Briefings in Bioinformatics, 19(5), 2018, 737–753

doi: 10.1093/bib/bbx013
Advance Access Publication Date: 8 March 2017
Paper

A survey of the approaches for identifying differential methylation using bisulfite sequencing data

Adib Shafi, Cristina Mitrea, Tin Nguyen and Sorin Draghici

Corresponding author. Sorin Draghici, Department of Computer Science and Department of Obstetrics and Gynecology, Wayne State University, 14th Floor, Suite 14200, 5057 Woodward, Detroit, MI 48202. Tel.: +1(313)577-2162; Fax: +1(313) 577-6868; E-mail: sorin@wayne.edu

Abstract

DNA methylation is an important epigenetic mechanism that plays a crucial role in cellular regulatory systems. Recent advancements in sequencing technologies now enable us to generate high-throughput methylation data and to measure methylation up to single-base resolution. This wealth of data does not come without challenges, and one of the key challenges in DNA methylation studies is to identify the significant differences in the methylation levels of the base pairs across distinct biological conditions. Several computational methods have been developed to identify differential methylation using bisulfite sequencing data; however, there is no clear consensus among existing approaches. A comprehensive survey of these approaches would be of great benefit to potential users and researchers to get a complete picture of the available resources. In this article, we present a detailed survey of 22 such approaches focusing on their underlying statistical models, primary features, key advantages and major limitations. Importantly, the intrinsic drawbacks of the approaches pointed out in this survey could potentially be addressed by future research.

Key words: DNA methylation; epigenetic modification; differentially methylated cytosines (DMCs); differentially methylated regions (DMRs); bisulfite sequencing

Introduction

Epigenetics is the field of study that provides information on how, where and when genes are switched on and off inside a living cell. DNA methylation is an intensively studied and well understood epigenetic mechanism that plays a vital role in many processes [1]. Due to its role in regulating gene expression, DNA methylation is an important part of cellular processes such as cell development and differentiation. Furthermore, patterns of hypermethylation have been identified in human cancers, which can provide novel insights into the development and progression of such complex diseases [2]. Specifically, in cancer, one of the causes of silenced tumor suppressor genes is hypermethylation.

The most studied form of DNA methylation, known as 5-methylcytosine (5-mc), involves the addition of a methyl group to the 5-carbon of the cytosine (C) base of a DNA strand. Although approximately only 5% of the cytosine bases in the human genome are methylated, cytosine (C) followed by a guanine (G), which is known as a CpG site, is methylated 70–80% of the time [3, 4]. Methylation can also occur in non-CpG context, such as CHG and CHH sites (where H = C, T or A),

Adib Shafi is a PhD candidate in the Department of Computer Science at Wayne State University, USA. His research interests include biological pathway analysis, finding mechanism using multi-omics data and variant analysis.
Cristina Mitrea is a PhD candidate in the Department of Computer Science at Wayne State University, USA. Her work is focused on research in data mining techniques applied to bioinformatics and computational biology. Other interests include network discovery and meta-analysis applied to pathway analysis.
Tin Nguyen received his PhD from the Computer Science Department at Wayne State University. His research interests include computational and statistical methods for analyzing high-throughput data. His current foci are meta-analysis and multi-omics data integration.
Sorin Draghici currently holds the Robert J. Sokol, MD Endowed Chair in Systems Biology, as well as appointments as Full Professor with the Department of Computer Science and the Department of Obstetrics and Gynecology, Wayne State University. He is also the head of the Intelligent Systems and Bioinformatics Laboratory in the Department of Computer Science. His work is focused on research in artificial intelligence, machine learning and data mining techniques applied to bioinformatics and computational biology. He has published 2 best-selling books on data analysis of high-throughput genomics data, 8 book chapters and over 190 peer-reviewed journal and conference papers.
Submitted: 1 September 2016; **Received (in revised form):** 14 January 2017

especially in plants and stem cells [5, 6]. Recent studies have also shown that the Ten-Eleven translocation (TET) proteins are involved in oxidizing 5-mc into 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine (5-fC) and 5-carboxylcytosine (5-caC). However, the abundance level of these methylation variants (5-hmC, 5-fC, 5-caC) is low compared with that of 5-mc [7]. Therefore, our survey focuses on 5-mc methylation in CpG context, considering most of the methods have been developed for analyzing this type of epigenetic modification. When a CpG site is methylated in the promoter regions, it typically represses the transcriptional activity of that region by restricting the binding of specific transcription factors (TFs). Alternatively, when a CpG site is unmethylated in promoter regions, it allows for the binding of those TFs [8–10]. Given its regulatory role in cellular activities, identifying changes in DNA methylation across multiple biological conditions is of great interest.

The availability of the reference genome and the advanced sequencing technologies have led to methods that provide high-resolution methylation profiles on a genome scale. Based on the resolution at which the methylation levels are measured, current sequencing-based technologies can be divided in two categories: (i) enrichment-based approaches, and (ii) bisulfite sequencing-based approaches [11, 12]. The former allows us to measure the methylation levels at 100–200 base resolution, while the latter allows us to measure the methylation levels at single-base resolution. One of the challenges in measuring genome-level methylation is the amount of biological material needed, which has only recently reached levels feasible for clinical samples [13]. Other challenges are related to processing data from new technologies and integrating them with different types of data in a meaningful way to provide biological insights (e.g. methylation and gene expression). In this review, we focus on bisulfite sequencing-based approaches.

Within the past few years, many tools have been developed for differential methylation (DM) analysis using bisulfite sequencing data (Figure 1), but only a few attempts have been made to provide a review of these approaches. Robinson et al. [14] provides a mini review of the approaches that identify DM, briefly discussing their methodologies and current challenges. This review not only includes the approaches that use bisulfite sequencing data but also the approaches that use DNA methylation arrays (Illumina's 27k or 450k) and enrichment assays (MeDIP-seq). Yet, the number of approaches based on bisulfite sequencing data and the number of features considered for each approach are low. Klein et al. [15] evaluates nine approaches that can possibly be used for DM analysis. However, the methods are limited to the scope of analyzing DM in predefined regions using only reduced representation bisulfite sequencing (RRBS) data. Among the nine approaches, only four of them are originally designed for analyzing DM. The other five approaches are general approaches that can be applied for RNA-Seq and gene expression data. Yu and Sun [16] evaluate only five approaches developed for the purpose of identifying differentially methylated regions (DMRs). Sun et al. [17] briefly summarize the commonly used platforms for methylation profiling, data preprocessing techniques and statistical approaches for DM analysis. This review provides a well-organized conceptual overview of approaches that identify DM using bisulfite sequencing data. However, this survey only includes seven such approaches. In summary, all previous attempts of reviewing the approaches that identify DM using bisulfite sequencing data are limited in at least one of the following aspects: (i) the total number of approaches covered in the survey (fewer than 10 methods reviewed), (ii) the applicability (e.g. only methods dealing with RRBS data) or (iii) a small number of biological features considered. To address these issues, a comprehensive survey of the approaches that identify DM using bisulfite sequencing data is greatly needed.

In this article, we review 22 different approaches for DM analysis, including approaches for identifying differentially methylated cytosines (DMCs), DMRs (both predefined and *de novo* regions) and methylation patterns using bisulfite sequencing data (whole genome bisulfite sequencing [WGBS] and RRBS). We classify these approaches into seven different categories based on the primary concepts and key techniques used to identify DM. In addition, we provide a short overview of several general hypothesis-based tests, which can also be applied for DM analysis. In the following sections, first, we will provide a brief overview of bisulfite sequencing technology and the workflow of analyzing bisulfite sequencing data. Next, we will provide a systematic review of the approaches highlighting their pros and cons, discussing their key characteristics.

Bisulfite sequencing

The gold standard for measuring cytosine methylation is bisulfite sequencing, which has the advantage of measuring methylation at single-base resolution. In this technique, DNA is treated with sodium bisulfite, which deaminates unmethylated cytosines (C) to uracils (U) leaving the methylated cytosines unchanged. Uracils are read as thymines (T) during the sequencing step. Methylation level at each CpG site is estimated by simply counting the ratio of C/(C + T). Thus, this process allows sequence-specific discrimination between methylated and unmethylated CpG sites [18].

Several technologies have been developed for measuring DNA methylation based on bisulfite sequencing conversion. The most comprehensive protocol among them is WGBS, which provides genome-wide DNA profiling. However, the application of this protocol on the whole genome is expensive when it comes to studying organisms with large genomes. More cost-effective protocols, such as RRBS and enhanced RRBS, have

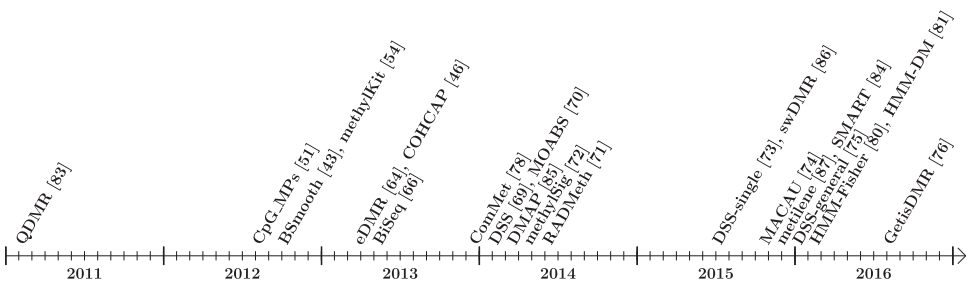

Figure 1. Timeline of the approaches that identify DM using bisulfite sequencing data.

allowed for methylation analysis with reduced sequencing requirements through a more targeted approach for CpG-rich genomic regions that meet specific length requirements [19]. These techniques therefore are more affordable for studies with multiple replicates.

The overall workflow for bisulfite sequencing data analysis is displayed in Figure 2. The overall pipeline consists of six major elements: (i) the input including methylation data (in FASTA/FASTQ format) and the reference genome, (ii) data processing and quality control, (iii) alignment of short reads to the reference genome, (iv) post-alignment analysis, (v) DM analysis and (vi) the output including DMCs, DMRs and methylation patterns. The details of each element will be described in the following sections.

Pre-analysis

Data preprocessing

Bisulfite sequencing data consist of short read sequences in the FASTA/FASTQ file format. Data processing starts with performing quality control operations on the raw sequencing reads, including quality trimming and adapter trimming. Quality trimming reduces methylation call errors by trimming the bases that have poor quality scores, whereas adapter trimming removes the known adapters from short reads to increase mapping efficiency. Existing tools for quality control include FASTX-Toolkit [20], PRINSEQ [21], SolexaQA [22], Cutadapt [23], Trimmomatic [24] and Trim Galore! [25]. Both the input and output of these tools are files in the FASTA/FASTQ format.

