Journal of Advances and Scholarly Researches in Allied Education
Vol. 16, Issue No. 6.1, May-2019 (Special Issue), ISSN 2230-7540

Impact of Strategies for Analyzing Bisulfate Sequencing Data

Dr. S. Saravanan*

Assistant Professor, Department of Biotechnology, Dhanalakshmi Srinivasan College of Arts and Science for Women, Perambalur, Tamil Nadu

research@dscollege.ac.in

ABSTRACT

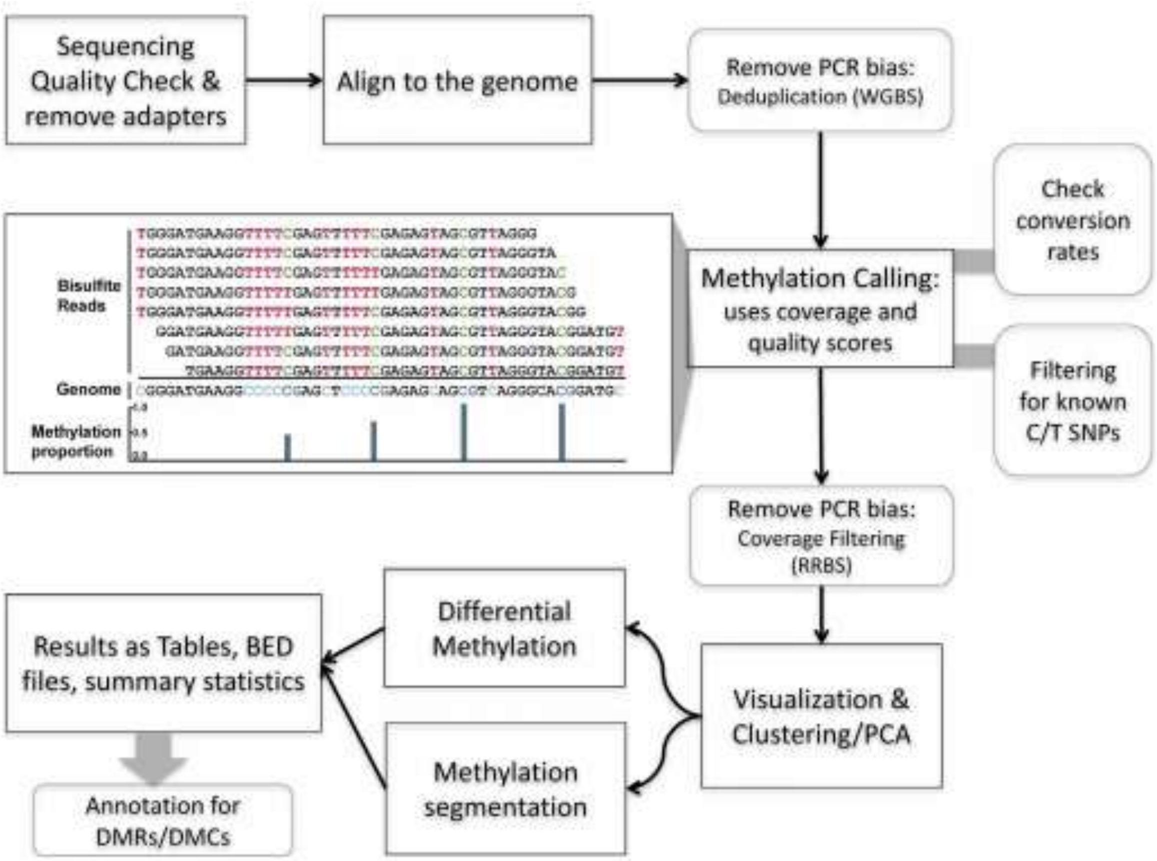
One of the key epigenetic modifications in the eukaryotic genome is DNA methylation; it has been shown to play a role in complex cell-type regulation of gene expression, and thus cell-type identity. The gold standard for calculating methylation across the genomes of interest is bisulfate sequencing. Here, for the study of high-throughput bisulfate sequencing, we examine many techniques used. To guarantee data accuracy, we implement advanced short-read alignment techniques as well as pre/post-alignment quality check methods. In addition, after alignment, we address subsequent review steps. We implement different methods of differential methylation and compare their output using datasets of simulated and actual bisulfate sequencing. We also address the techniques used to segment methylomes to classify regulatory regions. We implement methods of annotation that can be used for further classification by segmentation and differential methylation methods of regions returned. Finally, we analyses software packages that incorporate techniques to deal effectively locally with large bisulfate sequencing datasets and address workflows for online research that do not require any previous programming skills. The analysis techniques outlined in this review will direct researchers to the best bisulfate sequencing analysis practices at any stage.

