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

Article in Briefings in Bioinformatics · March 2017
DOI: 10.1093/bib/bbx013

4 authors:

Adib Shafi
Wayne State University
2 PUBLICATIONS 0 CITATIONS

Cristina Mitrea
Wayne State University
9 PUBLICATIONS 57 CITATIONS

Tin Nguyen
Wayne State University
16 PUBLICATIONS 18 CITATIONS

Sorin Draghici
Wayne State University
225 PUBLICATIONS 7,279 CITATIONS

Some of the authors of this publication are also working on these related projects:

- disease subtype discovery based on integration of multiple types of omics data View project

All content following this page was uploaded by Adib Shafi on 24 April 2017.

The user has requested enhancement of the downloaded file.
A survey of the approaches for identifying differential methylation using bisulfite sequencing data

Adib Shafi, Cristina Mitrea, Tin Nguyen, and Sorin Draghici

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 (DM) 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 paper, 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.

Keywords

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), 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 very low compared to 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 \[8, 9, 10\]. 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 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 differential methylation, 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 is low. Klein et al. \[15\] evaluates nine approaches that can possibly be used for differential methylation analysis. However, the methods are limited to the scope of analyzing differential methylation in predefined regions using only RRBS data. Among the nine approaches, only four of them are originally designed for analyzing differential methylation. The other five approaches are general approaches that can be applied for RNA-Seq and gene expression data. Yu and Sun. \[16\] evaluates only five approaches developed for the purpose of identifying DMRs. Sun et al. \[17\] briefly summarizes the commonly used platforms for methylation profiling, data preprocessing techniques, and statistical approaches for differential methylation analysis. This review provides a well organized conceptual overview of approaches that identify differential methylation using bisulfite sequencing data. However, this survey only includes seven such approaches. In summary, all previous attempts of reviewing the approaches that identify differential methylation 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. In order to address these issues, a comprehensive survey of the approaches that identify differential methylation using bisulfite sequencing data is greatly needed.

In this paper, we review 22 different approaches for differential methylation analysis, including approaches for identifying DMCs, DMRs (both predefined and de novo regions), and methylation patterns using bisulfite sequencing data (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 differential methylation 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 a 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 \[40\].

Several technologies have been developed for measuring DNA methylation based on bisulfite sequencing conversion. The most comprehensive protocol among them is whole genome bisulfite sequencing (WGBS),
which provides genome-wide DNA profiling. However, the application of this protocol on the whole genome is very expensive when it comes to studying organisms with large genomes. More cost-effective protocols, such as reduced representation bisulfite sequencing (RRBS) and enhanced RRBS (ERRBS), have allowed for methylation analysis with reduced sequencing requirements through a more targeted approach for CpG rich genomic regions that meet specific length requirements [41]. 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 6 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) differential methylation 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 pre-processing

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 in order to increase mapping efficiency. Existing tools for quality control include FASTX-Toolkit [42], PRINSEQ [43], SolexaQA [44], Cutadapt [45], Trimmomatic [46], and Trim Galore! [47]. 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 in order to estimate the methylation levels. Simply aligning these reads by using standard aligners results in poor mapping efficiency since 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 into two categories: three-letter aligners and wildcard aligners. Three-letter aligners, such as Bismark [48], BS Seeker [49], MethylCoder [50], BRAT [51], and GNU-Map-bs [52], 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 standard genome aligners such as Bowtie [53], and Bowtie2 [54]. In contrast, wildcard aligners, such as BS-Map [55], RRBS-Map [56], GSNAP [57], and RMAD [58], 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 prior to differential methylation analysis. Several post-alignment analysis tools have been developed, including BiQ Analyzer \[59\], QUMA \[60\], BRAT \[51\], MethyQA \[61\], BSPAT \[62\], and MethGo \[63\]. 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).

Differential methylation analysis

After obtaining the methylation information of the CpG sites, typically the next downstream analysis is to perform differential methylation analysis, which is usually done in the form of identifying differentially methylated cytosines (DMCs) or differentially methylated regions (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].

In order 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][64]. Incorporating spatial correlation in DM analysis can reduce the required sequencing depth and can estimate the methylation status of the missing CpG sites [20]. **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 since 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 [65][66]. Ignoring biological variation while detecting DMRs might lead to a high number of false positives in the results [14][20][26]. 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 [67][68][69][70].

