THE SCIENCE WORLD
A Monthly e Magazine
ISSN:2583-2212
Popular Article
December, 2025 Vol.5(12), 11295-11303

The Methylation Toolkit: What's new in Epigenomic Engineering?

Suruchi Sharma*[1] And Sahil Sharma[2]
[1]PhD Scholar Animal Biochemistry, ICAR-National Dairy Research Institute, Karnal, Haryana, [21]PhD Scholar Animal Nutrition, ICAR-National Dairy Research Institute, Karnal, Haryana
*Corresponding author's email: suruchis91@gmail.com
DOI:10.5281/ScienceWorld.18119851

Abstract

A potent method for identifying 5-methylcytosine in DNA, bisulfite sequencing has greatly advanced our knowledge of epigenetic control in both plants and animals. In the meanwhile, studies on further base alterations, such as 6-methyladenine and 4-methylcytosine, which are common in prokaryotes, has been hindered by the absence of a similar method. DNA degradation, lack of selectivity, and short reads with poor sequence diversity are just a few of the challenges that bisulfite sequencing faces. In this study, we examine the latest developments in bisulfite sequencing techniques that allow mapping single-cell methylomes, identifying derivatives of 5-methylcytosine, and focusing on genomic regions of interest. The special benefit of long-read sequencing in identifying base changes in native DNA is then discussed, along with the advantages and disadvantages of PacBio and Nanopore sequencing for this use. The capacity to identify different modified bases from a universal sample preparation, along with the mapping and phasing benefits of the longer read lengths, give long-read sequencing a clear advantage over short-read bisulfite sequencing for an increasing number of applications across kingdoms, even though analyzing epigenetic data from long-read platforms is still difficult.

Introduction

The classical DNA bases A, T, C, and G make up genomic DNA. Instead of altering the basic sequence, modified DNA bases convey an additional layer of information that frequently determines how that DNA sequence is used, such as designating sequences as endogenous or modifying transcription. The DNAmod database lists 43 DNA modifications seen in natural DNA, most of which are caused by DNA damage but some of which have regulatory functions. N6-methyladenine (6mA), 4-methylcytosine (4mC), and 5-methylcytosine (5mC) are common in bacteria and play roles in both cellular defense and gene expression control, which affects physiology and pathogenicity. The most common, researched, and well-understood alteration in both plants and animals is 5 mC. This is partly

because 5mC can be measured using precise bisulfite-based short-read sequencing methods. Nevertheless, bisulfite sequencing has some drawbacks and is difficult to use for other base alteration detection. However, with the benefits of long reads and single-molecule sequencing, new long-read sequencing methods present intriguing opportunities to investigate a variety of alterations.

In this study, we will go over the existing gold standard for detecting 5mC and its oxidized derivatives and contrast it with the potential that long-read sequencing from Pacific Biosciences and Oxford Nanopore technology offers, both now and in the future. However, we point out that there are numerous other ways to measure 5mC that we won't go over here, each of which might be helpful in particular situations and should be taken into account.

Bisulfite-converted DNA sequencing

After DNA is extracted, DNA methylation markers are still there. Methylated cytosines remain intact while unmethylated cytosines are deaminated to uracil when genomic DNA is treated with sodium bisulfite. In order to produce enough template for analysis, treated DNA is next PCR-amplified using a uracil-tolerant polymerase, which causes uracil to change into thymine. As a result, DNA methylation can be directly detected using conventional Sanger or Illumina short-read sequencing by comparing it to an untreated or reference sequence, yielding a highly quantitative readout with base-pair resolution.

Although 5mC is most frequently abundant in the CpG dinucleotide context in animals, it can also be found in the CHG and CHH contexts (where H is either A, C, or T), where research also points to potential functional significance. Methylation is prevalent in all CG, CHG, and CHH contexts in plants, however it varies greatly between species. Therefore, it can be very helpful that bisulfite sequencing provides information on all cytosines, regardless of context. Due to these characteristics, bisulfite-converted DNA sequencing has become the gold standard for 5mC detection methods; nonetheless, there are a few drawbacks to take into account. First of all, a lot of input DNA is frequently needed because the severe bisulfite treatment causes degradation and DNA that is more difficult to PCR amplify. Enzymatic Methyl-seq (EM-seq), a method recently described by New England Biolabs, leverages APOBEC's enzymatic deamination of cytosine to produce a sequence that is identical to bisulfite treatment without requiring the harsh chemical treatment. Furthermore, because bisulfite sequence data must be compared to a bisulfite-converted reference genome before methylation calls can be deduced, more advanced bioinformatic analysis approaches are needed than for unconverted DNA. There are several top-notch programs made especially to handle bisulfite sequence data. The same challenges that affect all short-read data are also