Read mapping

After quality control, bisulfite sequencing reads can be aligned to the reference genome to estimate the methylation levels. Simply aligning these reads by using standard aligners results in poor mapping efficiency because the bisulfite treatment introduces additional discrepancies between the sequencing reads and the reference genome by converting the unmethylated cytosines to thymines. Therefore, new strategies were proposed for bisulfite sequencing read alignment. Existing bisulfite sequencing alignment approaches can be divided in two categories: three-letter aligners and wildcard aligners. Three-letter aligners, such as Bismark [26], BS Seeker [27], MethylCoder [28], BRAT [29] and GNUMAP-bs [30], convert all Cs into Ts in the forward strand and all Gs into As in the reverse strand of the reference genome. Equivalently converted reads are then aligned to these pre-converted forms of the reference genomes using

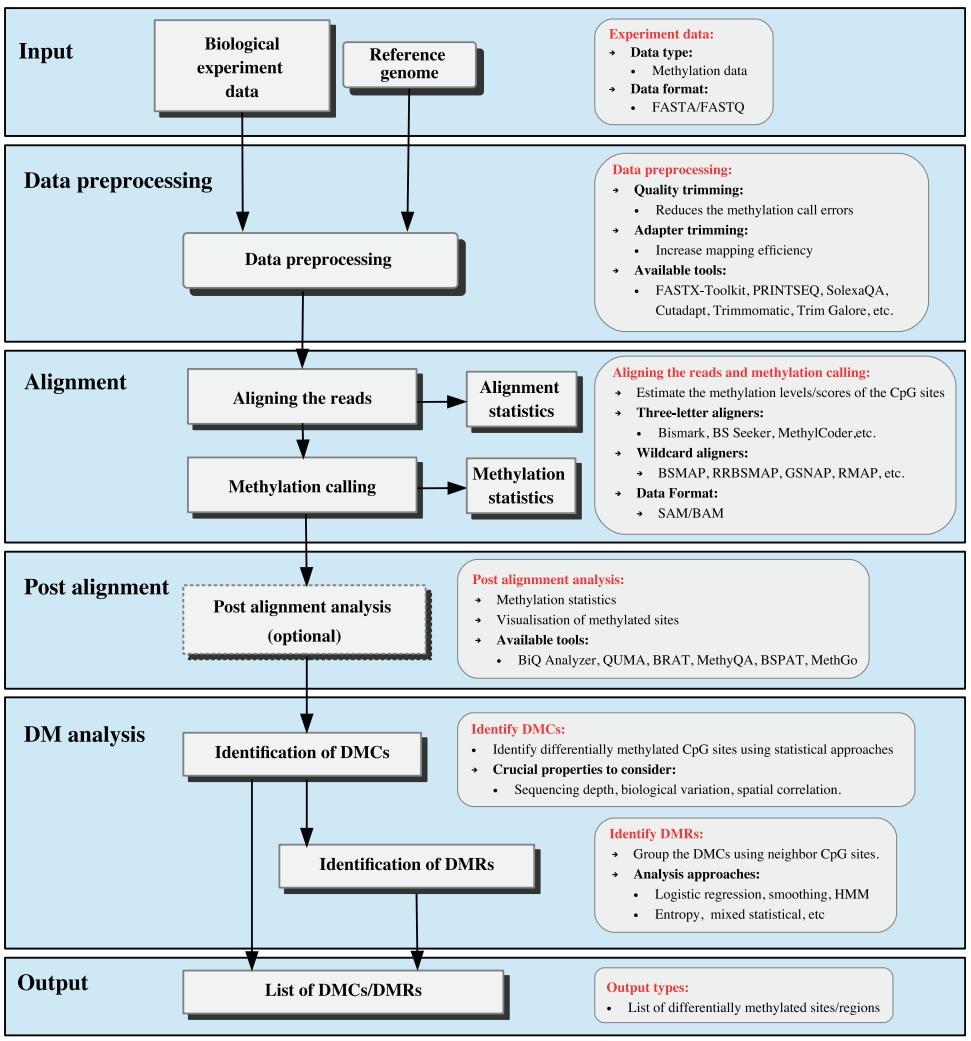

Figure 2. The workflow of analyzing DNA methylation using bisulfite sequencing data.

standard genome aligners such as Bowtie [31] and Bowtie2 [32]. In contrast, wildcard aligners, such as BSMAP [33], RRBSMAP [34], GSNAP [35] and RMAP [36], replace the Cs of the reference genome with the wildcard letter Y that matches both Cs and Ts in the sequencing reads. The alignment results are usually stored in SAM/BAM file format.

Post-alignment analysis

After mapping the reads, an optional post-alignment step can be performed to extract meaningful biological information from the alignment results before DM analysis. Several post-alignment analysis tools have been developed, including BiQ Analyzer [37], QUMA [38], BRAT [29], MethyQA [39], BSPAT [40] and MethGo [41]. Most of these tools provide summary statistics, quality assessment and visualization of the methylation data. Some of these tools include extra features such as read mapping (e.g. BSPAT and BRAT), identifying DNA methylation co-occurrence pattern (e.g. BSPAT), single nucleotide polymorphisms and copy number variation calling (e.g. MethGo) and detecting allele-specific methylation patterns (e.g. BSPAT).

DM analysis

After obtaining the methylation information of the CpG sites, typically the next downstream analysis is to perform DM analysis, which is usually done in the form of identifying DMCs or DMRs. Identification of DMCs involves comparing the methylation level at each CpG site across the phenotypes (two or more) and applying statistical tests for hypothesis testing. Identification of DMRs is usually a two-step process: (i) the identification of DMCs, and (ii) grouping the neighboring DMCs as contiguous DMRs by certain distance criteria. However, some approaches can directly identify DMRs. DMCs/DMRs occasionally can be linked to transcriptional repression of the associated genes; therefore, they provide crucial biological insights that may lead to the development of potential drug candidates [1].

To identify putative potential DMCs/DMRs from bisulfite sequencing data, some characteristics need to be considered. One such characteristic is the 'spatial correlation' between the methylation levels of the neighboring CpG sites, which plays an important role in getting an accurate estimation of the methylation levels [3, 42]. Incorporating spatial correlation in DM analysis can reduce the required sequencing depth and can estimate the methylation status of the missing CpG sites [43]. 'Sequencing depth' is another important characteristic that is directly related to the certainty of the methylation scores of CpG sites. Considering sequencing depth while identifying DMRs is crucial because it can take into account the sampling variability that occurs during sequencing. Another such characteristic is 'biological variation' among replicates, which is crucial in identifying the regions that consistently differ between groups of samples [44, 45]. Ignoring biological variation while detecting DMRs might lead to a high number of false positives in the results [14, 43, 46]. This is due to the fact that the methylation levels of the CpG sites are heterogeneous not only when the cell types are different but also when the cells are of the same type [47–50].

Classical hypothesis testing methods, such as Fisher's exact test (FET), chi-square (χ^2) test, regression approaches, *t*-test, moderated *t*-test, Goeman's global test and analysis of variance (ANOVA), can be used to identify DM using bisulfite sequencing data [3, 46, 51, 52, 53]. These approaches can be divided into two categories based on the data type they use: count-based hypothesis tests and ratio-based hypothesis tests.

Count-based hypothesis tests

Input of these hypothesis testing methods are count values, which can be either the number of reads or the number of CpG sites in a predefined genomic region. FET is a classical statistical test used to determine whether there are nonrandom associations between two categorical variables. In the context of methylation analysis, we can use the data to build a contingency table, where the two rows represent the two methylation states, and the two columns represent a pair of samples. When applying FET for two groups of samples, the counts for a methylation status within each group are aggregated into a single number [54]. Chi-square test is another classical method to test the relationship between two categorical variables (methylated versus unmethylated). In contrast with FET, it allows for testing across multiple samples. As pointed out by Sun et al. [17] and Hurlbert et al. [55], there are several issues related to the aggregation of read counts into a single number while applying tests of independence (FET and χ^2 test). First, the read counts are not independent; they represent different sets of interdependent or correlated observations. Thus, aggregating the counts violates the fundamental assumption underlying the test for independence. Second, due to uneven coverage of each individual site, the results are biased toward the samples with higher coverage. Third, by aggregating (summing) the counts, some of the biological variations (e.g. sample size, intra-group variance) is not taken into account by the hypothesis testing. Therefore, using FET and χ^2 test to compare two groups of samples could lead to a high number of false positives [14, 43, 46].

Regression approaches (e.g. Poisson, quasi-Poisson, negative binomial regression) are primarily used for detecting differentially expressed genes using RNA-Seq data, but they can also be applied in the context of DM analysis [15]. For example, the read counts can be modeled using a Poisson distribution and a modified Wald test can be used to detect DM as the difference between two Poisson means [56, 57].

Ratio-based hypothesis tests

These hypothesis tests use methylation percentage (methylation ratio) instead of count values. For a particular CpG site, methylation percentage is calculated by taking the ratio between the methylated read counts and the total read counts of that site. To compare the methylation difference level between two groups (phenotypes) of samples, classical tests such as *t*-test [58, 59], moderated *t*-test (limma) [60] or Goeman's global test [61] can be used. While *t*-test is a classical approach to compare the means, limma and Goeman's test are empirical Bayesian approaches that were primarily designed to detect differentially expressed genes using microarray data. When analyzing methylation levels across multiple groups of samples, ANOVA [62] can be used instead of multiple pair-wise comparisons. Compared with count-based hypothesis tests, the ratio-based tests take into account the biological variation across multiple replicates. However, because they only take into account the ratio of the reads (methylated reads versus all reads), they ignore the sequencing depth within the CpG sites.

Although classical hypothesis testing methods are somewhat useful, straightforward and easy to use, they are not efficient in more sophisticated methylation analysis, such as identifying *de novo* regions, considering spatial correlation among the methylation levels of the CpG sites and estimating

methylation levels of missing CpG sites. Over the past few years, several approaches have been developed to address these challenges, which are discussed and summarized in the following subsections.

Logistic regression-based approaches

Approaches in this category model the read counts of the CpG sites by using logistic regression to identify DM. One of the popular approaches in this category is 'methylKit' [54], which uses logistic regression to model the methylation proportion at a given base or region when biological replicates are available. In the absence of biological replicates, methylKit uses FET to identify DM. *P*-values are corrected using the false discovery rate (FDR) approach or the sliding linear model approach [63]. MethylKit is commonly used to identify DMCs from predefined regions (RRBS data). However, it can also be used to identify DMRs from WGBS data based on user-defined tiling windows. Major contribution of methylKit is that it can take into account the sequencing coverage. It can incorporate additional covariates into the model and work with CHG or CHH methylation. It also provides functionalities such as sample-wise methylation summary, sample clustering, annotation and visualization of DM, etc.