Classical hypothesis testing methods, such as Fisher’s exact test (FET), Chi-square ($\chi^2$) test, regression approaches, t-test, moderated t-test, Goeman’s global test, ANOVA, etc. can be used to identify differential methylation using bisulfite sequencing data [3][19][23][71][72]. 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. Fisher’s exact test (FET) is a classical statistical test used to determine if 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 [21]. Chi-square test ($\chi^2$) is another classical method to test the relationship between two categorical variables (methylated vs. unmethylated). In contrast with FET, it allows for testing across multiple samples. As pointed out by Sun et al. [17] and Hurlbert et al. [73], there are several issues related to the aggregation of read counts into a single number while applying tests of independence (FET and Chi-square test). First, the read counts are not independent; they represent different sets of inter-dependent 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 towards the samples with higher coverage. Third, by aggregating (summing) the counts, some of the biological variation (e.g., sample size, intra-group variance) is not taken into account by the hypothesis testing. Therefore, using FET and Chi-square test to compare two groups of samples could lead to a high number of false positives [14][20][26].

Regression approaches (e.g. Poission, 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 differential methylation analysis [15]. For example, the read counts can be modeled using a Poisson distribution and a modified Wald test can be used to detect differential methylation as the difference between two Poisson means [73][75].

**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 [76][77], moderated t-test (limma) [78] or Goeman’s global test [79] 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 [80] can be used instead of multiple pair-wise comparisons. Compared to count based hypothesis tests, the ratio based tests take into account the biological variation across multiple replicates. However, since they only take into account the ratio of the reads (methylated reads vs. 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, estimating methylation levels of missing CpG sites, etc. 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 in order to identify differential methylation. One of the popular approaches in this category is methylKit [21], 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 Fisher's exact test to identify differential methylation. P-values are corrected using the false discovery rate (FDR) approach or the sliding linear model approach [81]. 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 differential methylation, etc.

Another method named eDMR [22] 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 [82]. The p-values for DMRs are then corrected for multiple comparisons using the FDR method. eDMR provides 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 [20], which relies on smoothing across the genome within each sample. It looks for group differences via CpG-wise t-tests in order 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 [54]. 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 employed 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, minimum number of CpG sites, etc.

BSmooth was the first approach primarily developed for DMR identification that takes into account the biological variation among replicates. BSmooth reduces the required sequencing coverage by applying the local likelihood smoothing approach across the samples. It can identify de novo regions from WGBS datasets.
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) [24]. BSmooth is not applicable to those datasets 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) [24]. 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 in order 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 (WFDR) is applied to take into account the size of individual clusters [83]. 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 datasets.

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 [20]. Furthermore, they can estimate the methylation levels of missing CpG sites. On the other hand, smoothing based approaches can not detect the low CpG density regions where methylation has sharp changes such as transcription factor binding sites (TFBs). TFBs are usually very small (i.e., < 50 bp) which might consist of a single CpG that is differentially methylated [84]. 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 absence of any biological or technical variation, methylation proportion of a particular CpG site follows a binomial distribution since 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 in order to identify differential methylation, such as DSS [26], MOABS [27], RADMeth [30], methylSig [29], DSS-single [31], MACAU [33], DSS-general [36] and GetisDMR [39]. 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 differential methylation 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 employed to estimate the
dispersion parameter in order 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, minimum number of CpG sites, etc.

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 in order to define the DMR boundaries and select candidate DMRs.

An even more recent variation of DSS approach, named DSS-general, identifies differentially methylated loci from bisulfite sequencing data under general experiment design. DSS-general identifies differentially methylated loci 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 (GLS) 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 since it estimates the regression parameters in one iteration.

MOABS is another approach that relies on beta-binomial assumption to identify differential methylation. 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 differential methylation between two samples is represented by a single metric named credible methylation difference (CDIF), which incorporates both the biological and statistical significance of the differential methylation. MOABS can also work with CHG or CHH methylation.

RADMeth is another analysis pipeline that relies on the beta-binomial assumption. RADMeth employs 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 differential methylation 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 \[85\]) in order to obtain the differentially methylated regions. The key contribution of this approach is the ability to analyze WGBS data in multiple factor experiments.