Official Website
www.thescienceworld.net
thescienceworldmagazine@gmail.com
Published 28/12/2025

present in Illumina sequencing of bisulfite DNA, especially when it comes to mapping problems to low complexity or repeating areas, including 5mC-relevant regions like repetitive DNA and substantially GC rich regulatory regions. The loss of sequence variety brought on by the bisulfite conversion exacerbates these problems even more. Furthermore, because short reads are less likely to include a meaningful single-nucleotide polymorphism (SNP), they are challenging to haplotype. Long-read sequencing, which is covered in more detail below, eliminates these short-read sequencing restrictions. Whole-genome bisulfite sequencing (WGBS), which is Illumina sequencing of total genomic DNA, offers the most thorough and objective survey of 5mC currently feasible. However, obtaining such comprehensive data requires a high number of reads, with a coverage of 5–15× recommended (approximately 160–480M 100-bp reads for a haploid human genome, pooling both DNA strands).

Methods of enrichment for bisulfite sequencing

Techniques that enrich for regions of interest have been developed to reduce the high costs of getting the required reads for high-quality WGBS data in big genomes. Amplicon sequencing is simple and economical for a limited number of genomic loci (≤ 20).

Before being amplified using certain primers and barcoding, DNA is first bisulfite-treated. It is then sequenced as a multiplex. Capture-sequencing eliminates the labor-intensive primer pair design for more regions, but it necessitates the synthesis of a probe panel. Either before (Agilent Sure-Select Methyl-Seq, TruSeq Methyl Capture) or after bisulfite conversion (Roche SeqCap Epi), capture via hybridization to certain probes can be carried out. In the latter scenario, there is a chance that biases in the measurement of methylation could be introduced by preferential binding of probes to specific methylation states of the target fragments. There are commercially available panels that work well for the human genome, but custom panels can be costly for one-off applications.

Reduced representation bisulfite sequencing (RRBS) provides a cost-effective method for enriching regions in mammals where mC regulation is more likely, such as CpG Islands, which are areas of high CpG density that frequently correspond to differentially methylated gene regulatory regions like enhancers and promoters. RRBS has been shown to be informative for 85% of CpG islands, which comprise less than 3% of the genome and significantly lower sequencing costs, by using MspI, a 5mC agnostic restriction enzyme that cuts at CCGG patterns.

The clear disadvantage of RRBS is that it is restricted to loci with MspI cut sites by design. CpG island beaches, or areas of moderate CpG density that border CpG islands, are also frequently shown to be differentially methylated. These can be identified by sequencing

Official Website
www.thescienceworld.net
thescienceworldmagazine@gmail.com
Published 28/12/2025

the longer restriction fragments using a method called improved RRBS. Another thing to keep in mind when using RRBS is that the MspI cut reduces diversity at the beginning of sequencing reads, which might cause problems for cluster detection and calibration on the newest Illumina sequencers. Nevertheless, this can be circumvented by employing adapters with diversity bases, spiking in libraries with high diversity, or hiding the initial bases of each read from the sequencer (a technique known as dark sequencing).

Using bisulfite sequencing to identify oxidized forms of methylation

In mammals, the TET family of dioxygenases oxidizes 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC). The DNA glycosylase TDG then excises 5fC or 5caC to complete the demethylation process. While 5fC and 5caC are susceptible to deamination and hence read as unmethylated cytosine, 5hmC, like 5mC, is shielded from bisulfite-induced deamination, making it impossible for bisulfite sequencing to distinguish between the two forms. Although modified bisulfite sequencing techniques have been developed to differentiate these oxidized forms, this is a significant caution to be taken into account for any bisulfite sequencing work. Only 5mC is shielded from bisulfite-induced deamination by oxidative bisulfite sequencing (OxBS-seq), which uses a particular chemical oxidation of 5hmC to 5fC. As a result, OxBS-seq provides precise information on 5mC without confusing 5hmC. By subtracting OxBS-seq from unoxidized bisulfite sequencing, positions of 5hmC can be found. By using the TET1 enzyme to oxidize 5mC to 5caC while shielding 5hmC from oxidation by adding glucose, TET-assisted bisulfite sequencing (TAB-seq) makes sites of 5mC, but not 5hmC, vulnerable to bisulfite-induced deamination. As a result, TAB-seq offers a direct measurement of 5hmC, and locations of 5mC can be ascertained by deducting the TAB-seq signal from conventional bisulfite sequencing. Bisulfite sequencing can be used to identify 5fC sites by either selectively reducing to 5hmC (redBS-Seq) or chemically shielding 5fC from bisulfite-induced deamination (fCAB-seq). Chemical changes can shield carboxylcytosine from deamination, making it identifiable by bisulfite sequencing (CAB-seq).