Another method named 'eDMR' [64] was proposed as an extension of methylKit. eDMR models the distances between the neighboring CpG sites using a bimodal normal distribution and estimates DMR boundaries using a weighted cost function. After estimating the regional boundaries, DMRs are filtered based on the mean methylation difference, the number of DMCs and the number of CpG sites. Significance of the DMRs are calculated by combining the *P*-values of the DMCs using Stouffer-Liptak method [65]. The *P*-values for DMRs are then corrected for multiple comparisons using the FDR method. eDMR provides a list of DMRs and their annotation as output.

Approaches in this category take sequencing coverage into account. They can incorporate additional covariates into the model as well. However, they do not consider the biological variation among the replicates. Although eDMR estimates the significance of the identified regions based on spatial auto correlation, it does not consider the spatial correlation among the CpG sites when estimating the methylation levels.

Smoothing-based approaches

Approaches in this category assume that methylation levels of the CpG sites vary smoothly across the genome. They perform 'smoothing' across the samples or predefined regions, which is a technique to estimate the methylation levels of the CpG sites by borrowing information from their neighbors. Group differences across different conditions are computed based on the estimated methylation values of the CpG sites. Finally, different statistical tests are used to identify the differentially methylated sites or regions.

One of the most commonly used smoothing-based approaches is 'BSmooth' [43], which relies on smoothing across the genome within each sample. It looks for group differences via CpG-wise *t*-tests to identify DMRs between two groups. The BSmooth algorithm begins with aligning the sequencing reads to the reference genome. Two alternative pipelines are available for the users to align the reads. The first pipeline, which supports gaped alignment and the alignment of the paired-end bisulfite-treated reads, is based on *in silico* bisulfite conversion that uses the 'Bowtie-2' aligner to align the reads [32]. The second pipeline is based on a newly developed aligner named 'Merman', which supports the alignment of the colorspace bisulfite reads. After aligning the reads, sample-specific quality assessment metrics are compiled. Local likelihood smoothing is applied within a smoothing window across the samples to estimate the methylation levels of the CpG sites. A signal-to-noise statistic similar to *t*-test is used to identify the DMCs. Finally, DMRs are defined by merging the consecutive DMCs based on some defined criteria, such as a cutoff value of the *t*-statistic, maximum distance between the CpG sites and minimum number of CpG sites.

BSmooth was the first approach primarily developed for DMR identification that takes into account the biological variation among replicates. It reduces the required sequencing coverage by applying the local likelihood smoothing approach across the samples. It can also identify *de novo* regions from WGBS data sets. On the other hand, BSmooth lacks suitable error measurement criteria within the identified DMRs. As a result, there is no way to check whether the identified CpG sites inside the predicted DMRs are true DMCs or selected erroneously. BSmooth predicts methylation values of the CpG sites based on the last observed slope. Hence, for the genomic regions that are not covered by the reads, previously observed methylation level will continue, resulting in a biased estimation of the methylation level (i.e. extrapolated methylation values of 0 and 1) [66]. BSmooth is not applicable to those data sets that do not have biological replicates. In addition, BSmooth is limited to comparisons between two groups of conditions.

Another approach in this category, 'BiSeq', performs the smoothing of methylation data across defined candidate regions instead of across the samples (like BSmooth) [66]. The pipeline begins with defining CpG clusters within the genome based on a minimum number of 'frequently covered CpG sites' (CpG sites that are covered by the majority of samples) and a proximity distance threshold defined by the user. A smoothing function is modeled for each defined cluster. While modeling the smoothing function, the coverage information for each CpG site is taken into account to make sure that the CpG site with high coverage has a greater impact on the estimated methylation level than the CpG site with low coverage. Group effects of the CpG sites are modeled using beta regression with probit link function. DMCs are identified using Wald test procedure. Next, a hierarchical testing procedure is applied to identify significant clusters containing at least one DMC. While testing the target regions, weighted FDR is applied to take into account the size of individual clusters [67]. A location-wise FDR approach is applied to trim the CpG sites that are not differentially methylated within the selected significant clusters.

One of the major contributions of BiSeq approach is that it provides region-wise error control measurement to test the target regions. This approach is also capable of adding additional covariates to the regression model. In contrast, one of the limitations of the BiSeq approach is that it is only suitable for analyzing experiments that have predefined regions such as RRBS data sets.

In general, smoothing-based approaches have the advantage of considering the spatial correlation between the methylation levels of the CpG sites. By performing smoothing, the required sequencing coverage and the variance of the methylation levels can be reduced [43]. Furthermore, they can estimate the methylation levels of missing CpG sites. On the other hand, smoothing-based approaches cannot detect the low CpG density regions where methylation has sharp changes such as transcription factor binding sites (TFBS). TFBS are usually small

(i.e. <50 bp), which might consist of a single CpG that is differentially methylated [68]. Thus, biological events involving a single CpG site might not be detected by the smoothing approaches. In addition, these approaches are not appropriate for biological systems whose true methylation levels of the CpG sites are not spatially correlated.

Beta-binomial-based approaches

Approaches in this category characterize the methylation read counts as a beta-binomial distribution. In the absence of any biological or technical variation, methylation proportion of a particular CpG site follows a binomial distribution because sequencing reads over a CpG site can be either methylated or unmethylated. Whenever biological and technical variation are present in the data, methylation proportions of the CpG sites are assumed to follow a beta distribution. Therefore, in the presence of biological replicates, an appropriate statistical model for methylation analysis is the beta-binomial model, as it can take into account both sampling and biological variability.

Over the past few years, several beta-binomial-based approaches have been developed to identify DM, such as DSS [69], MOABS [70], RADMeth [71], methylSig [72], DSS-single [73], MACAU [74], DSS-general [75] and GetisDMR [76]. These approaches differ from each other in the way they estimate regression parameters, calculate *P*-values, estimate DMR boundaries, etc.

'DSS' is one of the approaches in this category that relies on a beta-binomial hierarchical model to identify DM using bisulfite sequencing data. In this model, the prior distribution is constructed from the whole genome, which is either methylated or unmethylated. True methylation proportions of the CpG sites among the replicates are then modeled using the beta distribution parameterized by group mean and a dispersion parameter. The biological variability is captured by the beta distribution, whereas the sampling variability is captured by the binomial distribution. Variation across the methylation proportion of the CpG sites relative to the group mean is captured by the dispersion parameter, which is estimated by an empirical Bayes approach. When the sample size is small, a shrinkage approach is used to estimate the dispersion parameter to improve the overall performance. Differentially methylated CpG sites are determined by using *P*-values from the Wald test, which is performed by comparing the mean methylation levels between two groups. Lastly, candidate DMRs are defined by applying user-specified thresholds on DMR characteristics among which are *P*-value, minimum length and minimum number of CpG sites.

The key contribution of the DSS approach is the shrinkage procedure that improves the dispersion parameter estimation. For this reason, this approach is particularly useful when the sample size is small. By applying the Wald test procedure, this approach takes into consideration the biological variation and sequencing coverage.

A more recent method, named 'DSS-single', is an improved version of the DSS approach, which can take into account the spatial correlation among the CpG sites across the genome. In addition, DSS-single considers the within-group variation without biological replicates by using the neighboring CpG sites as 'pseudo-replicates'. Similar to DSS, DSS-single captures the technical variability using binomial distribution and the biological variability using beta distribution. The beta distribution is parameterized with the group mean and dispersion parameter. DSS-single estimates the group mean using a smoothing function and the dispersion parameter using an empirical Bayes procedure. Hypothesis testing is performed using the Wald test to identify the DMCs. Later, user-defined thresholds are applied to define the DMR boundaries and select candidate DMRs.

An even more recent variation of DSS approach, named 'DSS-general', identifies differentially methylated loci (DML) from bisulfite sequencing data under general experiment design. DSS-general identifies DML by modeling the methylation count data for each locus using the beta-binomial regression with the 'arcsine' link function. The 'arcsine' link function is applied to perform a data transformation that decreases the dependency of the data variance on the mean and prepares it for the next step. Due to this data transformation, the regression coefficient and the variance matrix can be estimated by applying the generalized least square method, as opposed to the beta-binomial generalized linear model or logistic regression, which are limited when values are separable (e.g. values for unmethylated sites are close to 0, values for methylated sites are close to 1). Finally, Wald test is used to perform hypothesis testing.

The key advantage of DSS-general approach is that it is applicable to bisulfite sequencing data with multiple groups or covariates. In addition, it uses 'arcsine' link function, which is more efficient than other widely used 'logit' and 'probit' functions because it estimates the regression parameters in one iteration.

'MOABS' is another approach that relies on beta-binomial assumption to identify DM. Similar to DSS, the prior distribution is constructed from the whole genome, resulting in a bimodal distribution. The posterior distribution follows a beta distribution, which is estimated using an empirical Bayes approach. When biological replicates are available, the posterior distribution is generated using the maximum likelihood approach. The significance of the DM between two samples is represented by a single metric named 'credible methylation difference', which incorporates both the biological and statistical significance of the DM. MOABS can also work with CHG or CHH methylation.

'RADMeth' is another analysis pipeline that relies on the beta-binomial assumption. RADMeth uses a beta-binomial regression approach using 'logit' link function to model the methylation levels of the CpG sites across the samples. Regression parameters are estimated using a standard maximum likelihood approach. In the beta-binomial regression model, RADMeth incorporates the experimental factors using a model matrix. The DM of a particular site is determined by comparing two fitted regression models (i.e. reduced model without factors and full model with factors) using the log-likelihood ratio. Subsequently, *P*-values of the neighboring CpG sites are combined using the weighted Z-test (i.e. Stouffer-Liptak test [77]) to obtain the DMRs. The key contribution of this approach is the ability to analyze WGBS data in multiple factor experiments.

'MethylSig' is another analysis pipeline that uses beta-binomial model across the samples to identify either DMCs or DMRs. The pipeline begins with taking the number of Cs and Ts as input. The approach uses the beta-binomial model to estimate the methylation levels at each CpG site or region, which involves the two following steps: (i) estimate the dispersion parameter for each CpG site or region, which accounts for biological variation among the samples within a group; and (ii) calculate the group methylation level at each CpG site or region using the estimated dispersion parameters. In each step, local information can be incorporated from nearby CpG sites or regions to increase statistical power. The significance level of the

methylation difference is calculated using the likelihood ratio test. Similar to DSS, MethylSig is useful when the sample size is small. MethylSig uses local information and a maximum likelihood estimator to compute both the methylation level and the variance.