MethylSig is another analysis pipeline that utilizes 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 employs 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 dataset. This model is a generalized beta-binomial model consisting of an extra term to model the population structure. In 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 binomial mixed model 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 dataset.

**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 differential methylation 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 Fisher’s exact test. 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 non-random association between a data point and it’s 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 non random 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 can not work with enriched regions, such as RRBS data.

Beta-binomial based approaches are useful since they take into account both sampling variability among the read counts and biological variability among the replicates. Furthermore, these approaches are able to identify differential methylation 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 (HMM) based approaches**

Approaches in this category employ 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 differential methylation 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** [22], included in the Bisulfighter methylation analysis suite [23, 86], 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 dataset. 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, 20, 26].

Another approach in this category is HMM-Fisher [37], 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 Fisher’s exact test. Identified DMCs are further grouped into DMRs if the distance between the CpG sites is less than 100 bases. Non-consecutive CpG sites are reported as DMCs in the output.

One of the major contribution 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 [38] is another approach that employs HMM to identify differential methylation. HMM-DM directly estimates the differential methylation 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. Differential methylation states for the CpG sites are estimated using the Markov Chain Monte Carlo (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 neighbor 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 datasets.

Entropy based approaches

Entropy based approaches identify the methylation difference across multiple samples using Shannon entropy [87], 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 [18] was the first approach that utilizes Shannon entropy [87] 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 towards 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 [19], have been proposed from the same research
group which can identify methylation patterns across paired or multiple samples using WGBS data. This approach identifies \textit{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 the regions that are differentially methylated (i.e. DMRs) or conservatively unmethylated (i.e. CUMRs).

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 in order 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 (CUMRs), conservatively methylated regions (CMRs) and differentially methylated regions (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 [35], 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 \textit{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 Fisher’s exact test, t-test, ANOVA, etc. 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 [23], 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 \[48\]) 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., Fisher’s exact test, 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 differential methylation analysis. In addition, it takes into account the biological variation among the replicates.

**DMAP** \[28\], 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 \[48\] format. In order to identify candidate genomic regions from WGBS data, DMAP defines fixed-size windows (i.e. default 1000bp). For RRBS data, it defines fragments of variable sizes (40-220bp). Next, a p-value is calculated for each region or fragment based on the methylated CpG counts using a chosen statistical test (Chi-square test, Fisher’s exact test and ANOVA). Fisher’s exact test is recommended for pairwise comparison, Chi-Squared 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** \[32\], another approach in this category, integrates multiple commonly used statistical approaches in order to identify DMRs from WGBS data. The pipeline begins with taking the methylated read counts of each CpG site (preferably from the Bismark aligner \[48\]) 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., Fisher’s exact test, t-test, Chi-square, 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, Chi-square or Fisher’s exact test. 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, etc.

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 transcription factor binding sites (TFBs) that can contain a single differentially methylated CpG site \[84\]. 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 in order to identify candidate regions from bisulfite sequencing data. The only approach in this category, metilene [34], utilizes 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 in order to identify a window with the maximum mean difference signal (MDS). 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.

Figure 3: Pros and cons of the seven categories used to identify differential methylation discussed in this survey.
Discussion

In this survey, we briefly summarize 22 approaches that identify differential methylation (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, identifying de novo regions, etc. 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). On that case, the methods will be distributed among the categories as follows: i) Count data: MethylKit, eDMR, DSS, DSS-single, DSS-general, MOABS, RADmethyl, MethylSig, MACAU, GetisDMR, ComMet; ii) Ratio data: BSsmooth, BiSeq, qDMR, CpG_MPs, SMART, HMM-Fisher.
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.