Bisulfite sequencing in a single cell

Until recently, genomic techniques, such as bisulfite sequencing, could only yield average values across bulk cell populations due to their high input needs. Single-cell genomics is now starting to provide previously unheard-of insights into cell-to-cell variation. In this case, bisulfite sequencing offers significant advantages over alternative 5mC detection techniques.

Official Website
www.thescienceworld.net
thescienceworldmagazine@gmail.com
Published 28/12/2025

First, the input need was significantly lowered to that of a single nucleus by switching the order of adapter ligation to after bisulfite treatment (also known as post bisulfite adapter tagging, or PBAT). However, adapter ligation is not without biases, and chimeric readings frequently arise. Second, bisulfite sequence data incorporates the measurement within the read, so every mappable read is informative, while methods that use read count-based statistics as measurements suffer from the limited coverage typical of single-cell sequencing. There are methods for single-cell RRBS and single-cell WGBS. Interestingly, bisulfite sequencing of the nucleus from within a single cell is being paired with RNA-seq from the cytoplasmic fraction to provide an unparalleled assessment of the functional relationship between the transcriptome and the epigenome. Currently available techniques to quantify combinations of 5mC, RNA, copy number, and nucleosome positioning are referred to as single-cell multi-omics. A thorough explanation of multi-omic techniques may be found in.

Using single molecule long-read sequencing to identify DNA methylation

The mappability issue with short reads can be resolved by long-read sequencing. There are now two long-read sequencing technologies available: single molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio) and nanopore sequencing from Oxford Nanopore Technologies (ONT). Although both approaches can produce a readout akin to short-read bisulfite sequencing when applied to bisulfite-treated and amplified DNA, their primary benefit is their capacity to sequence native DNA and deduce base modifications from their effects on the raw sequencing signal. Then, both amplification biases and DNA degradation via bisulfite conversion are prevented. In general, the range of assayable changes is increased and experimental complexity is decreased since enzymatic or chemical treatments unique to each base modification of interest are no longer required. This feature's drawback is that the DNA cannot be amplified, hence input amounts may be limited (200 fmol for Nanopore, or roughly 1 µg of 8 kb fragments; 5 µg for a typical PacBio library, however 100 ng may be adequate). Nanopore and PacBio native DNA sequencing will not be possible in situations where only small amounts of DNA are available due to experimental design (e.g., single small organism, microdissected tissue, single cells) or the value of the original tissue (e.g., biopsies or paleogenomic samples). As a result, these methods work best with large samples, limiting the granularity of the analyses and making it challenging to assess cell-to-cell variation.

PCR-free, CRISPR-based enrichment methods are available for both PacBio and ONT when only specific regions of the genome are of interest. It is possible to achieve enrichments

Official Website
www.thescienceworld.net
thescienceworldmagazine@gmail.com
Published 28/12/2025

of hundreds or thousands of times, offering tailored genetic and epigenetic assays that are affordable and sequence-effective.

Calling base changes at a single-base, single-read level from long-read sequencing is inaccurate compared to short-read bisulfite sequencing. Thus, accurate estimations are obtained by summarizing at genomic positions (requiring enough sequencing depth), regions, or motifs when appropriate, or by combining results over numerous passes (PacBio only). The unique advantages and disadvantages of PacBio and ONT affect their respective use cases.

SMRT sequencing

PacBio's SMRT sequencing is based on sequencing-by-synthesis, which uses a series of fluorescence pulses to determine the circular DNA template's sequence. Each pulse is produced when a polymerase attached to the bottom of a well adds one labeled nucleotide. Therefore, base changes alter the polymerase's kinetics but have no effect on the base-called sequence. Base modifications can be deduced from comparing a modified template to an unmodified template or an in silico model by taking the inter-pulse length into account. For instance, the polymerase prefers to incorporate the complementary T more slowly when a 6mA is present in the template strand. Kinetic perturbation patterns can be more intricate and context-dependent, and base alteration also affects the perturbations' magnitude.