***Keywords** – Strategies, Analyzing, Bisulfate*

INTRODUCTION

One of the key covalent base modifications in eukaryotic genomes is cytosine methylation (5-methylcytosine, 5mC). In a cell-type specific manner, it is involved in epigenetic control of gene expression. It is reversible and, by cell division, can remain stable. The classical understanding of DNA methylation is that when it occurs in a CpG rich promoter region, it silences gene expression (Bock et al., 2012). It occurs mainly in metazoan genomes on CpG dinucleotides and occasionally on non-CpG bases. In human embryonic stem and neuronal cells, non-CpG methylation has been predominantly observed (Lister et al., 2009) (Lister et al., 2013). In the human genome, there are about 28 million CpGs, 60-80% of which are normally mentholated. In CG-dense regions called CpG islands in the human genome, less than 10 percent of CpGs occur in.

It has been shown that DNA methylation is often not distributed evenly across the genome, but rather is correlated with the density of CpG. In vertebrate genomes, in CpG-rich regions such as CpG islands, cytosine bases are typically unmethylated and appear to be methylated in CpG-deficient regions. Except for the CpG islands, vertebrate genomes are mostly CpG deficient. Invertebrates such as Drosophila melanogaster and elegant Caenorhabditis, on the other hand, do not display cytosine methylation and thus do not have CpG rich and poor regions, but rather a steady CpG frequency over the genome (Deaton and Bird, 2011). DNA methylation is formed in combination with DNMT3L by the DNA methyltransferases DNMT3A and DNMT3 B and is sustained by the methyltransferase DNMT1 and associated proteins through/after cell division. During early development, DNMT3a and DNMT3b were in charge of de novo methylation. During replication or exclusion of DNMT1 from the nucleus, loss of 5mC can be accomplished passively by dilution. Recent results of the protein family ten-eleven translocation (TET) and their ability to transform 5-methylcytosine (5mC) into 5-hydroxymethylcytosine (5hmC) in vertebrates provide a mechanism for the demethylation of catalysed active DNA (Tahiliani et al., 2009). Iterative TET-catalyzed 5hmC oxidation leads to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). G/T mismatch-specific thymine-DNA glycosylase (TDG) excises the 5caC mark from DNA, which results in cytosine residue returning to its unmodified state (He et al., 2011). Besides these, mostly bacteria, but probably


Sequencing Quality Check & remove adapters
Align to the genome
Remove PCR bias: Deduplication (WGBS)
Check conversion rates
Bisulfite Reads
TGGGATGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGG
TGGGATGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGGTA
TGGGATGAAGGTTTTCGAGTTTTTTGAGAGTAGCGTTAGGGTAC
TGGGATGAAGGTTTTTGAGTTTTTTGAGAGTAGCGTTAGGGTACG
TGGGATGAAGGTTTTTGAGTTTTTCGAGAGTAGCGTTAGGGTACGG
GGATGAAGGTTTTTGAGTTTTTCGAGAGTAGCGTTAGGGTACGGATGT
GATGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGGTACGGATGT
TGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGGTACGGATGT
Genome
CGGGATGAAGGCCCCCGAGCTCCCCCGAGAGCAGCGTCAGGGCACGGATGC
Methylation proportion
1.0
0.5
0.0
Methylation Calling: uses coverage and quality scores
Filtering for known C/T SNPs
Remove PCR bias: Coverage Filtering (RRBS)
Results as Tables, BED files, summary statistics
Differential Methylation
Visualization & Clustering/PCA
Methylation segmentation
Annotation for DMRs/DMCs

Higher eukaryotes contain base modifications on bases other than cytosine, such as mentholated adenine or guanine (Clark et al., 2011). One of the most reliable and popular ways to measure DNA methylation is bisulfate sequencing. This method, and related ones, allows measurement of

DNA methylation at the single nucleotide resolution. In this review, we describe strategies for analyzing data from bisulfate sequencing experiments. First, we introduce high-throughput sequencing techniques based on bisulfate treatment. Next, we summarize algorithms and tools for detecting differential methylation and methylation profile segmentation. Finally, we discuss management of large datasets and data analysis workflows with a guided user interface. The computational workflow summarizing all the necessary steps is shown in Fig. 1.