'MACAU' is based on binomial mixed model (BMM) that takes into account the population structures from a data set. This model is a generalized beta-binomial model consisting of an extra term to model the population structure. In the absence of that extra term, this model can be reduced to a beta-binomial model. In this approach, the prior distribution is constructed from a BMM, whereas the posterior distribution is constructed from a log-normal distribution. Model parameters are estimated by using a Markov chain Monte Carlo (MCMC) algorithm-based approach. Hypothesis testing is performed by using Wald test. Finally, DMRs are constructed by merging the DMCs using empirical thresholds.

One advantage of this approach is that it can add a predictor variable of interest in the model to check the association with any genetic background. In addition to considering biological variability among the replicates and the sampling variability among the sequencing reads, this method also takes into consideration the population variability. Furthermore, it can be applied to both WGBS and RRBS data sets.

'GetisDMR', a recent beta-binomial-based approach, identifies variable-size DMRs directly from WGBS data by using a local Getis-Ord statistic, which is commonly used to identify statistically significant spatial clusters (hotspots). By incorporating this statistic into DM analysis, GetisDMR accounts for spatial correlation among the methylation levels of the CpG sites, along with the biological and sampling variability. When biological replicates are available, beta-binomial regression with logistic link function is used to model the methylation level of each CpG site. Model parameters are estimated by using the maximum likelihood function. Hypothesis testing is performed by using the likelihood ratio test. In the absence of biological replicates, methylation levels are modeled by using binomial distribution, and hypothesis testing is performed by using FET. *P*-values from the hypothesis testing are further used to calculate z-scores. Finally, a local Getis-Ord statistic is used based on the z-scores to identify DMRs using the information from the neighboring CpG sites. The Getis-Ord statistic uses the distribution of the data (i.e. z-scores) to compute a score of the nonrandom association between a data point and its neighbors, where a positive score shows a positive association and a negative score shows a negative association. This statistic is then used to identify data regions with points that exhibit nonrandom associations (i.e. DMRs).

One of the primary strengths of GetisDMR is that it can detect DMRs with variable length, instead of depending on user-specified threshold parameters. It can take into account the spatial correlation between the neighboring CpG sites. Additionally, it can incorporate additional confounding factors into the model. Furthermore, it can work with multiple groups, with or without biological replicates. One drawback of this approach is that it cannot work with enriched regions, such as RRBS data.

Beta-binomial-based approaches are useful because they take into account both sampling variability among the read counts and biological variability among the replicates. Furthermore, these approaches are able to identify DM at single-base resolution from low CpG-density regions (e.g. TFBS). On the other hand, most of the beta-binomial-based approaches (except DSS-single, MACAU and GetisDMR) do not take into account the spatial correlation between the methylation levels of the CpG sites.

Hidden Markov model-based approaches

Approaches in this category use hidden Markov model (HMM) to identify differentially methylated patterns from bisulfite sequencing data. These approaches model the methylation levels of the CpG sites as methylation states (i.e. hypermethylation, hypomethylation and no change) instead of continuous methylation values. Transition probabilities among the methylation states represent the distance distribution among the DMCs, whereas emission probabilities represent the likelihood of DM for the CpG sites. High transition probabilities and low transition probabilities are used to model the neighboring CpG sites that have high similarities and low similarities within their methylation levels, respectively. Parameters are estimated usually by using established learning algorithms, whereas potential DMRs are identified using different statistical approaches.

One of the approaches in this category named 'ComMet' [64], included in the Bisulfighter methylation analysis suite [78, 79], combines all the samples within a group into one sample and identifies the DMRs by comparing a pair of two samples. This method captures the probability distribution of distances between the neighboring DMCs and adjusts the DMC chaining criteria automatically for each data set. Transition probabilities are estimated using an expectation maximization algorithm, whereas emission probabilities are estimated from a beta-binomial mixture model. Parameters of the beta-binomial model are estimated by incorporating an unsupervised learning algorithm. DMRs are identified by using a dynamic programming algorithm.

One of the advantages of ComMet is that it does not require biological replicates to identify DMRs. It takes into account the sequencing coverage and the spatial distribution of the neighboring CpG sites. On the other hand, one of the limitations of this approach is that it does not take into account the biological variation across replicates, which might lead to higher number of false positives in the results [14, 43, 46].

Another approach in this category is 'HMM-Fisher' [80], which estimates the methylation status of the CpG sites for each sample instead of combining all the samples. Similar to ComMet, HMM-Fisher models both the similarity and dissimilarity of the methylation levels of the neighboring CpG sites using transition probability. HMM-Fisher estimates the transition probabilities using a Dirichlet distribution, whereas emission probabilities are computed using a truncated normal distribution. After estimating the methylation levels of all the CpG sites for each sample, differentially methylated CpG sites are identified using FET. Identified DMCs are further grouped into DMRs if the distance between the CpG sites is $<$100 bases. Non-consecutive CpG sites are reported as DMCs in the output.

One of the major contributions of HMM-Fisher is that it can identify DMRs of variable size, instead of depending on user-defined boundary thresholds. It takes the biological variation among the replicates into account and can provide both DMCs and DMRs as output. It can also be used to identify sample-wise methylation patterns.

'HMM-DM' [81] is another approach that uses HMM to identify DM. HMM-DM directly estimates the DM states of the CpG sites for each sample across the groups. In this approach, the transition probability of each CpG site only depends on the methylation state of the immediate previous CpG site. Like HMM-Fisher and ComMet, the transition probabilities are

estimated from a Dirichlet distribution. In contrast, emission probabilities are estimated from a beta distribution. DM states for the CpG sites are estimated using the MCMC method. Finally, consecutive CpG sites with same methylation status are grouped together based on user-defined thresholds to form DMRs. Similar to HMM-Fisher, HMM-DM can identify variable size DMRs from WGBS and RRBS data. It also takes into account the biological variation among the replicates.

In general, one of the key advantages of HMM-based approaches is that they can identify DMRs with variable size in contrast to the approaches that use a fixed window size. They consider the spatial correlation of the CpG sites by borrowing methylation information from their neighboring sites. These approaches can also identify independent DMCs or short DMRs; therefore, they can identify sharp methylation changes among the CpG sites. In addition, all the three approaches discussed above are applicable to both WGBS and RRBS data sets.

Entropy-based approaches

Entropy-based approaches identify the methylation difference across multiple samples using Shannon entropy [82], which is a quantitative measure of the variation or change in a series of events. Approaches in this category are capable of providing sample-specific methylation information.

'QDMR' [83] was the first approach that used Shannon entropy [82] for the purpose of identifying DMRs from bisulfite sequencing data. It quantitatively identifies DMRs from predefined regions based on the average methylation levels of the CpG sites of the regions. The probability that a sample is methylated at a specific location is calculated by taking the ratio of the methylation level of that sample and the total methylation level across all samples. The original entropy formula can be used to measure the methylation difference across samples, where lower entropy represents higher methylation difference. However, this way of calculating entropy is biased toward hypermethylation in minor samples. Therefore, QDMR introduces a one-step Tukey biweight weighted mean to make their approach less sensitive to such outliers. Finally, a region is differentially methylated if the weighted entropy for that region is smaller than a certain cutoff, which is determined by using a probability model. QDMR takes into account the biological variability across the samples. In addition to the list of DMRs, QDMR provides quantification, visualization and annotation of the DMRs for each sample. One of the limitations of this approach is that it can identify DMRs only from predefined regions (RRBS); therefore, it is unable to identify *de novo* regions.

An improved approach in this category, 'CpG_MPs' [51], has been proposed from the same research group, which can identify methylation patterns across paired or multiple samples using WGBS data. This approach identifies *de novo* methylated and unmethylated regions using hotspot extension algorithm based on the methylation status of the neighboring CpG sites. It combines a combinatorial algorithm with Shannon entropy to identify DMRs.

The overall workflow of CpG_MPs is divided into four modules. The first module normalizes the sequencing reads of the CpG sites into methylation levels. The second module categorizes the methylation states of the CpG sites based on their normalized methylation levels into four categories such as unmethylated CpGs, partially unmethylated CpGs, methylated CpGs and partially methylated CpGs. CpGs are then scanned from 5′ to 3′ end to extract a certain number of methylated (unmethylated) CpGs to create methylated (unmethylated) hotspots. Next, the hotspots are extended both upstream and downstream to incorporate partially methylated or partially unmethylated CpGs into their corresponding hotspots. Neighboring regions with the same patterns are then combined based on a given threshold. Also, the mean value and the standard deviation of the methylation levels of the CpG sites within each region are computed. The third module identifies conservatively unmethylated regions, conservatively methylated regions and DMRs by using a combinatorial algorithm with Shannon entropy. At first, the identified methylated and unmethylated regions are mapped to the reference genome and then overlapping regions (ORs) are recorded in the reference genome. Next, the hotspot extension technique is used to merge the neighboring ORs with the same methylation patterns across multiple samples. A modified Shannon entropy-based method is used to identify the regions that are significant across multiple samples. The fourth module analyzes sequencing features and visualizes the identified regions.

One key advantage of CpG_MPs is that it determines the DMR boundaries by applying combinatorial algorithm instead of depending on empirical thresholds to identify DMRs; hence, it can detect variable-length boundaries. It can also be used to identify methylation patterns for each sample. In addition, CpG_MPs considers biological variation among the replicates. However, CpG_MPs does not include any error control measurement among the identified regions.

A more recent approach, 'SMART' [84], extends the weighted entropy concept introduced by QDMR to determine cell type-specific methylation patterns from a large number of DNA methylomes. The input of SMART is the sample-wise methylation status of the CpG sites. SMART first quantifies the methylation specificity across the samples using Shannon entropy with a one-step Tukey biweight weighted mean. Next, it incorporates methylation similarities between neighboring CpG sites by estimating the methylation level of the sites based on Euclidean distance. These similarity metrics and methylation specificity states are then used to segment the genome into groups of CpG sites. Finally, a group of CpG sites is called hypermethylated (hypomethylated) if the methylation levels of that group is significantly higher (lower) than the average methylation levels of all samples determined by one sample *t*-test.

Major contribution of SMART is that it can identify cell type-specific methylation marks (i.e. HyperMark and HypoMark) from a large sample cohort. Instead of depending on user-defined thresholds, it determines DMR boundaries of variable sizes by quantifying the methylation levels of the CpG sites. It also provides functional annotation of the identified methylation marks. It considers the biological variation among the replicates and spatial correlation among the methylation levels of the CpG sites across the genome. In addition, it can be applied to both WGBS and RRBS data.

One of the key benefits of the entropy-based approaches is that they can directly identify DMRs without identifying DMCs. As a result, entropy-based approaches that can detect *de novo* regions (i.e. CpG_MPs and SMART) do not depend on empirical boundary estimations. Furthermore, these approaches take into account the biological variation within replicates.

