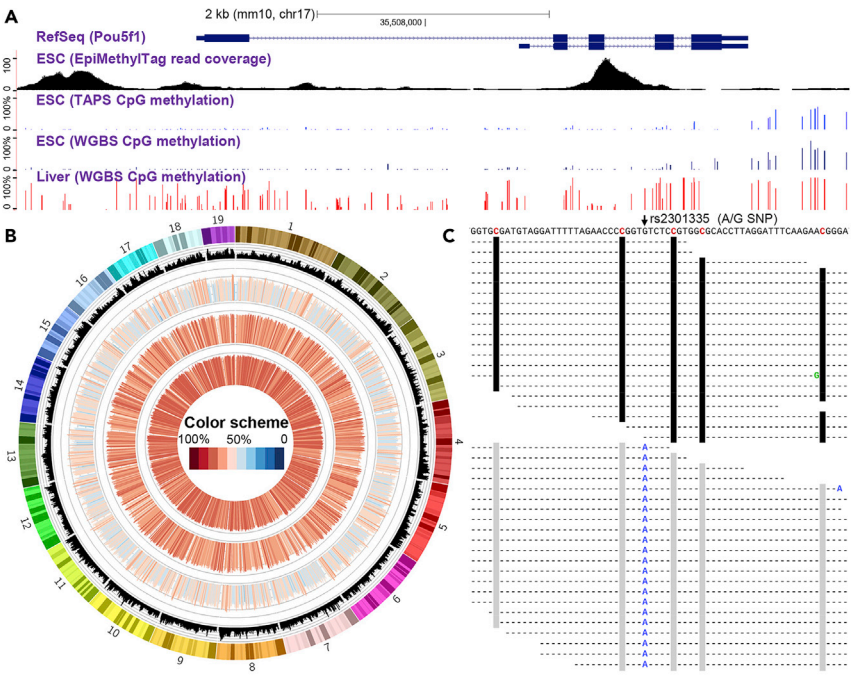


Figure 3. Illustration of Msuite's Data-Visualization Functions

(A) UCSC genome browser snapshot of DNA methylation densities around *Pouf51* (also known as *Oct4*) gene in murine ESCs and liver tissue.

(B) Circos plot of DNA methylation densities in murine placental tissue, ESCs, and liver tissue. The mouse genome is divided into 1-Mbp bins. DNA methylation level for each color is indicated in "Color scheme." From outmost to innermost circles: CpG densities, placental tissue, ESCs, and liver tissue.

(C) Mviewer output on a WGBS dataset from human placental tissue. An A/G SNP site (rs2301335 in dbSNP) is present in this sample. Black and gray bars represent methylated and unmethylated cytosines, respectively; hyphens and colored letters represent sequenced nucleotides that are the same as and different from the reference genome, respectively.

all of the current mainstream DNA methylation profiling assays. In fact, we have analyzed sequencing data generated from conventional WGBS, emerging TAPS, and EpiMethylTag assays in this work (Figure 3). Moreover, Msuite outperforms current state-of-the-art software in terms of better mapping efficiency, higher accuracy, faster speed, and lower computing resource requirements. Even though Msuite and Bismark both utilize bowtie2 as the underline aligner, Msuite shows higher speed and lower memory usage. This is mostly because Msuite only calls one bowtie2 instance and runs it in multi-thread mode while Bismark calls multiple bowtie2 instances and runs them in single-thread mode, i.e., Bismark loads multiple copies of the genome indices and therefore requires more time and memory. The unique 4-letter analysis mode of Msuite is designed for emerging bisulfite-free assays, such as TAPS and 5hmC-CATCH, which indeed demonstrates improved performance on an *in silico* simulated dataset, especially in the CT/GA-rich regions. The 3-letter mode, on the other hand, is also essential as it is more generic and could handle datasets generated using conventional bisulfite treatment approaches or from species/tissues with gross non-CpG methylations (e.g., plants, the brain).