Bisulfate sequencing for detection of methylation and other base modifications

Techniques for profiling genome-wide DNA methylation fall into four categories: methods focused on DNA-methylation-sensitive restriction enzymes (such as MRE-seq), methyl cytosine-specific antibodies (such as MeDIP-seq methylated DNA immunoprecipitation (Weber et al., 2005)), methyl-CpG-binding domains for interest-site enrichment of mentholated DNA (Brinkman et al., 2010) and bas-specific domains (Brinkman et al., 2010) However, over measured regions ranging in size from 100 to 1000 bp, the first three methods allow methylation detection. DNA methylation at single nucleotide resolution is determined by methods using sodium bisulfate treatment that converts unmethylated cytosine to thymine (via uracil) while mentholated cytosine remains safe.

We will concentrate on bisulfite-conversion based sequencing techniques for the remainder of this segment. The 'gold standard' for assaying DNA methylation is called whole genome bisulfite sequencing (WGBS) because it offers global coverage at single-base resolution. Briefly, with high-throughput sequencing, it incorporates bisulfite conversion of DNA molecules. The genomic DNA is first randomly fragmented to the appropriate size to conduct WGBS (200 bp). By ligation to adaptors containing 5mCs, the fragmented DNA is translated into a sequencing library. The library of the series is then treated with bisulfate. This treatment transforms unmethylated cytosine into uracil effectively. It is sequenced using high-throughput sequencing after amplification of the library treated with bisulfate by PCR. Uracils will be portrayed as thymine after the PCR. Not only does an effective recall of cytosine methylation require adequate sequencing depth, but it also relies heavily on the quality of bisulfite conversion and amplification of the library. The advantage of this shotgun approach is that, in unbiased representation, it usually achieves coverage of over 90% of the CpGs in the human genome. It enables the detection of non-CG methylation as well as the identification of partly mentholated domains in embryonic stem cell valleys (PMDs, and regions with low methylation distal regulatory elements (LMRs, and DNA methylation valleys (DMVs).

WGBS remains the most costly procedure, considering its benefits, and standard library preparation involves relatively large amounts of DNA (100ng-5 ug); as such, it is typically not applied to large numbers of samples. High sequencing depth is needed to achieve high sensitivity in the detection of methylation differences between samples, leading to substantial increases in sequencing costs. Another technique is reduced representation bisulfate sequencing (RRBS), which can also profile DNA methylation at single-base resolution. It combines genomic DNA digestion with restriction enzymes and bisulfate sequencing in order to enrich areas with a high CpG content. It then relies first on genomic DNA digestion with restriction enzymes, such as MspI, which recognises 5'-CCGG-3' sequences and cleaves upstream CpG dinucleotide phosphodiester bonds. Only CpG dense regions can be sequenced and CpG-deficient regions, such as functional enhancers, intronic regions, interagency regions or typically low methylated regions (LMRs) of the genome, are not interrogated. In CpGpoor areas, it has broad genome coverage and explores about 4% to 17% of the approximately 28 million CpG dinucleotides distributed throughout the human genome, depending on the depth of sequencing and which variant of RRBS is used.

A combination of bisulfite sequencing with high-throughput sequencing is often used for Targeted Bisulfite sequencing, but a previous set of predefined genomic regions of interest is needed. PCR amplification of regions of interest padlock probes (Ball et al., 2009), hybridization-based goal enrichment and convert-then-capture approaches are commonly used protocols The fact that bisulfite sequencing cannot differentiate between hydroxymethylation (5hmC) and methylation (5mC) is one of the major assay-specific problems (Huang et al., 2010). Upon bisulfite treatment, hydroxymethylation converts to cyto-5-methanesulfonate which then reads as a C when sequenced In addition, a mechanism for non-passive DNA demethylation is 5hmC mediated by TET proteins. Measurements of methylation for tissues with high 5-hydroxymethylation would therefore be inaccurate, at least in some genomic regions. The design of Tet-assisted bisulfite sequencing (TAB-seq) (Yu et al., 2012) and oxBS-Seq (Booth et al., 2012) has made it possible to differentiate between the two single-base resolution modifications. In addition to 5hmC, mammalian genomes have recently achieved single-base resolution mapping of 5caC using CAB-seq (Lu et al., 2013) and detection of 5 fc (fCAB-seq (Song et al., 2013a; Booth et al., 2014) and redBS-Seq (Song et al., 2013a; Booth et al., 2014)).