Mixed statistical tests-based approaches

Approaches in this category rely on established statistical tests, such as FET, *t*-test and ANOVA, to identify DMCs/DMRs. These statistical tests are applied to CpG sites across the samples or

within predefined genomic regions (i.e. fixed/variable size windows).

One of the approaches in this category, 'COHCAP' [46], identifies differentially methylated CpG islands from two or more groups using predefined regions. It also provides integration with gene expression data and visualization of the results. The pipeline starts with taking aligned read counts (e.g. output of Bismark aligner [26]) as input. CpG sites are marked as methylated or unmethylated based on a user-defined threshold. *P*-values of the CpG sites are first calculated by using different statistical approaches (i.e. FET, ANOVA and *t*-test) based on the chosen experimental design. Later the *P*-values are corrected using the FDR approach. CpG sites are filtered based on *P*-value of the CpG site, average methylation proportion across all the samples and FDR value. CpG islands with a minimum number of filtered CpG sites are considered as candidate DMRs. In the 'average by CpG site' pipeline, *P*-values of the CpG sites within candidate DMRs are calculated by the previously selected statistical method. In the 'average by CpG island' pipeline, beta values of the filtered CpG sites within each candidate DMR are averaged, and then a *P*-value is calculated based on the averaged beta value. The major contribution of COHCAP is that it provides integration of gene expression data with DM analysis. In addition, it takes into account the biological variation among the replicates.

'DMAP' [85], another approach in this category, is a fragment-based approach primarily designed for the RRBS protocol to identify differentially methylated fragments (DMFs). Nonetheless, this approach can also detect DMRs from WGBS data. In addition to the identification of DMRs/DMFs, DMAP provides information about nearby genes and CpG sites.

The input of DMAP is methylated read counts in Bismark aligner [26] format. To identify candidate genomic regions from WGBS data, DMAP defines fixed-size windows (i.e. default 1000 bp). For RRBS data, it defines fragments of variable sizes (40–220 bp). Next, a *P*-value is calculated for each region or fragment based on the methylated CpG counts using a chosen statistical test (χ^2 test, FET and ANOVA). FET is recommended for pairwise comparison, χ^2 test is recommended for testing variability across multiple samples and ANOVA is recommended for comparing groups of samples. Candidate regions are selected as DMRs (for WGBS data) and DMFs (for RRBS data) based on a user-defined *P*-value threshold. Options to correct for multiple comparisons are also provided. The output is a list of candidate regions/fragments with their *P*-values and information regarding the statistical test that was applied. Furthermore, DMAP provides gene annotation features of the identified regions/fragments. Major contribution of this approach is that it can detect variable-size fragments (DMFs) from predefined regions.

'swDMR' [86], another approach in this category, integrates multiple commonly used statistical approaches to identify DMRs from WGBS data. The pipeline begins with taking the methylated read counts of each CpG site (preferably from the Bismark aligner [26]) as input, which are later converted to methylation ratios. Next, it divides the genome into multiple overlapping fragments or windows of equal length based on user-defined thresholds. A statistical approach is chosen from a list of commonly used approaches (i.e. FET, *t*-test, χ^2, Wilcoxon, ANOVA and Kruskal–Wallis test) to perform hypothesis testing within each window across two or more samples. For two samples, methylation levels of the CpG sites are compared using *t*-test, Wilcoxon test, χ^2 test or FET. For more than two samples, methylation levels are compared using either ANOVA or Kruskal–Wallis test. Therefore, for each window, swDMR provides a *P*-value generated using the selected statistical test. The resulting *P*-values are corrected for multiple comparisons using the FDR approach. The regions with corrected *P*-values lower than a predefined threshold are selected as potential DMRs. Using an extension function, two potential DMRs are merged if the distance between them is less than a predefined threshold. The merged DMRs are tested with the previously selected statistical test, and *P*-values are corrected with respect to the new DMR boundaries. Finally, the merged DMRs with the corrected *P*-values less than the user-defined threshold are selected as candidate DMRs. swDMR approach can be used without biological replicates and can work with CHG or CHH methylation. It also provides functionalities such as DMR cluster analysis, visualization and annotation of DMRs.

The key advantage of the approaches in this category is that they provide flexibility in selecting different statistical tests, and methods for multiple test correction. In contrast, these approaches do not take into account the spatial correlation between the methylation levels of the neighboring CpG sites. In addition, these approaches either work on predefined regions or divide the genome into windows of fixed/variable size. Hence, they miss the low CpG density regions where methylation has sharp changes such as TFBS that can contain a single differentially methylated CpG site [68]. Importantly, they depend on user-defined thresholds to estimate the DMR boundaries.

Binary segmentation-based approaches

Approaches in this category use binary segmentation algorithm to recursively divide the genome to identify candidate regions from bisulfite sequencing data. The only approach in this category, 'metilene' [87], uses a circular binary segmentation algorithm to identify DMRs. It can be used to analyze both WGBS and RRBS experiments across multiple samples with or without replicates.

The pipeline starts with a pre-segmentation step that divides the genome into primary regions based on the available methylation information. The pre-segmented regions are then iteratively segmented using a circular binary segmentation algorithm to identify a window with the maximum mean difference signal. The segmentation is terminated when a segment has less number of CpGs than a predefined threshold, or it does not show any improvement in the two-dimensional Kolmogorov–Smirnov test results. The identified window is marked as a potential DMR. The output of metilene is a list of DMRs with their *P*-values, adjusted *P*-values and the *P*-value from a Mann–Whitney U test.

Metilene can detect *de novo* regions of various lengths without relying on user-defined boundary thresholds. It takes into account the variation among biological replicates. In addition, it can predict methylation levels of the missing CpG sites using beta distribution. One of the limitations of metilene is that the result greatly depends on the minimum segment size parameter, which can lead to false negatives (if it is too high) or false positives (if it is too low). In addition, it does not consider the spatial correlation of the methylation levels of the CpG sites across biological replicates.

Discussion

In this survey, we briefly summarize 22 approaches that identify DM using bisulfite sequencing data focusing on their important features, such as concept used, protocol used, biological variability, spatial distribution, additional covariates, error correction, sequencing coverage and identifying *de novo* regions. The

approaches are categorized into seven different categories based on their primary concepts or techniques used to identify DM. Some of the approaches involve multiple concepts to identify DM; hence, they could be assigned to multiple categories. On such cases, we categorize the approach based on the concept that the authors highlighted. Pros and cons of these categories are summarized in Figure 3. The important features of the approaches covered in this survey are summarized in Table 1. Moreover, the workflow of the approaches, including the information about genome segmentation, difference quantification and DMR calling, are described in Figure 4.

Note that there are other possible ways to categorize these approaches. For instance, this can be done based on the data type used to estimate the methylation levels of the CpG sites (count data, ratio data and both count and ratio data). In that case, the methods will be distributed among the categories as follows: (i) count data: MethylKit, eDMR, DSS, DSS-single, DSS-general, MOABS, RADmeth, MethylSig, MACAU, GetisDMR, ComMet; (ii) ratio data: BSmooth, BiSeq, qDMR, CpG_MPs, SMART, HMM-Fisher, HMM-DM, COHCAP, metilene; (iii) both count and ratio data: DMAP, swDMR. A graphical representation of this classification is shown in Figure 5. Similarly, the approaches can be categorized based on the number of groups allowed (one group of samples, two groups without replicates and two groups with replicates), based on the protocol used (WGBS, RRBS and both WGBS and RRBS), etc.

Biological variability within the replicates is a crucial factor to consider because it can reduce the number of false positives in the results [14, 43, 46]. If an approach takes into account each biological replicate within a group separately when modeling the methylation levels of the CpG sites, then biological variability is considered. On the other hand, biological variability is lost if an approach combines the read counts of the CpG sites across the replicates. Although classical hypothesis testing methods (e.g. *t*-test and ANOVA) take biological variation into account, BSmooth was the first approach primarily developed for DMR identification that takes into account the biological variation among replicates. Within the surveyed approaches, smoothing-based approaches, beta-binomial-based approaches, entropy-based approaches, etc. (see Table 1 for full list) take the biological variation among the replicates into account.

Spatial correlation is another factor to consider, which provides a better estimation of the methylation levels of the CpG sites by borrowing information from their neighbors. A common way of considering spatial correlation is to perform 'smoothing' operation before the detection of DM. In this survey, smoothing-based approaches (BSmooth and BiSeq) and a few beta-binomial-based approaches (DSS-single, MACAU and GetisDMR) fall into this category. Performing smoothing when identifying DMRs can reduce the required sequencing depth and estimate the methylation status of missing CpG sites [43]. Additionally, smoothing procedure helps to identify relatively longer DMRs. However, this procedure is only applicable for the genome whose methylation profile is known to be smooth. Also smoothing is not suitable for the data sets whose CpG sites are sparse (commonly seen in RRBS protocol) due to extrapolated methylation values of 0 and 1. Besides smoothing, other techniques can be applied to take spatial correlation into account. For instance,

LOGISTIC REGRESSION BASED APPROACHES	PROS: (i) Consider additional covariates; (ii) Provide error correction for multiple tests; (iii) Annotate the DMCs or DMRs. CONS: (i) Do not consider biological variation.
SMOOTHING BASED APPROACHES	PROS: (i) Consider biological variation; (ii) Consider spatial correlation; (iii) Can reduce required sequencing depth; (iv) Can estimate methylation level of missing CpGs. CONS: (i) Can not detect sharp methylation changes.
BETA-BINOMIAL BASED APPROACHES	PROS: (i) Consider biological variation; (ii) Some consider additional covariates; (iii) Some provide error correction for multiple tests; (iv) Mostly can detect *de novo* regions; (v) Can detect sharp methylation changes; (vi) Consider sequencing coverage. CONS: (i) Usually do not consider spatial correlation.
HMM BASED APPROACHES	PROS: (i) Can detect variable size DMRs; (ii) Consider spatial correlation; (iii) Usually consider biological variation. CONS: (i) Do not consider additional covariates; (ii) Do not provide error correction for multiple tests; (iii) Usually do not consider sequencing coverage.
ENTROPY BASED APPROACHES	PROS: (i) Directly identify DMRs; (ii) Can identify sample specific methylation patterns/DMRs; (iii) Consider biological variation. CONS: (i) Do not identify DMCs; (ii) Do not consider sequencing coverage.
MIXED STATISTICAL BASED APPROACHES	PROS: (i) Mostly can detect DMR boundaries; (ii) Provide error correction for multiple tests; (iii) Provide flexibility to chose different statistical approaches. CONS: (i) Usually do not consider biological variation; (ii) Do not consider spatial correlation; (iii) Do not consider additional covariates.
BINARY SEGMENTATION BASED APPROACHES	PROS: (i) Considers biological variation; (ii) Provides error correction for multiple tests; (iii) Can detect variable size DMRs. CONS: (i) Does not identify DMCs; (ii) Does not consider spatial correlation; (iii) Does not consider additional covariates.