On the real dataset generated using the TAPS protocol, despite imperfect chemical treatment, the 4-letter mode still shows high-quality results and completes in much less time. In addition, on this real dataset, Msuite also shows certain advantages over the original method used by Liu et al., which directly aligns the reads to the genome without any modifications (i.e., they treat the data as normal DNA sequencing during alignment).[15] As shown in Figure S3, Msuite shows much better coverage for reads originated from hypermethylated CpG islands (e.g., suppressed regulator elements in the specific cell type); such reads usually contain various methylated cytosines that are converted into thymines in the TAPS assay, and therefore contains too many "mismatches" compared with the reference genome to be aligned efficiently; by contrast, such reads do not affect Msuite because both cytosines and converted thymines in CpG sites are accepted as "matches" after the cytosine-to-thymine conversion of the reference genome and sequencing reads. Together, these results demonstrate the advantage and rationale of Msuite's 4-letter analysis mode for better support to the emerging assays. Interestingly, 4- and 3-letter modes generate rather consistent results, although the deduced methylation levels show a large difference for a small proportion of CpG sites, which are enriched in repeat regions and indeed show much lower coverage (Figure S4), suggesting that they are located in the genomic regions that are difficult to align. Although the 4-letter analysis mode has demonstrated higher mapping accuracy on the benchmark datasets, we do not have strong evidence that the 4-letter mode is more accurate on those inconsistent CpG sites; therefore, it is meaningful to further explore and/or validate the accuracy of 4- and 3-letter analysis modes, as well as the limitation of current sequencing-based protocols, using additional methods (e.g. microarrays) on such loci.

In addition, Msuite integrates quality control, sequencing read alignment, and downstream analyses (e.g., methylation call) into one pipeline, thus providing an easy-to-use, all-in-one solution for DNA methylation data analysis. Msuite provides multiple data-visualization functions, which could further help users to inspect and interpret their data. For instance, its accompanying tool, Mviewer, provides favorable characteristics that can be specifically meaningful in scenarios with allele-specific DNA methylation, such as imprinting gene (Figure 3) and tissue-specific signatures.[42,47] Msuite also provides all the features required for a modern data analyzer, including multi-file support, outputs in standardized format, and parallelization (Table 1). Hence, Msuite holds the full potential to serve as an optimal data-analysis toolkit to facilitate DNA methylation studies.

Conclusion

In conclusion, we have designed and implemented Msuite, a versatile and high-performance DNA-analysis toolkit, with

dedicated support for emerging bisulfite-free assays and enhanced performance over the state-of-the-art tools, providing an easy-to-use and all-in-one solution for analysis of DNA methylation data.

EXPERIMENTAL PROCEDURES

Resource Availability

Lead Contact

Kun Sun, Ph.D., sunkun@szbl.ac.cn.

Materials Availability

This study did not generate any new unique reagents or materials.

Data and Code Availability

Source codes of Msuite and scripts to reproduce the results described in this paper are freely available at https://github.com/hellosunking/Msuite/, distributed under the GPL v3 license. Accession numbers for third party data used in this study: GEO: GSE112520, GSE129673 (murine ESCs), GSM2191922 (murine liver tissue), GSM1545829 (murine placental tissue), and GSM1186665 (human placental tissue).

Aim, Design, and Setting of the Study

The aim of this study is to develop an all-in-one DNA methylation data-analysis toolkit that is easy to use, powerful, and supports all of the current assays (especially the emerging bisulfite-free ones). The schematic workflow of Msuite is shown in Figure 1B; detailed information for alignment algorithm (the core component) and performance benchmarking are explained in the following sections.

Sequencing Read Alignment Strategy

Msuite provides two alignment modes: 3- and 4-letter. The 3-letter mode of Msuite is similar to conventional methods: it first converts all the genomic cytosines to thymines (hence, the converted genome only contains three letters: adenine, guanine, and thymine), then builds two indices for Watson and Crick chains separately. Msuite then converts all the cytosines to thymines in the sequencing reads and aligns them against the pre-built 3-letter genome indices. The 4-letter mode of Msuite, however, only converts the cytosines in the CpG context to thymines, while leaving the rest of the cytosines untouched, and builds indices for Watson and Crick chains separately. In mammalian genomes, only a very minor proportion of cytosines are within a CpG context, therefore the converted genome still contains a high proportion of cytosines (i.e., still a 4-letter genome). During sequencing read alignment, only the cytosines followed by guanines are converted to thymines and aligned against the pre-built 4-letter genome indices. After the initial alignment, Msuite screens for multiple-mapping reads and only those with unique best hits are kept: reads that could be mapped to multiple locations in the same strand with equal best scores are discarded; for a read that has 1 unique best hit on Watson strand and 1 unique best hit on Crick strand, if the two hits have the same score, the read will be discarded; otherwise the hit with a higher score will be reported (along with a reduced mapping score). In addition, Msuite also looks for aligned reads that have identical start and end positions and strand information as PCR duplicates, and only keeps the one with the best sequencing read quality.