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 since it can reduce the number of false positives in the results [14, 20, 26]. 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 better estimation of the methylation levels of the CpG sites by borrowing information from the neighboring CpG sites. A popular way of considering spatial correlation is to perform “smoothing” operation prior to the detection of differential methylation. In this survey, smoothing based approaches (BSMOOTH, and BiSeq) and few beta-binomial based approaches (DSS-single, MACAU, and GetisDMR) fall into this category. By performing smoothing when identifying DMRs can reduce the required sequencing depth and estimate the methylation status of missing CpG sites [20]. 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 dataset whose CpG sites are sparse (commonly seen in RRBS protocol) due to extrapolated methylation values of 0 and 1. Besides smoothing, there are other techniques as well that can take spatial correlation into account. For instance, eDMR uses auto correlation of the methylation data,
HMM based approaches (ComMet, HMM-Fisher, and HMM-DM) use hidden Markov model, 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., Fisher’s exact test, Chi-square test) take into account sequencing coverage by simply pooling the read counts; however, these tests require grouping of read counts, and this is biased towards the samples with higher sequencing coverage. For other differential methylation 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 methylation status of CpG sites to identify DMR. 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 additional filters to remove low coverage CpG sites before estimating methylation.

Identifying de novo regions is another very 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, binary segmentation, etc. to estimate DMR boundaries (see Figure 4). While empirical thresholds allow for more flexibility to the users, proper tuning of these parameters is necessary in order 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, visualization of the DMRs, etc.

Error control is an important factor to consider 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. In our survey, we also provide information whether an approach allows additional covariates or not.

Identification of the fittest approach, among all that are available, is a challenging task in differential methylation analysis. If biological replicates are available, beta-binomial approaches are suitable since 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., transcription factor binding sites). Within the beta-binomial based approaches DSS-single, MACAU, and GetisDMR accounts 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 dataset, 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 employs 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 Fisher’s exact test if the distance between them is less than 100 bases. If multiple experimental factors are available in the dataset, approaches such as methylKit, eDMR, BiSeq, RADmeth, MACAU, DSS-general and GetisDMR are recommended since 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:
<table>
<thead>
<tr>
<th>Method &amp; Ref.</th>
<th>Concept used</th>
<th>Protocol</th>
<th>Primary purpose</th>
<th>Spatial var.</th>
<th>Addl. cov.</th>
<th>Error corr.</th>
<th>Seq. cov.</th>
<th>Identify de novo regions</th>
<th>Total citation</th>
<th>Cit./year</th>
</tr>
</thead>
<tbody>
<tr>
<td>methylKit</td>
<td>Logistic regression</td>
<td>Both</td>
<td>Identify DMCs and annotate</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>175</td>
<td>43.75</td>
</tr>
<tr>
<td>eDMR</td>
<td>Logistic regression</td>
<td>Both</td>
<td>Identify DMCs and DMRs</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>28</td>
<td>8</td>
</tr>
<tr>
<td>BSmooth</td>
<td>Smoothing</td>
<td>Both</td>
<td>Identify DMCs with replicates</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>156</td>
<td>39</td>
</tr>
<tr>
<td>DSS</td>
<td>Smoothing</td>
<td>Both</td>
<td>Identify DMCs with replicates</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>62</td>
<td>18</td>
</tr>
<tr>
<td>MOABS</td>
<td>Beta-binomial</td>
<td>Both</td>
<td>Identify DMCs with replicates</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>49</td>
<td>14.4</td>
</tr>
<tr>
<td>RADMeth</td>
<td>Beta-binomial</td>
<td>Both</td>
<td>Identify DMCs with replicates</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>31</td>
<td>11.3</td>
</tr>
<tr>
<td>methylSig</td>
<td>Beta-binomial</td>
<td>Both</td>
<td>Identify DMCs with replicates</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>17.7</td>
<td>17.7</td>
</tr>
<tr>
<td>DSS-single</td>
<td>Beta-binomial</td>
<td>Both</td>
<td>Identify DMCs without replicates</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>MACAU</td>
<td>Beta-binomial</td>
<td>Both</td>
<td>Identify DM with population structure</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>GetisDMR</td>
<td>Gamma distribution</td>
<td>Both</td>
<td>Identify DMCs directly</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>ComMet</td>
<td>Mixed statistics</td>
<td>Both</td>
<td>Identify DMCs directly</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>24</td>
<td>8.7</td>
</tr>
<tr>
<td>COHCAP</td>
<td>Mixed statistics</td>
<td>Both</td>
<td>Identify DMCs and consistent CpG islands</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>61</td>
<td>10.7</td>
</tr>
<tr>
<td>SMART</td>
<td>Mixed statistics</td>
<td>Both</td>
<td>Identify DMCs and consistent CpG islands</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>30</td>
<td>7.2</td>
</tr>
<tr>
<td>DMAP</td>
<td>Mixed statistics</td>
<td>Both</td>
<td>Identify DMCs and consistent CpG islands</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>swDMR</td>
<td>Mixed statistics</td>
<td>Both</td>
<td>Identify DMCs and consistent CpG islands</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>methylene</td>
<td>Binary segmentation</td>
<td>Both</td>
<td>Identify DMCs and consistent CpG islands</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>31</td>
<td>12.4</td>
</tr>
</tbody>
</table>