Figure 3. Pros and cons of the seven categories discussed in this survey.

Table 1. Summary of the important characteristics of the 22 surveyed approaches

	Method and reference	Concept used	Protocol	Primary purpose	Biological variation	Spatial distribution	Additional covariates	Error correction	Sequencing coverage	Identify *de novo* region	Total citations	Citation/ year
1	methylKit [54]	Logistic regression	Both	Identify DMCs and annotate	✗	✗	✓	✓	✓	✓	175	43.75
2	eDMR [64]	Logistic regression	Both	Identify DMCs and DMRs	✗	✓	✓	✓	✓	✓	28	8
3	BSmooth [43]	Smoothing	WGBS	Identify DMRs with replicates	✓	✓	✗	✗	✗	✓	156	39
4	BiSeq [66]	Smoothing	RRBS	Identify DMRs with FDR correction	✓	✓	✓	✓	✓	✗	62	18
6	DSS [69]	Beta-binomial	Both	Identify DMLs for small samples	✓	✗	✗	✗	✓	✓	43	16.1
5	MOABS [70]	Beta-binomial	Both	Identify DMCs with replicates	✓	✗	✗	✗	✓	✓	49	18.4
7	RADMeth [71]	Beta-binomial	WGBS	Identify DMLs and DMRs	✓	✗	✓	✗	✓	✓	31	13.3
8	methylSig [72]	Beta-binomial	Both	Identify DMCs and DMRs	✓	✗	✗	✗	✓	✓	42	17.4
9	DSS-single [73]	Beta-binomial	Both	Identify DMRs without replicates	✓	✓	✗	✓	✓	✓	15	12
10	MACAU [74]	Beta-binomial	Both	Identify DM using population structure	✓	✓	✓	✓	✓	✓	8	8
11	DSS-general [75]	Beta-binomial	RRBS	Identify DMLs	✓	✗	✓	✓	✓	✗	3	3
12	GetisDMR [76]	Beta-binomial	WGBS	Identify DMRs directly	✓	✓	✓	✓	✓	✓	0	0
13	ComMet [78]	HMM	Both	Identify DMRs	✗	✓	✗	✗	✓	✓	24	8.7
14	HMM-Fisher [80]	HMM	Both	Identify DM patterns	✓	✓	✗	✗	✗	✓	4	4
15	HMM-DM [81]	HMM	Both	Identify DMRs	✓	✓	✗	✗	✗	✓	4	4
16	QDMR [83]	Shannon entropy	RRBS	Identify DMRs	✓	✗	⊘	⊘	✗	✗	61	10.7
17	CpG_MPs [51]	Shannon entropy	WGBS	Identify DM patterns	✓	✓	⊘	⊘	✗	✓	30	7.2
18	SMART [84]	Shannon entropy	WGBS	Identify cell type-specific methylation marks	✓	✓	⊘	⊘	✗	✓	9	9
19	COHCAP [46]	Mixed statistics	RRBS	Identify DMCs and consistent CpG islands	✓	✗	✗	✓	✗	✗	27	7.7
20	DMAP [85]	Mixed statistics	Both	Identify DMRs and DMFs	✓	✗	✗	✓	✓*	✓	31	12.4
21	swDMR [86]	Mixed statistics	WGBS	Identify DMRs without replicates	✗	✗	✗	✓	✓*	✓	4	3.2
22	metilene [87]	Binary segmentation	Both	Identify DMRs in large groups of samples	✓	✗	✗	✓	✗	✓	0	0

For columns 5–10, ✓ means that the method considers the characteristic and ✗ means that the method does not consider the characteristic. For the 9th column, ✓* means that the method considers sequencing coverage when count-based hypothesis tests are performed. For the 10th column, identify *de novo* regions, ✓ means that the method can and ✗ means that the method cannot identify *de novo* regions. For columns 5–10, ⊘ means the characteristic is not applicable. Total citations and citations per year represent the number of citations and the average number of citations per year, respectively, as shown on google scholar as of 24 October 2016.

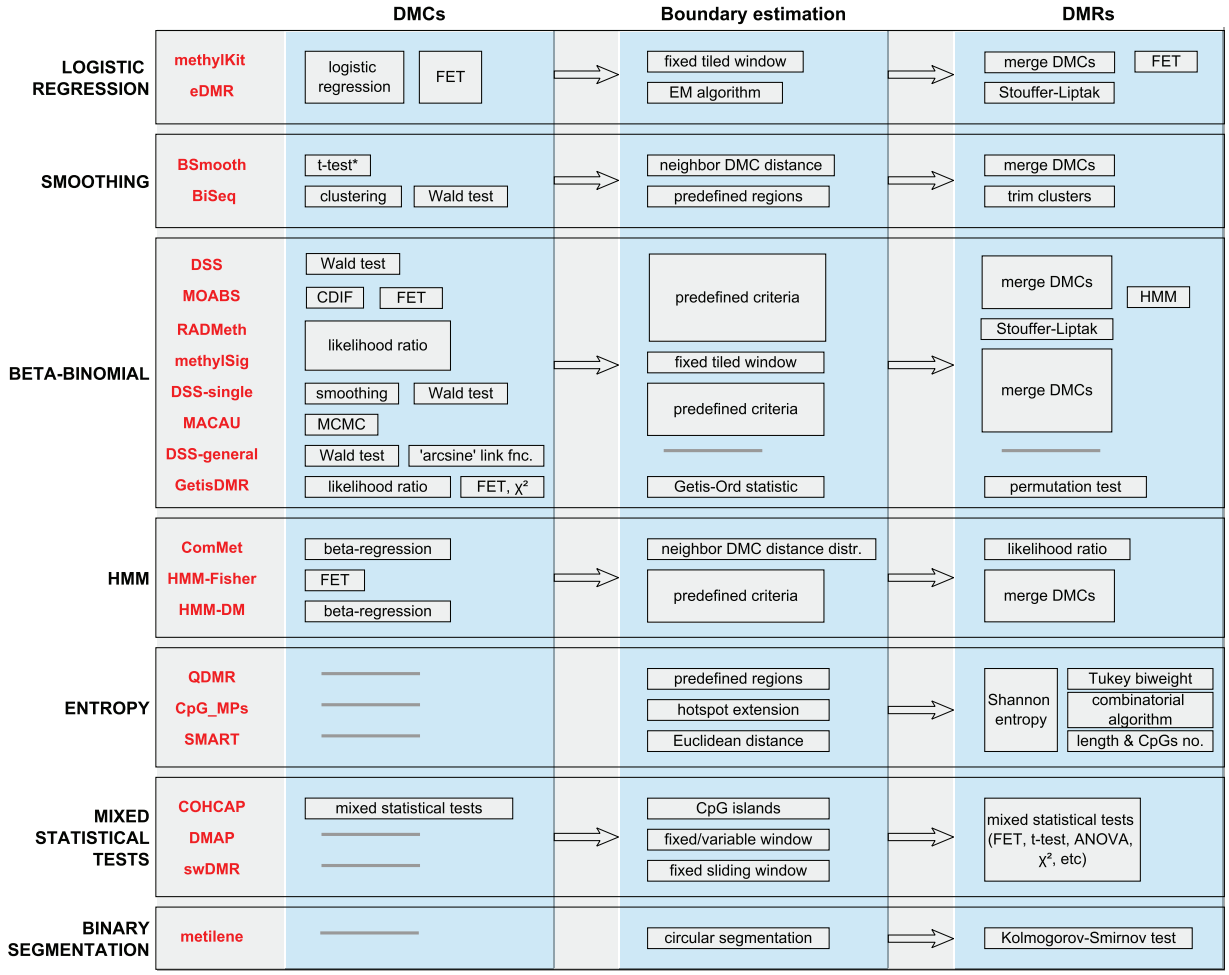

Figure 4. The workflow of 22 approaches developed for DM analysis. *t*-test* denotes a signal-to-noise statistic similar to the classical *t*-test. Predefined criteria represent user-defined thresholds such as *P*-value cutoff of the DMCs, length of the DMRs, distance between neighbor DMRs, minimum number of DMCs per DMR, cutoff value of CDIF (only for MOABS), etc. FET denotes Fisher's exact test, HMM denotes hidden Markov model, MCMC denotes Markov Chain Monte Carlo and CDIF denotes credible methylation difference.

eDMR uses autocorrelation of the methylation data, HMM-based approaches (ComMet, HMM-Fisher and HMM-DM) use HMM, CpG_MPs uses hotspot extension algorithm and SMART uses Euclidean distance based on methylation similarity to take into account spatial correlation of the CpG sites.

Sequencing coverage is another important factor that affects the accuracy of the methylation estimation. Count-based hypothesis tests (e.g. FET, χ^2 test) take into account sequencing coverage by simply pooling the read counts; however, these tests require grouping of read counts, and this is biased toward the samples with higher sequencing coverage. For other DM analysis approaches, consideration of coverage information is not merely dependent on the hypothesis tests but dependent on whether coverage information is incorporated when modeling the methylation levels of the CpG sites. For example, HMM-Fisher uses methylation ratios to estimate the methylation status at each CpG sites and then applies FET on the count of the methylation states to identify DMCs. Therefore, HMM-Fisher does not take into account read coverage despite using FET as the hypothesis test. Among the surveyed approaches, BiSeq, ComMet, DMAP, swDMR, logistic regression-based and beta-binomial-based approaches are able to take the coverage information into account. Some approaches also include

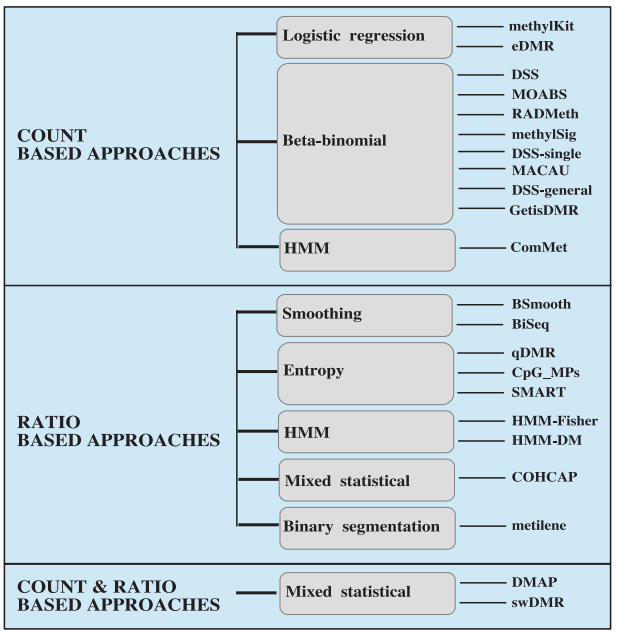

Figure 5. A higher level classification of the approaches discussed in this survey based on the data type used when modeling the methylation levels of the CpG sites.