Base alteration identification in single molecules is unreliable due to the poor signal-to-noise ratio and frequently necessitates summarizing at the genomic position-level. A coverage of 25× per strand is advised since 6mA and 4mC produce substantial kinetic fingerprints. However, unless they are enriched or altered to generate a bigger kinetic effect by glycosylation or TET-conversion to 5-carboxylcytosine, the modest effects of 5mC and 5hmC raise the needs to 250×. Therefore, PacBio sequencing can only reach single-molecule resolution for specific markers and on relatively short segments (≤ 2 kb) that the polymerase can read many times. Longer pieces make it impossible to thoroughly examine cell-to-cell variability. PacBio is especially well-suited to bacterial genomes, where 6mA and 4mC are common and frequently centered on certain motifs, because to its cost per Gb, sensitivity to specific mutations, resolution, and high coverage needs. Since 2012, the number of known methyltransferases has significantly increased because to the use of SMRT sequencing. Additionally, base J (β-d-glucosyl-hydroxymethyluracil) in Leishmania has been detected using SMRT sequencing, which may reveal undiscovered alterations. The Sequel II sequencer and v2 chemistry from PacBio significantly increased the throughput and cost-effectiveness of SMRT sequencing in 2019, producing up to 160 Gb per SMRT cell.

Official Website
www.thescienceworld.net
thescienceworldmagazine@gmail.com
Published 28/12/2025

Nanopore sequencing

When a single-stranded nucleic acid is ratcheted through a biological nanopore, ONT's nanopore sequencing monitors the change in ionic current. Basecalling is the technique by which neural networks convert the current trace into nucleotides. DNA base changes cause variations in the raw signal, which can be identified.

Three steps are often involved in detecting base modifications in ONT data: (1) basecalling with canonical bases, (2) anchoring the raw signal to a genomic reference, and (3) assessing the evidence that a base is modified. A well-known program called Nanopolish uses a pre-trained algorithm to identify 5mCG and has demonstrated a strong association with bisulfite data on the genomes of humans and mice. It is not necessary to sequence a PCR-amplified, unaltered control in addition to the sample of interest because Nanopolish includes a model for 5mCG. At a single-read, nearly single-nucleotide (really single k-mer) resolution, Nanopolish outputs the likelihood that a base is altered. The underlying algorithms and the changes they are taught to identify are different for other accessible tools: 6mA, 5mC, and 5hmC were detected by signal Align; 6mA and 5mCG were detected by mCaller, DeepSignal, DeepMod, and Megalodon; and BrdU detection is the focus of D-Nascent and RepNano. ONT-created Tombo offers a 5mC and 6mA model. Although 6mA detection is typically less reliable than 5mCG detection, improvements can still be made with better algorithms and training data. Similar to PacBio's base modification detection principles, in the absence of pre-trained algorithms for the base modification of interest, it can be deduced by comparing it to an in silico reference signal or, more efficiently, to a PCR-amplified control that is unaltered. It has only lately become feasible to basecall changes directly from the raw signal without the need for genomic anchoring. Although this method is currently not benchmarked and limited to 5mC in the CG and CC(A/T)GG contexts and 6mA in the GATC context, it is highly promising and reduces the need for computationally demanding downstream analysis.

The training data, which is usually made up of a fully unmodified sample (PCR-amplified or synthesized) and a fully modified sample (synthesized or modified in vitro by enzymes), significantly affects the performance of algorithms that rely on prior knowledge about the expected deviations in signal. Subpar performance results from motifs that are either absent from the training set or contain combinations of modified and unmodified bases. For instance, Nanopolish only shows the probability that the entire group is methylated on a pattern like CGCGT, not the percentages for individual cytosines.

Whole-genome bisulfite sequencing and whole-genome nanopore sequencing are similar in cost. Nanopore sequencing for base modification detection is currently in its early

Official Website
www.thescienceworld.net
thescienceworldmagazine@gmail.com
Published 28/12/2025

stages of research. The whole range of distinguishable modifications, the limits of sensitivity, and the lack of generalized algorithms capable of simultaneously calling numerous modifications are still unknowns. To comprehend the consistency of performance across species and sequencing batches, independent benchmarks of both established and new technologies are required. The raw signal varies each time ONT modifies the pore chemistry, necessitating a new training of the algorithms. Thankfully, a lot of solutions allow users to train the algorithms at the basecalling stage using their own data. It is possible that genetic and epigenetic information could soon emerge straight from the sequencer without additional processing thanks to recent developments in basecalling.