Table 1: Summary of the important characteristics of the 22 surveyed approaches. For columns 5–9, ⊘ means that the method considers the characteristic when count based hypothesis tests are performed. For the 9th column, identify de novo regions, ⊘ means that the method cannot identify de novo regions. For columns 5–9, ± means that the method cannot identify de novo regions. For the 10th column, total citations, and Citations per year represents the number of citations and the average number of citations per year, respectively, as shown on google scholar as of 10/24/2016. DM denotes differential methylation. DMRs, DMFs, DMCs and DMLs denote differentially methylated regions, differentially methylated fragments, differentially methylated cytosines and differentially methylated loci, respectively.
Table 2: Comparison of the available implementations of the 22 surveyed approaches. *RADMeth is now part of the MethPipe tool released on 09/06/2013 with the latest update on 10/21/2016. **CCANS: Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. ***Custom license stating that the software is free of charge to researchers working at academic, non-profit organizations on non-commercial projects. ****PSFL: Python Software Foundation License. CLI: command line interface.

<table>
<thead>
<tr>
<th>Method &amp; Tool ref.</th>
<th>Platform</th>
<th>Availability</th>
<th>License</th>
<th>Output</th>
<th>Published date</th>
<th>Updated date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 methylKit [21]</td>
<td>Biconductor R package</td>
<td>standalone</td>
<td>Artistic v2</td>
<td>DMCs/DMRs list (table) DMRs per chromosome (graph)</td>
<td>11/09/2011</td>
<td>10/22/2016</td>
</tr>
<tr>
<td>2 eDMR [21, 22]</td>
<td>R package</td>
<td>standalone</td>
<td>Artistic/GPL</td>
<td>DMRs list (table)</td>
<td>01/04/2013</td>
<td>04/04/2014</td>
</tr>
<tr>
<td>3 BSmooth (bsseq) [20]</td>
<td>Biconductor R package</td>
<td>standalone</td>
<td>Artistic v2</td>
<td>DMRs list (table) DMR /locus methylation level (graph)</td>
<td>07/20/2012</td>
<td>10/14/2016</td>
</tr>
<tr>
<td>4 BiSeq [38]</td>
<td>Biconductor R package</td>
<td>standalone</td>
<td>LGPL v3</td>
<td>DMRs list (table) DMR mean methylation (graph)</td>
<td>04/02/2013</td>
<td>10/17/2016</td>
</tr>
<tr>
<td>5 MOABS [27]</td>
<td>C++ package &amp; Perl script</td>
<td>standalone</td>
<td>GNU GPL v3</td>
<td>DMRs list (table)</td>
<td>06/12/2013</td>
<td>05/30/2015</td>
</tr>
<tr>
<td>7 RADMeth [30]</td>
<td>C++ package</td>
<td>standalone</td>
<td>GNU GPL v3</td>
<td>DMCs/DMRs list (table) DMR % methylation /locus (graph)</td>
<td>03/27/2014</td>
<td>05/01/2014*</td>
</tr>
<tr>
<td>8 methylISig [29]</td>
<td>R package</td>
<td>standalone</td>
<td>GNU GPL v3</td>
<td>DMRs list (table)</td>
<td>06/17/2014</td>
<td>06/10/2016</td>
</tr>
<tr>
<td>9 DSS-single (DSS) [89, 26, 31, 36]</td>
<td>Biconductor R package</td>
<td>standalone</td>
<td>GNU GPL</td>
<td>DMCs/DMRs list (table) DMR % methylation /locus (graph)</td>
<td>04/29/2015</td>
<td>10/17/2016</td>
</tr>
<tr>
<td>11 DSS-general (DSS) [89, 26, 31, 36]</td>
<td>Biconductor R package</td>
<td>standalone</td>
<td>GNU GPL</td>
<td>DMCs/DMRs list (table) DMR % methylation /locus (graph)</td>
<td>04/29/2015</td>
<td>10/17/2016</td>
</tr>
<tr>
<td>13 ComMet (Bisulfighter) [25]</td>
<td>C++ package &amp; Python</td>
<td>standalone</td>
<td>CCANS**</td>
<td>DMRs list (table)</td>
<td>12/12/2014</td>
<td>09/29/2015</td>
</tr>
<tr>
<td>14 HMM-Fisher [37]</td>
<td>R scripts</td>
<td>standalone</td>
<td>none</td>
<td>DMRs list (table) DMR /locus methylation level (graph)</td>
<td>04/25/2014</td>
<td>02/29/2016</td>