Table 2. Comparison of the available implementations of the 22 surveyed approaches

	Method (tool) and tool reference	Platform	Availability	License	Output	Published date	Updated date
1	methylKit [54]	Biconductor R package	Standalone	Artistic v2	DMCs/DMRs list (table) DMCs/DMRs per chromosome (graph)	9 November 2011	22 October 2016
2	eDMR [54, 64]	R package	Standalone	Artistic/GPL	DMRs list (table)	4 January 2013	4 April 2014
3	BSmooth (bsseq) [43]	Biconductor R package	Standalone	Artistic v2	DMRs list (table) DMR/locus methylation level (graph)	20 July 2012	14 October 2016
4	BiSeq [88]	Biconductor R package	Standalone	LGPL v3	DMRs list (table) DMR mean methylation (graph)	2 April 2013	17 October 2016
6	DSS [69, 73, 75, 89]	Biconductor R package	Standalone	GNU GPL	DMCs/DMRs list (table) DMR % methylation/locus (graph)	04 June 2012	17 October 2016
5	MOABS [70]	C ++ package and Perl script	Standalone	GNU GPL v3	DMRs list (table)	12 June 2013	30 May 2015
7	RADMeth [71]	C ++ package	Standalone	GNU GPL v3	DMCs/DMRs list (table)	27 March 2014	1 May 2014[a]
8	methylSig [72]	R package	Standalone	GNU GPL v3	DMCs/DMRs list (table) CpG sites methylation rate (graph)	17 June 2014	10 June 2016
9	DSS-single (DSS) [69, 73, 75, 89]	Biconductor R package	Standalone	GNU GPL	DMCs/DMRs list (table) DMR % methylation/locus (graph)	16 April 2015	17 October 2016
10	MACAU [74]	C ++ package and R script	Standalone	GNU GPL	DMCs list (table)	5 June 2015	9 December 2015
11	DSS-general (DSS) [69, 73, 75, 89]	Biconductor R package	Standalone	GNU GPL	DMCs/DMRs list (table) DMR % methylation/locus (graph)	29 April 2015	17 October 2016
12	GetisDMR [76]	C ++ package and R scripts	Standalone	GNU GPL	DMRs list (table)	28 April 2016	28 September 2016
13	ComMet (Bisulfighter) [78]	C ++ package and Python	Standalone	CCANS	DMRs list (table)	12 December 2014	29 September 2015
14	HMM-Fisher [80]	R scripts	Standalone	None	DMRs list (table) DMR/locus methylation level (graph)	25 April 2014	29 February 2016
15	HMM-DM [81]	R scripts	Standalone	None	DMRs list (table) DMR/locus methylation level (graph)	27 March 2014	24 March 2016
16	QDMR [83]	Java package	Standalone web, CLI	Custom[b]	DMRs list (table) DMR in UCSC Genome Browser (graph)	10 May 2010	17 October 2012
17	CpG_MPs [51]	Java package and Perl script	Standalone web, CLI	None	DMRs list (table)	20 June 2011	1 September 2015
18	SMART (SMART-BS-Seq) [84]	Python package	Standalone	PSFL	DMRs list (table)	17 May 2015	17 May 2015
19	COHCAP [46]	Biconductor R package	Standalone	GNU GPL v3	DMCs and DM CpG islands list (table) DM CpG islands methylation average (graph)	9 January 2014	17 October 2016
20	DMAP (meth_progs_dist) [85]	C package	Standalone	None	DMRs list (table)	14 May 2013	28 August 2016
21	swDMR [86]	Perl and R scripts	Standalone	GNU GPL v3	DMRs list (table) DMR methylation level (graph)	6 January 2013	15 June 2014
22	metilene [87]	C package	Standalone	GNU GPL v2	DMRs list (table)	8 May 2015	29 April 2016

[a]RADMeth is now part of the MethPipe tool released on 6 September 2013 with the latest update on 21 October 2016.
[b]Custom license stating that the software is free of charge to researchers working at academic, non-profit organizations on non-commercial projects.
GNU, general public license; LGPL, lesser general public license; CCANS, creative commons attribution-NonCommercial-ShareAlike 3.0 unported license; PSFL, python software foundation license; CLI, command line interface.

additional filters to remove low coverage CpG sites before estimating methylation.

Identifying *de novo* regions is another important feature of the approaches that identify DM. Approaches that identify *de novo* regions use various techniques such as merging DMCs using empirical thresholds, entropy-based algorithms and binary segmentation to estimate DMR boundaries (see Figure 4). While empirical thresholds allow for more flexibility to the users, proper tuning of these parameters is necessary to get robust results. Some of the approaches, in addition to the list of DMRs, provide information such as the list of DMCs, genetic annotations and visualization of the DMRs.

Error control is another important factor in DM analysis, as it reduces the number of false positives in the results. Approaches control errors by correcting *P*-values for each CpG site across the genome, correcting *P*-values for each region, correcting the *P*-values within the identified regions, etc.

Identification of the fittest approach, among all that are available, is a challenging task in DM analysis. If biological replicates are available, beta-binomial approaches are suitable because they take both coverage information and biological variability among the replicates into account. In addition, they can identify low CpG density regions where methylation has sharp changes (e.g. TFBS). Within the beta-binomial-based approaches, DSS-single, MACAU and GetisDMR take spatial correlation into account. Therefore, these three approaches are more appropriate if the methylation levels of the CpG sites are known to be spatially correlated and biological replicates are available. Smoothing-based approaches, entropy-based approaches, HMM-Fisher, HMM-DM and metilene can also be applied when biological replicates are available. Similarly, if the methylation levels of the CpG sites are known to be spatially correlated, approaches that take spatial distribution into consideration, such as smoothing-based approaches, HMM-based approaches, DSS-single, MACAU, GetisDMR, CpG_MPs and SMART, should be used.

When sample size is small in the data set, DSS, MethylSig and HMM-Fisher are appropriate. While DSS uses information from all CpG sites and an empirical Bayes estimate to achieve variation shrinkage, methylSig uses local information and a maximum likelihood estimator to compute both the methylation level and the variance. HMM-Fisher, on the other hand, combines two CpG sites while conducting FET if the distance between them is $<$100 bases. If multiple experimental factors are available in the data set, approaches such as methylKit, eDMR, BiSeq, RADMeth, MACAU, DSS-general and GetisDMR are more appropriate because they allow additional covariates in their model.

Suitable approaches can also be chosen based on their primary purposes. For example, QDMR, CpG_MPs or HMM-Fisher can be used to identify methylation patterns from a single sample. To identify cell type-specific methylation marks from large sample cohorts, SMART is a suitable choice. To identify DM patterns (hypermethylation and hypomethylation) across two groups of samples, HMM-Fisher and HMM-DM are more appropriate. Approaches can be chosen based on the input data type as well. For instance, if the data protocol is RRBS, and the purpose is to identify DMRs, then QDMR, BiSeq, DSS-general or COHCAP can be applied. To work with CHG or CHH methylation, methylKit, eDMR, MOABS, DSS, RADMeth and swDMR are recommended because they are not limited to CpG methylation.

Comparison of some of the approaches can be found from two existing review papers, Klein et al. [15] and Yu and Sun. [16]. Klein et al. compared four tools that are originally developed for DM analysis: BiSeq [88], COHCAP [46], methylKit [54] and RADMeth [71]. This review evaluates the trade-off between the sensitivity and specificity for individual methods using the receiver operator characteristic (ROC) based on the regional *P*-values of the identified regions. The performance of each method is then assessed by computing and comparing the area under the ROC curve. According to this review, BiSeq and RADMeth outperform COHCAP and methylKit. Yu and Sun [16] compared BSmooth, methylKit, BiSeq, HMM-Fisher and HMM-DM. According to this review, HMM-Fisher and HMM-DM achieved higher sensitivity and specificity than the other three methods. To assess the performance of all of the available approaches, a benchmark analysis is needed. Due to the complex nature of the methylation data and lack of a gold standard for performance evaluation and standardized format of the input data, building a benchmark for assessing the efficiency of these approaches is a challenging task and out of the scope of this survey.

In addition to the conceptual overview, we also summarized the implementations of the approaches in Table 2. The summary includes platform information, license information, output format, published date and last update date. While this is a condensed view of the capabilities of these tools, it could still be expanded to include information such as consistency in the input and output formats. Such details as well as a simulated, noise-free data set with known results are further requirements toward creating a comprehensive benchmark for assessing the practical performance of DM detection tools.

Conclusion

Epigenetic modifications are thought to play a role in developmental disorders and cancer, are likely to be influenced by environmental factors and are known to regulate gene expression. Identification of DM using bisulfite sequencing data is a crucial step in the analysis of epigenetic data. Several statistical methods have been developed to address this challenge. In this study, we survey 22 methods that identify DM from bisulfite sequencing data. All the approaches surveyed in this article were developed within the past 5 years, which shows great interest for progress in this area. Our main objective in this survey is to provide the community a comprehensive view of the existing approaches that identify DM from bisulfite sequencing data. To do that, we classify the approaches into seven categories based on their primary concepts and features. We summarize the distinguishing characteristics, benefits and limitations of each approach and category. This survey is intended to help potential users to choose the best DM analysis method based on their requirements. It will help the researchers to design experiments to generate data that are better suited for the community. In addition, this survey will guide the developers to develop new efficient statistical models that identify DM by considering key characteristics described here.

Key points

- Identification of the fittest approach, among all that are available, is a challenging task in DM analysis.
- A comprehensive benchmark of the available approaches that identify DM is greatly needed.
- Due to the high computation cost, only a few web-based implementations of the approaches are currently available.