Discussion

A sensitive and quantitative method for base-resolution DNA methylation is offered by bisulfite sequencing. Targeted sequencing methods can avoid the high cost of deep coverage. However, 5mC and 5hmC cannot be distinguished by ordinary bisulfite conversion, and unique techniques are needed to identify each of these markers. This also applies to 5caC and 5fmC, and each mark of interest requires its own library preparation and sequencing. In contrast, a variety of base changes can be concurrently detected by SMRT and nanopore long-read sequencing without the need for extra sample preparation. When it comes to bacterial epigenomics, SMRT is most sensitive at detecting 4mC and 6mA. Although accuracy on m6A is comparable with PacBio, nanopore sequencing currently performs best at detecting 5mCG. Its cheaper cost per Gb when compared to SMRT makes it appropriate for bigger genomes. Hypothesis-free testing for novel base alterations is compatible with both technologies. Long-read methods continue to be less accurate at detecting 5mC than bisulfite sequencing; this trend is probably going to continue for other markers. Outside of motifs, this becomes especially troublesome because a high false positive rate could mask the signal in background noise. For this reason, orthogonal validation is advised. An effective way to raise the signal-to-noise ratio is to concentrate on particular themes where the mark of interest is prevalent. The creation of suitable training data, where DNA with known base modifications is present at known places in all biologically relevant motifs, is necessary to increase the accuracy and range of detectable alterations. Unfortunately, we are not as good at synthesizing DNA as we are at sequencing it. However, improvements in machine learning and long-read sequencing throughput indicate that PacBio and ONT sequencing accuracy can still be improved. A near future in which base alterations are a common feature of DNA and RNA sequencing is hinted up by the emergence of nanopore basecallers that incorporate changed bases.

Official Website
www.thescienceworld.net
thescienceworldmagazine@gmail.com
Published 28/12/2025

Bisulfite sequencing has been optimized for low input requirements, making it appropriate for single-cell sequencing. Although SMRT and nanopore sequencing are single-molecule methods, they currently need more than 100 ng of DNA and are not single-cell methods. A significant obstacle to using long-read sequencing to single-cell epigenomics is the loss of base modifications during PCR.

A significant benefit of long-read sequencing is its capacity to phase genetic and epigenetic information, producing allele-specific 5mC patterns that provide insight into the impact of mutations, structural variations, or parental origin on gene regulation, in addition to the detection of base modifications not amenable to bisulfite sequencing. Additionally, across repeat-rich regions that are resistant to short-read sequencing, long-read sequencing yields genetic and epigenetic information. It has proven challenging to investigate some clinical disorders associated with repeat expansions and unsuccessful epigenetic control using short reads. It is anticipated that long-read sequencing would significantly advance our understanding of these illnesses' biological causes and diagnosis.

It's fascinating and difficult to quickly iterate over ONT nanopore chemistry, methods, and software. In contrast to bisulfite sequencing, long-read epigenetics has few well-established analytical pipelines. In order to assess the effectiveness of the various tools, benchmarking activities are essential. Long-read sequencing takes us closer to acquiring full-length, complete, and phased epigenomes, even though we are still far from reading out the 43 base modifications described in DNAmod.

Conclusion

5mC, 5hmC, 5fC, and 5caC can be mapped at base resolution using variations of short-read bisulfite sequencing.

Bisulfite sequencing can be used to acquire single-cell methylomes, which can then be coupled with other omics to explore epigenetic regulation at single-cell resolution.

PacBio is best suited for bacterial epigenomics and identifies 6mA and 4mC with the highest sensitivity.

Nanopore sequencing is currently being developed and is capable of resolving 6mA, 5mC, 5hmC, and BrdU in single molecules.

Compared to bisulfite sequencing, long reads enhance phasing, genomic coverage, and epigenome completeness without the need for specialized chemicals.

References

Gouil Quentin and Keniry Andrew. Latest techniques to study DNA methylation, Essays in Biochemistry (2019), vol 63: pp 639-648 (https://doi.org/10.1042/EBC20190027)

Official Website
www.thescienceworld.net
thescienceworldmagazine@gmail.com
Published 28/12/2025