</tr>
<tr>
<td>15 HMM-DM [38]</td>
<td>R scripts</td>
<td>standalone</td>
<td>none</td>
<td>DMRs list (table) DMR /locus methylation level (graph)</td>
<td>03/27/2014</td>
<td>03/24/2016</td>
</tr>
<tr>
<td>16 QDMR [18]</td>
<td>Java package</td>
<td>standalone</td>
<td>custom***</td>
<td>DMRs list (table) DMR in UCSC Genome Browser (graph)</td>
<td>05/10/2010</td>
<td>10/17/2012</td>
</tr>
<tr>
<td>17 CpG-MPs [19]</td>
<td>Java package &amp; Perl script</td>
<td>standalone</td>
<td>none</td>
<td>DMRs list (table)</td>
<td>06/20/2011</td>
<td>09/01/2015</td>
</tr>
<tr>
<td>18 SMART (SMART-BS-Seq) [35]</td>
<td>Python package</td>
<td>standalone</td>
<td>PSFL****</td>
<td>DMRs list (table)</td>
<td>05/17/2015</td>
<td>05/17/2015</td>
</tr>
<tr>
<td>20 DMAP (meth progs dist) [28]</td>
<td>C package</td>
<td>standalone</td>
<td>none</td>
<td>DMRs list (table)</td>
<td>05/14/2013</td>
<td>08/28/2016</td>
</tr>
<tr>
<td>21 swDMR [32]</td>
<td>Perl &amp; R scripts</td>
<td>standalone</td>
<td>GNU GPL v3</td>
<td>DMRs list (table) DMR methylation level (graph)</td>
<td>01/06/2013</td>
<td>06/15/2014</td>
</tr>
<tr>
<td>22 metilene [34]</td>
<td>C package</td>
<td>standalone</td>
<td>GNU GPL v2</td>
<td>DMRs list (table)</td>
<td>05/08/2015</td>
<td>04/29/2016</td>
</tr>
</tbody>
</table>
BiSeq [88], COHCAP [23], methylKit [21], and RADMeth [30]. 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. In order to assess the performance of all of the available approaches, a benchmark analysis is needed. Due to the complex nature of the methylation data, the lack of a gold standard for performance evaluation, and the lack of 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, last update date, etc. 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 dataset with known results are further requirements towards 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 they are known to regulate gene expression. Reliable identification of differential methylation using bisulfite sequencing data is an important problem in the analysis of epigenetic data and statistical methods have employed to address the problem. In this study, we survey 22 methods that identify differential methylation from bisulfite sequencing data. All the approaches surveyed in this paper were developed within the last five 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 differential methylation 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 differential methylation analysis method based on their requirements. It will help the researchers to design experiments to generate data that is better suited for the community. In addition, this survey will guide the developers to develop new efficient statistical models that identify differential methylation 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.
Short description of the authors

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 student 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 (ISBL) 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 two best-selling books on data analysis of high throughput genomics data, 8 book chapters and over 160 peer-reviewed journal and conference papers.

Funding

This work was partially supported by the following grants: National Institutes of Health [RO1 DK089167, STTR R42GM087013]; National Science Foundation [DBI-0965741]; and by the Robert J. Sokol M.D. Endowment in Systems Biology (to SD). Any opinions, findings, and 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


[26] Hao Feng, Karen N Conneely, and Hao Wu. “A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data”. In: *Nucleic Acids Research* 42.8 (2014), e69.