Alignment and data processing for bisulfite sequencing

Since BS-seq changes unmethylated cytosine (C) to thymine (T), subsequent steps of analysis focus on counting the number of conversions from C to T and quantifying the per-base methylation ratio. This is simply achieved by recognizing C-to-T conversions in the aligned reads and dividing the amount of Cs for each cytosine in the genome by the sum of Ts and Cs. Being able to do the quantification accurately depends on quality control prior to alignment, the methods of alignment and quality control post alignment. Since the quality of base calling is not constant and could change between sequencing runs and within the same reading, it is necessary to check the quality of the base (which represents the level of confidence in the base calls). Miscalled bases may erroneously be counted as C-T conversions and such errors should, if possible, be avoided. You can perform this simple quality check via fast QC software

In addition, adapters may often be sequenced and they can either lower the alignment rates or trigger incorrect C-T conversions if not properly removed. To mitigate problems with false C-T conversions and to improve alignment rates, we suggest trimming poor quality bases on sequence ends and eliminating adapters. Using trimming programmers such as Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim galore/), this can be done. When quality control and processing of pre-alignment is completed, the next step is to align where future C-T conversions should be treated. The methods of BS-seq alignment depend mainly on modifications of established methods of short-read alignment. Bismarck, for example, relies on Bowtie and conversion of reads and genomes in silico C-T (Krueger and Andrews, 2011). In silico conversion strategy, many other aligners use this, such as: Methyl Coder (Pedersen et al., 2011), BS-seeker2 BRAT-BW and Bison Other strategies, such as Last (Frith et al., 2012), use a particular score matrix that can tolerate C-T mismatches or masks Ts in the reads and matches them to genomic Cs, such as BSMAP (Xi and Li, 2009). As new alternatives always appear, there are few detailed benchmarks of the aligners, but previous attempts to compare the aligners' output did not find sufficient differences between aligners to exclude any from consideration.

In addition, in some aspects of the benchmark, recent tools are typically only better; they may, for instance, outperform competing tools in terms of computation time, but show a much higher memory footprint or have a worse mapping quality, some of these performance differences also disappear by different tool parameters and we do not see convincing evidence that an existing tool such as Bismar We also use Bismark for our own work because it offers BAM files, as well as additional methylation call-related metrics and files. There is still a need for more quality control after the alignment and methylation calls. To be highlighted here, there are possible concerns. Unmethylated Cs are inserted at the ends of the DNA fragments during the final repair phase following fragmentation (Bock, 2012). This leads to a large decrease in the average degree of methylation that can be observed at certain ends in a methylation bias (M-bias) map. In addition, incomplete conversion can occur during bisulfate treatment, where not all unmethylated Cs are converted to Ts. A simple solution would be to ignore the affected positions in the sequenced reads (Genereux et al., 2008).

Owing to the perception of the non-converted unmethylated cytosine as mentholated, incomplete conversion produces false positive results. We can calibrate the conversion rate for organisms without significant non-CpG methylation, such as humans, by using the percentage of non-CpG methylation. We expect the conversion rate to be as close to 100 percent as possible for a high quality experiment, average values for a successful experiment would be higher than 99.5 percent . The addition of spike-in sequences with unmethylated Cs and the number of Ts for unmethylated Cs is another way to calculate the conversion rate. A further possible issue is the degradation of DNA during bisulfite therapy. Long incubation time and high concentration of bisulfite will lead to approximately 90% of the incubated DNA degradation (Grunau et al., 2001). Testing specific alignment rates and reading lengths after trimming is therefore critical. In addition, the majority of CpGs with high inter population differences have been shown to contain common genomic SNPs (lower allele frequency > 0.01). (Daca-Roszak et al., 2015).