Benchmark Data Generation

Considering that both Bismark and BWA-meth perform 3-letter alignment, whereby all the cytosines in the sequencing read and reference genome are converted to thymines irrelevant to their sequence context, we thus generated *in silico* simulation datasets following the BS-seq and TAPS protocols separately for performance evaluations as well as investigating the advantage of Msuite's unique 4-letter mode. The simulated data following the BS-seq protocol was analyzed using 3-letter mode only, and the simulated data following the TAPS protocol was analyzed using both 3- and 4-letter modes. Four *in silico* datasets containing paired-/single-end 36-/100-bp reads, respectively, were generated using SHERMAN script (https://www.bioinformatics.babraham.ac.uk/projects/sherman/) against the human reference genome (NCBI assembly GRCh38). Each dataset was composed of 11 levels of 1 million reads with cytosine-to-thymine conversion rates on the CpG loci ranging from 0% to 100% in 10% increments. In the meantime, cytosine-to-thymine conversion on non-CpG loci was set to 99.5% and 0.5% for BS-seq and TAPS protocols, respectively, as the vast majority of them are unmethylated in mammalian genome. A sequencing error rate of 0.1% was also incorporated into the simulated data. Considering that other benchmarked software does not contain built-in support for adaptor trimming, sequencing adaptors are not incorporated during simulation. The experiments for each of the 11-level cytosine-to-thymine conversation rates were repeated ten times and the averaged mapping efficiency and accuracy, as well as running time and peak memory usage for the benchmarked software were reported.

For the CT/GA-rich regions, we first divided the human reference genome into 500-bp bins and searched for the bins with CT or GA proportion higher than 70% (or 80%). During *in silico* reads simulation, we set the cytosine-to-thymine conversion rate on CpG loci to 50% and kept other settings identical to the previous simulation. We then adapted the SHERMAN script only to generate reads overlapping the CT/GA-rich regions. Ten repeat experiments, each with 1 million paired-end 36-bp reads simulated, were performed for CT/GA proportion larger than 70% and 80% regions, respectively. To annotate the CT/GA-rich regions, we extracted the transcription start sites (TSSs) in RefSeq genes[48] and defined (TSS − 2K, TSS + 1K) as promoter regions. The scripts for *in silico* benchmark data generation, performance evaluation, mined CT/GA-rich regions, and testing environment information were publicly available at https://github.com/hellosunking/Msuite/.

Implementation and System Requirement

Msuite is implemented in C++/Perl and runs on GNU/Linux systems. To use Msuite, a working C++ compiler (e.g., g++) and Perl interpreter (usually distributed along with the Linux system) are required. In addition, Msuite employs bowtie2[31] for read alignment, samtools[35] for converting SAM format files into BAM format, and R for data visualizations. When called, Msuite will look for the dependencies in the system automatically.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.patter.2020.100127.

ACKNOWLEDGMENTS

We thank the Shenzhen Bay Laboratory and National Supercomputing Center in Shenzhen for providing computing support. This work was supported by Shenzhen Bay Laboratory and Guangdong Basic and Applied Basic Research Foundation (2019A1515110173).

AUTHOR CONTRIBUTIONS

K.S. and H.S. conceived of the study; K.S. designed and implemented the software, and supervised the overall study; K.S., L.L., L.M., Y.Z., L.D., H.W., and H.S. performed research; K.S. and L.L. wrote the paper. All the authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 1, 2020
Revised: September 8, 2020
Accepted: September 16, 2020
Published: October 15, 2020

REFERENCES

1. Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nat. Rev. Genet. *14*, 204–220.
2. Moore, L.D., Le, T., and Fan, G. (2013). DNA methylation and its basic function. Neuropsychopharmacology *38*, 23–38.
3. Wang, L., Zhao, Y., Bao, X., Zhu, X., Kwok, Y.K., Sun, K., Chen, X., Huang, Y., Jauch, R., Esteban, M.A., et al. (2015). LncRNA Dum interacts with

Dnmts to regulate Dppa2 expression during myogenic differentiation and muscle regeneration. Cell Res. *25*, 335–350.