Funding

National Institutes of Health (RO1 DK089167, STTR R42GM087013); National Science Foundation (DBI-0965741); and Robert J. Sokol M.D. Endowment in Systems Biology (to S.D.). Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

References

1. Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Dev* 2011;**25**(10):1010–22.
2. Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet* 2007;**8**(4):286–98.
3. Lister R, Pelizzola M, Dowen RH, *et al*. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;**462**(7271):315–22.
4. Krueger F, Kreck B, Franke A, *et al*. DNA methylome analysis using short bisulfite sequencing data. *Nat Methods* 2012;**9**(2):145–51.
5. Feng S, Jacobsen SE, Reik W. Epigenetic reprogramming in plant and animal development. *Science* 2010;**330**(6004):622–7.
6. Lindroth AM, Cao X, Jackson JP, *et al*. Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science* 2001;**292**(5524):2077–80.
7. Breiling A, Lyko F. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics Chromatin* 2015;**8**(1):24.
8. Hendrich B, Bird A. Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol* 1998;**18**(11):6538–47.
9. Bird AP, Wolffe AP. Methylation-induced repression–belts, braces, and chromatin. *Cell* 1999;**99**(5):451–4.
10. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Rev Genet* 2012;**13**(7):484–92.
11. Harris RA, Wang T, Coarfa C, *et al*. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* 2010;**28**(10):1097–105.
12. Taiwo O, Wilson GA, Morris T, *et al*. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat Protoc* 2012;**7**(4):617–36.
13. Gu H, Bock C, Mikkelsen TS, *et al*. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat Methods* 2010;**7**(2):133–6.
14. Robinson MD, Kahraman A, Law CW, *et al*. Statistical methods for detecting differentially methylated loci and regions. *Front Genet* 2014;**5**:324.
15. Klein HU, Hebestreit K. An evaluation of methods to test predefined genomic regions for differential methylation in bisulfite sequencing data. *Brief Bioinform* 2016;**17**:769–807.
16. Yu X, Sun S. Comparing five statistical methods of differential methylation identifi- cation using bisulfite sequencing data. *Stat Appl Genet Mol Biol* 2016;**15**(2):173–91.
17. Sun Z, Cunningham J, Slager S, *et al*. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics* 2015;**7**(5):813–28.
18. Clark SJ, Statham A, Stirzaker C, *et al*. DNA methylation: bisulphite modification and analysis. *Nat Protoc* 2006;**1**(5):2353–64.
19. Meissner A, Gnirke A, Bell GW, *et al*. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* 2005;**33**(18):5868–77.
20. FASTX-Toolkit: FASTQ/A short-reads pre-processing tools. http://hannonlab.cshl.edu/fastx_toolkit/. 2010.
21. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011;**27**(6):863–4.
22. Cox MP, Peterson DA, Biggs PJ. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 2010;**11**(1):485.
23. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;**17**(1):10.
24. Bolger AM, Lohse M, Usadel B. Trimmomatic: a exible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**(15):2114–20.
25. Trim Galore! http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
26. Krueger F, Andrews SR. Bismark: a exible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* 2011;**27**(11):1571–2.
27. Chen PY, Cokus SJ, Pellegrini M. BS seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics* 2010;**11**(1):203.
28. Pedersen B, Hsieh TF, Ibarra C, *et al*. MethylCoder: software pipeline for bisulfitetreated sequences. *Bioinformatics* 2011;**27**(17):2435–6.
29. Harris EY, Ponts N, Levchuk A, *et al*. BRAT: bisulfite-treated reads analysis tool. *Bioinformatics* 2010;**26**(4):572–3.
30. Hong C, Clement NL, Clement S, *et al*. Probabilistic alignment leads to improved accuracy and read coverage for bisulfite sequencing data. *BMC Bioinformatics* 2013;**14**(1):337.
31. Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**(3):R25.
32. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**(4):357–9.
33. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* 2009;**10**:232.
34. Xi Y, Bock C, Müller F, *et al*. RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. *Bioinformatics* 2012;**28**(3):430–2.
35. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 2010;**26**(7):873–81.
36. Smith AD, Chung WY, Hodges E, *et al*. Updates to the RMAP short-read mapping software. *Bioinformatics* 2009;**25**(21):2841–2.
37. Bock C, Reither S, Mikeska T, *et al*. BiQ analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics* 2005;**21**(21):4067–8.
38. Kumaki Y, Oda M, Okano M. QUMA: quantification tool for methylation analysis. *Nucleic Acids Res* 2008;**36(Suppl 2)**:W170–5.
39. Sun S, Noviski A, Yu X. MethyQA: a pipeline for bisulfite-treated methylation sequencing quality assessment. *BMC Bioinformatics* 2013;**14**(1):259.
40. Hu K, Ting AH, Li J. BSPAT: a fast online tool for DNA methylation co-occurrence pattern analysis based on high-throughput bisulfite sequencing data. *BMC Bioinformatics* 2015;**16**(1):220.
41. Liao WW, Yen MR, Ju E, *et al*. MethGo: a comprehensive tool for analyzing wholegenome bisulfite sequencing data. *BMC Genomics* 2015;**16**(12):S11.
42. Eckhardt F, Lewin J, Cortese R, *et al*. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* 2006;**38**(12):1378–85.

43. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* 2012;**13**(10):R83.
44. Jaffe AE, Feinberg AP, Irizarry RA, *et al*. Significance analysis and statistical dissection of variably methylated regions. *Biostatistics* 2012;**13**(1):166–78.
45. Feinberg AP, Irizarry RA. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc Natl Acad Sci USA* 2010;**107(Suppl 1)**:1757–64.
46. Warden CD, Lee H, Tompkins JD, *et al*. COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. *Nucleic Acids Res* 2013;**41**(11):e117.
47. Cameron EE, Baylin SB, Herman JG. p15INK4B CpG island methylation in primary acute leukemia is heterogeneous and suggests density as a critical factor for transcriptional silencing. *Blood* 1999;**94**(7):2445–51.
48. Smallwood SA, Lee HJ, Angermueller C, *et al*. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* 2014;**11**(8):817–20.
49. Varley KE, Mutch DG, Edmonston TB, *et al*. Intra-tumor heterogeneity of MLH1 promoter methylation revealed by deep single molecule bisulfite sequencing. *Nucleic Acids Res* 2009;**37**(14):4603–12.
50. Singer ZS, Yong J, Tischler J, *et al*. Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Mol Cell* 2014;**55**(2):319–31.
51. Su J, Yan H, Wei Y, *et al*. CpG_MPs: identification of CpG methylation patterns of genomic regions from high-throughput bisulfite sequencing data. *Nucleic Acids Res* 2013;**41**(1):e4.
52. Bibikova M, Chudin E, Wu B, *et al*. Human embryonic stem cells have a unique epigenetic signature. *Genome Res* 2006;**16**(9):1075–83.
53. Byun HM, Siegmund KD, Pan F, *et al*. Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue-and individual-specific DNA methylation patterns. *Hum Mol Genet* 2009;**18**(24):4808–17.
54. Akalin A, Kormaksson M, Li S, *et al*. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* 2012;**13**(10):R87.
55. Hurlbert SH. Pseudoreplication and the design of ecological field experiments. *Ecol Monogr* 1984;**54**(2):187–211.
56. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 2013;**14**(1):91.
57. Tony Ng HK, Tang ML. Testing the equality of two Poisson means using the rate ratio. *Stat Med* 2005;**24**(6):955–65.
58. Gosset WS. The probable error of a mean. *Biometrika* 1908;**6**:1–25.
59. Pearson ES, Hartley HO. *Biometrika tables for statisticians* (vol. 2). Biometrika Trust, page 385, 1976.
60. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004;**3**(1):Article3.
61. Goeman JJ, Van De Geer SA, De Kort F, *et al*. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004;**20**(1):93–9.
62. Gelman A. Analysis of variance—why it is more important than ever. *Ann Stat* 2005;**33**(1):1–53.
63. Wang HQ, Tuominen LK, Tsai CJ. SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics* 2011;**27**(2):225–31.
64. Li S, Garrett-Bakelman FE, Akalin A, *et al*. An optimized algorithm for detecting and annotating regional differential methylation. *BMC Bioinformatics* 2013;**14(Suppl 5)**:S10.
65. Pedersen BS, Schwartz DA, Yang IV, *et al*. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics* 2012;**28**(22):2986–8.
66. Hebestreit K, Dugas M, Klein HU. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* 2013;**29**(13):1647–53.
67. Benjamini Y, Hochberg Y. Multiple hypotheses testing with weights. *Scand J Stat* 1997;**24**(3):407–18.
68. Rhee HS, Franklin Pugh B. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* 2011;**147**(6):1408–19.
69. Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res* 2014;**42**(8):e69.
70. Sun D, Xi Y, Rodriguez B, *et al*. MOABS: model based analysis of bisulfite sequencing data. *Genome Biol* 2014;**15**(2):R38.
71. Dolzhenko E, Smith AD. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics* 2014;**15**(1):215.
72. Park Y, Figueroa ME, Rozek LS, *et al*. MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics* 2014;**30**:2414–22.
73. Wu H, Xu T, Feng H, *et al*. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res* 2015;**43**(21):e141.
74. Lea AJ, Tung J, Zhou X. A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. *PLoS Genet* 2015;**11**(11):e1005650.
75. Park Y, Wu H. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics* 2016;**32**(10):1446–53.
76. Wen Y, Chen F, Zhang Q, *et al*. Detection of differentially methylated regions in whole genome bisulfite sequencing data using local Getis-Ord statistics. *Bioinformatics* 2016;**32**:3396–404.
77. Zaykin DV. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *J Evol Biol* 2011;**24**(8):1836–41.
78. Saito Y, Tsuji J, Mituyama T. Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions. *Nucleic Acids Res* 2014;e45.
79. Saito Y, Mituyama T. Detection of differentially methylated regions from bisulfite-seq data by hidden Markov models incorporating genome-wide methylation level distributions. *BMC Genomics* 2015;**16**(12):S3.
80. Sun S, Yu X. HMM-Fisher: identifying differential methylation using a hidden Markov model and Fisher's exact test. *Stat Appl Genet Mol Biol* 2016;**15**(1):55–67.
81. Yu X, Sun S. HMM-DM: identifying differentially methylated regions using a hidden Markov model. *Stat Appl Genet Mol Biol* 2016;**15**(1):69–81.
82. Shannon CE. A mathematical theory of communication. *ACM SIGMOBILE Mobile Comput Commun Rev* 2001;**5**(1):3–55.
83. Zhang Y, Liu H, Lv J, *et al*. QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res* 2011;**39**(9):e58.
84. Liu H, Liu X, Zhang S, *et al*. Systematic identification and annotation of human methylation marks based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific

hypomethylation in the regulation of cell identity genes. *Nucleic Acids Res* 2016;**44**(1):75–94.
85. Stockwell PA, Chatterjee A, Rodger EJ, *et al*. DMAP: differential methylation analysis package for RRBS and WGBS data. *Bioinformatics* 2014;**30**(13):1814–22.
86. Wang Z, Li X, Jiang Y, *et al*. swDMR: a sliding window approach to identify differentially methylated regions based on whole genome bisulfite sequencing. *PloS One* 2015;**10**(7):e0132866.
87. Jühling F, Kretzmer H, Bernhart SH, *et al*. metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res* 2016;**26**(2):256–62.
88. Hebestreit K, Klein HU. BiSeq: processing and analyzing bisulfite sequencing data. R package version 1.14.0. 2015.
89. Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* 2013;**14**(2):232–43.