We suggest that known C/T SNPs that can interfere with methylation calls be removed to ensure more accurate analysis of the results. PCR bias is resolved by the last post-alignment consistency process. A easy approach might be to delete reads on the same strand that are aligned to the exact same genomic location. You may use the "samtools rmdup" command or the Bismark tools to perform this de-duplication. For RRBS, it is not advisable to eliminate PCR duplicates by looking at overlapping reading coordinates. Instead, by eliminating regions with exceptionally high coverage, one can try to eliminate PCR bias; this approach generates concomitant methylation measurements with orthogonal methods such as pyrosequencing (Akalin et al., 2012a).

Segmentation of the methyl me

The study of methylation dynamics is not limited solely to sample-to-sample differentially methylated areas, except that there is also an interest in analyzing the methylation profiles of the same sample. Depressions in methylation profiles usually recognize regulatory regions such as gene promoters that co-localize with islands of CG-dense CpG. Many gene-body regions, on the other hand, are heavily methylated and CpG-poor (Bock et al., 2012). Such observations will establish a bimodal model based on the local density of CpGs of either hyper- or hypomethylated regions (Lövkvist et al., 2016). However, given the discovery of CpG-poor regions with locally reduced levels of methylation (on average 30 percent) in pluripotent embryonic stem cells and in neuronal progenitors in both mouse and human, a different model seems also appropriate (Stadler et al., 2011). Such low-methylated regions (LMRs) are distal to promoters, have no overlap with islands of CpG and are associated with enhancer marks such as p300 binding sites and enrichment with H3K27ac.

Using computational methods, the detection of these LMRs can be accomplished by segmenting the methyl me. One of the well-known segmentation methods is based on a three-state Hidden Markov Model (HMM) without awareness of any additional genomic information such as CpG density or functional annotations, taking only DNA methylation into account (Stadler et al., 2011). Completely methylated regions (FMRs), unmethylated regions (UMRs) and low-mentholated regions were the three states that the authors targeted (LMRs).

This segmentation is a description of methylome properties and characteristics in which unmethylated CpG islands correspond to UMRs (Deaton and Bird, 2011), most of which are categorised as FMR since most of the genome is methylated (Bird, 2002) and LMRs are a new function with intermediate methylation levels, poor CpG content and shorter duration compared to CpG islands (Stadler et al., 2011). A two model state HMM is assumed by other segmentation techniques such as MethPipe and does not distinguish between LMRs and UMRs. The authors of the MethylSeekR R package (Burger et al., 2013) adapt the concept of a three-state methyl me and also recognise partially methylated domains (PMDs), another characteristic of methyl me found in human fibroblast, for example, but not in embryonic stem cells of H1.

These wide regions are characterised by highly disordered methylation, spanning hundreds of kilobases, with average methylation levels below 70% and covering almost 40% of the genome (Lister et al., 2009; Gaidatzis et al., 2014). PMDs do not generally occur in each methyl me, but a sliding window statistics can detect their existence (Burger et al., 2013). The genome wide identification is achieved by training a two-state HMM in both MethylSeekR and MethPipe, to distinguish PMDs from context regions. Prior to the characterization of UMRs/LMRs or hyper-/hypo ethylated areas, the PMDs are then masked (Song et al., 2013b) (Burger et al., 2013). There are also other strategies of segmentation based on change-point analysis, where a genome-wide signal's change points are reported and the genome is divided between consecutive change points into regions. In the sense of copy number variation detection, this technique is usually used (Klambauer et al., 2012), but can also be extended to methylome segmentation. A package that implements this segmentation approach based on change points is methylKit. It defines segments that use a mixture modelling approach to be further clustered. This clustering is focused only on the segments' average degree of methylation and enables the identification of distinct methyl me characteristics comparable to UMRs, LMRs and FMRs. This strategy offers a more robust segmentation approach where one can settle on the number of segmentation groups after segmentation. Whereas in HMM-based models, the numbers of segmentation groups must be identified, a priori, or several rounds of HMMs with different numbers must be run to identify which model matches the data best.