4. Li, L., Zhang, Y., Fan, Y., Sun, K., Su, X., Du, Z., Tsao, S.W., Loh, T.K., Sun, H., Chan, A.T., et al. (2015). Characterization of the nasopharyngeal carcinoma methylome identifies aberrant disruption of key signaling pathways and methylated tumor suppressor genes. Epigenomics *7*, 155–173.
5. Zhao, Y., Yang, Y., Trovik, J., Sun, K., Zhou, L., Jiang, P., Lau, T.S., Hoivik, E.A., Salvesen, H.B., Sun, H., et al. (2014). A novel wnt regulatory axis in endometrioid endometrial cancer. Cancer Res. *74*, 5103–5117.
6. Chan, K.C.A., Jiang, P., Chan, C.W., Sun, K., Wong, J., Hui, E.P., Chan, S.L., Chan, W.C., Hui, D.S., Ng, S.S., et al. (2013). Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. Proc. Natl. Acad. Sci. U S A *110*, 18761–18768.
7. Sun, K., Jiang, P., Chan, K.C.A., Wong, J., Cheng, Y.K., Liang, R.H., Chan, W.K., Ma, E.S., Chan, S.L., Cheng, S.H., et al. (2015). Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc. Natl. Acad. Sci. U S A *112*, E5503–E5512.
8. Xu, R.H., Wei, W., Krawczyk, M., Wang, W., Luo, H., Flagg, K., Yi, S., Shi, W., Quan, Q., Li, K., et al. (2017). Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. Nat. Mater. *16*, 1155–1161.
9. Lam, W.K.J., Gai, W., Sun, K., Wong, R.S.M., Chan, R.W.Y., Jiang, P., Chan, N.P.H., Hui, W.W.I., Chan, A.W.H., Szeto, C.C., et al. (2017). DNA of erythroid origin is present in human plasma and informs the types of anemia. Clin. Chem. *63*, 1614–1623.
10. Gai, W., Ji, L., Lam, W.K.J., Sun, K., Jiang, P., Chan, A.W.H., Wong, J., Lai, P.B.S., Ng, S.S.M., Ma, B.B.Y., et al. (2018). Liver- and colon-specific DNA methylation markers in plasma for investigation of colorectal cancers with or without liver metastases. Clin. Chem. *64*, 1239–1249.
11. Gai, W., and Sun, K. (2019). Epigenetic biomarkers in cell-free DNA and applications in liquid biopsy. Genes (Basel) *10*, 32.
12. Sun, K., Jiang, P., Cheng, S.H., Cheng, T.H.T., Wong, J., Wong, V.W.S., Ng, S.S.M., Ma, B.B.Y., Leung, T.Y., Chan, S.L., et al. (2019). Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. Genome Res. *29*, 418–427.
13. Raiber, E.-A., Hardisty, R., van Delft, P., and Balasubramanian, S. (2017). Mapping and elucidating the function of modified bases in DNA. Nat. Rev. Chem. *1*, 0069.
14. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.
15. Liu, Y., Siejka-Zielinska, P., Velikova, G., Bi, Y., Yuan, F., Tomkova, M., Bai, C., Chen, L., Schuster-Bockler, B., and Song, C.X. (2019). Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. Nat. Biotechnol. *37*, 424–429.
16. Zeng, H., He, B., Xia, B., Bai, D., Lu, X., Cai, J., Chen, L., Zhou, A., Zhu, C., Meng, H., et al. (2018). Bisulfite-free, nanoscale analysis of 5-hydroxymethylcytosine at single base resolution. J. Am. Chem. Soc. *140*, 13190–13194.
17. Schutsky, E.K., DeNizio, J.E., Hu, P., Liu, M.Y., Nabel, C.S., Fabyanic, E.B., Hwang, Y., Bushman, F.D., Wu, H., and Kohli, R.M. (2018). Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. Nat. Biotechnol. https://doi.org/10.1038/nbt.4204.
18. Olova, N., Krueger, F., Andrews, S., Oxley, D., Berrens, R.V., Branco, M.R., and Reik, W. (2018). Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol. *19*, 33.
19. Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. (2013). Characterizing and measuring bias in sequence data. Genome Biol. *14*, R51.
20. Pedersen, B.S., Eyring, K., De, S., Yang, I.V., and Schwartz, D.A. (2014). Fast and accurate alignment of long bisulfite-seq reads. arXiv, 1401.1129.
21. Jiang, P., Sun, K., Lun, F.M.F., Guo, A.M., Wang, H., Chan, K.C.A., Chiu, R.W.K., Lo, Y.M.D., and Sun, H. (2014). Methy-pipe: an integrated bioinformatics pipeline for whole genome bisulfite sequencing data analysis. PLoS One *9*, e100360.
22. Krueger, F., Kreck, B., Franke, A., and Andrews, S.R. (2012). DNA methylome analysis using short bisulfite sequencing data. Nat. Methods *9*, 145–151.
23. Sun, X., Han, Y., Zhou, L., Chen, E., Lu, B., Liu, Y., Pan, X., Cowley, A.W., Jr., Liang, M., Wu, Q., et al. (2018). A comprehensive evaluation of alignment software for reduced representation bisulfite sequencing data. Bioinformatics *34*, 2715–2723.
24. Chen, H., Smith, A.D., and Chen, T. (2016). WALT: fast and accurate read mapping for bisulfite sequencing. Bioinformatics *32*, 3507–3509.
25. Sun, K., Jiang, P., Wong, A.I.C., Cheng, Y.K.Y., Cheng, S.H., Zhang, H., Chan, K.C.A., Leung, T.Y., Chiu, R.W.K., and Lo, Y.M.D. (2018). Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing. Proc. Natl. Acad. Sci. U S A *115*, E5106–E5114.
26. Sun, K. (2020). Ktrim: an extra-fast and accurate adapter- and quality-trimmer for sequencing data. Bioinformatics *36*, 3561–3562.
27. Barnett, K.R., Decato, B.E., Scott, T.J., Hansen, T.J., Chen, B., Attalla, J., Smith, A.D., and Hodges, E. (2020). ATAC-Me captures prolonged DNA methylation of dynamic chromatin accessibility loci during cell fate transitions. Mol. Cell *77*, 1350–1364.e6.
28. Spektor, R., Tippens, N.D., Mimoso, C.A., and Soloway, P.D. (2019). methyl-ATAC-seq measures DNA methylation at accessible chromatin. Genome Res. *29*, 969–977.
29. Lhoumaud, P., Sethia, G., Izzo, F., Sakellaropoulos, T., Snetkova, V., Vidal, S., Badri, S., Cornwell, M., Di Giammartino, D.C., Kim, K.T., et al. (2019). EpiMethylTag: simultaneous detection of ATAC-seq or ChIP-seq signals with DNA methylation. Genome Biol. *20*, 248.
30. Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.
31. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.
32. Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.
33. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
34. Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966–1967.
35. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.
36. Hansen, K.D., Langmead, B., and Irizarry, R.A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. *13*, R83.
37. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.
38. Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.
39. Sun, K., Wang, H., and Sun, H. (2017). mTFkb: a knowledgebase for fundamental annotation of mouse transcription factors. Sci. Rep. *7*, 3022.
40. Altun, G., Loring, J.F., and Laurent, L.C. (2010). DNA methylation in embryonic stem cells. J. Cell. Biochem. *109*, 1–6.

41. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

42. Sun, K., Lun, F.F.M., Jiang, P., and Sun, H. (2017). BSviewer: a genotype-preserving, nucleotide-level visualizer for bisulfite sequencing data. Bioinformatics *33*, 3495–3496.

43. Lun, F.M.F., Chiu, R.W.K., Sun, K., Leung, T.Y., Jiang, P., Chan, K.C.A., Sun, H., and Lo, Y.M.D. (2013). Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA. Clin. Chem. *59*, 1583–1594.

44. Sun, K., Lun, F.M.F., Leung, T.Y., Chiu, R.W.K., Lo, Y.M.D., and Sun, H. (2018). Noninvasive reconstruction of placental methylome from maternal plasma DNA: potential for prenatal testing and monitoring. Prenat. Diagn. *38*, 196–203.

45. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

46. Tanaka, K., and Okamoto, A. (2007). Degradation of DNA by bisulfite treatment. Bioorg. Med. Chem. Lett. *17*, 1912–1915.

47. Chan, K.C.A., Jiang, P., Sun, K., Cheng, Y.K., Tong, Y.K., Cheng, S.H., Wong, A.I., Hudecova, I., Leung, T.Y., Chiu, R.W.K., et al. (2016). Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. Proc. Natl. Acad. Sci. U S A *113*, E8159–E8168.

48. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745.