Strategies for dealing with large datasets

With growing amounts of epigenetic data accessible to the public, it is tempting for several purposes to recreate the findings of published articles, e.g. to better understand the rationale behind the measures taken by the writers or to improve general data intuition. In the case of bisulfite sequencing results, using whole genome methylation data from multiple samples, we might want to perform differential methylation analysis in R. The issue is that file sizes may easily range from hundreds of megabytes to gigabytes for genome-wide studies, and processing numerous instances of those memory files (RAM) can become unworkable unless we have access to a high-performance cluster (HPC) with comprehensive RAM. If we want to use a small RAM desktop computer or laptop, we either need to restrict our research to a subset of data or use packages that can cope with this situation.

In order to allow multiple WGBS samples to be processed within a reasonable period, the developers of the RAD meth package for differential methylation analysis suggest running the programme on a "computing cluster with a few hundred available nodes." On a personal workstation, the same analysis can also be conducted with the downside of increasing the computational time, which is usually based on three factors: the coverage of the sample, the number of locations analysed and the number of samples. If one's workstation is a multicore machine, there is one way to speed up the time-consuming regression process. The authors provided a script for splitting the input data into smaller parts that could be separately processed and then combined using UNIX commands. RnBeads (Assenov et al., 2014), which internally relies on the 'ff' package, is a package for the systematic analysis of genome-wide DNA methylation data that can accommodate massive data. The 'ff' R package (Adler et al., 2014) makes it possible to work with datasets larger than usable RAM by storing them as temporary files and providing an interface to allow flat files to be read and written and run on the parts loaded into R. By leveraging flat file databases, the methylKit package offers very similar functionality to replace in-memory objects if the objects become too big. In addition to meta data, the internal data has a tabular structure that stores chromosome, start/end location, strand data of the related CpG base, as well as several other biological formats such as BED, GFF or SAM. It can be indexed using the generic Tabix tool by exporting this tabular information into a TAB-delimited file and ensuring that it is position-sorted accordingly (Lövkvist et al., 2016). Tabix indexing in general is a generalisation of BAM indexing for TAB-delimited generic directories. In terms of the few search function calls per query, it inherits all the benefits of BAM indexing, including data compression and effective random access (Li, 2011). MethylKit relies on Rsamtools (http://bioconductor.org/packages/release/bioc/html/Rsamtools.html), which incorporates R tabix features, so that internal methylKit artifacts can be stored easily on the disc as a compressed file and can still be accessed quickly. Another benefit is that existing compressed files can be loaded in collaborative sessions, enabling intermediate analysis results to be backed up and transferred.

Annotation of DMRs/DMCs and segments

The regions of interest obtained through differential methylation or segmentation analysis often need to be integrated with genome annotation datasets. Without this type of integration, differential methylation or segmentation results will be hard to interpret in biological terms. The most common annotation task is to see where regions of interest land in relation to genes and gene parts and regulatory regions: Do they mostly occupy promoter, intronic or exonic regions? Do they overlap with repeats? Do they overlap with other epigenomic markers or long-range regulatory regions? These questions are not specific to methylation −nearly all regions of interest obtained via genome-wide studies have to deal with such questions. Thus, there are already multiple software tools that can produce such annotations. One is the Bioconductor package genomation (Akalin et al., 2015). It can be used to annotate DMRs/DMCs and it can also be used to integrate methylation proportions over the genome with other quantitative information and produce meta-gene plots or heatmaps. Another similar package is ChIPpeakAnno (Zhu et al., 2010), which is designed for ChIP-seq peak

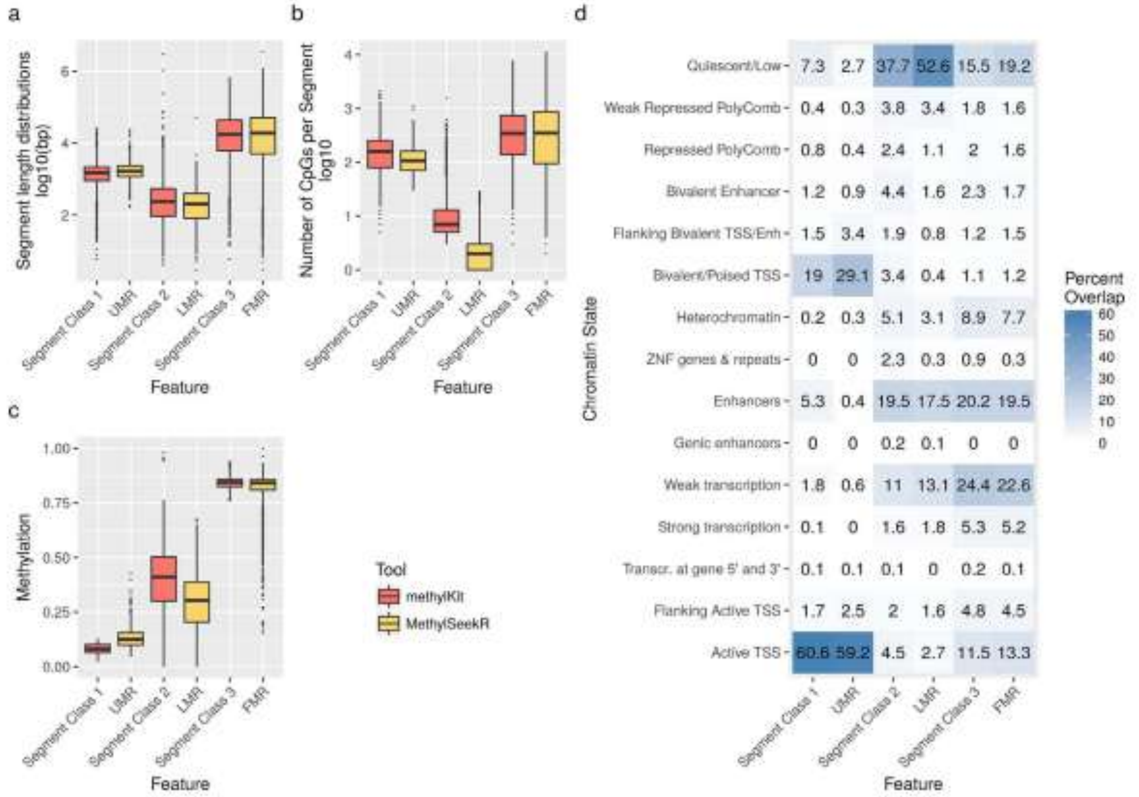


Fig. 4. Comparison of characteristics found by segmentation instruments examining methyl me chromosome 2 of the H1 embryonic stem cells. The distribution of (a) segment lengths in log10 transformed base pairs (bp) (b) CpG location covered by each segment in log10 transformed numbers (c) average methylation score per segment is shown by box plots for each feature. (a) − (c) Boxplot colours show either methylKit or MethylSeekR as the method generating the features. (d) Heatmap showing the percentage of segments of methylSeeker and methylKit overlapping H1 embryonic stem cells with chromatin state annotations.

Annotation but could also be used for DMR/DMC annotation to a certain degree.

CONCLUSIONS

In this review article, we explored the experimental and analytical methods of genome-wide or targeted measurement and analysis of DNA methylation. For bisulfate sequencing experiments, starting from read alignment and quality check, we presented all the required steps of downstream study. Differential methylation and methyl-me segmentation techniques were explored and compared. Our attempts to compare differential methods of methylation have shown that different methods have comparable efficiency. Based on the ultimate aim of their study, one may choose methods. For subsequent validation studies (DSS, limma, BS mooth, MethylKit with F-test and over dispersion correction), the techniques that are rigorous and restrict the false positive rates are fine, but these methods compromise sensitivity (true positive rate) in order to reduce false positives. A very relaxed technique has the best overall accuracy but the highest false positive rate, such as the default methylKit process. Chi-square checking after over dispersion correction is a successful alternative to rigid and relaxed techniques (implemented in methylKit).

REFERENCES

1. Akalin, A., Garrett-Bakelman, F.E., Kormaksson, M., Busuttil, J., Zhang, L., Khrebtukova, I., et. al. (2012a). Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. PLoS Genet. 8, e1002781. http://dx.doi.org/10.1371/journal.pgen.1002781.

2. Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., et. al. (2012b). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 13, pp. R87. http://dx.doi.org/10.1186/gb2012-13-10-r87.

3. Akalin, A., Franke, V., Vlahoviček, K., Mason, C.E. (2015). Schübeler D. genomation: a toolkit to summarize, annotate and visualize genomic intervals. Bioinformatics 31, pp. 1127–1129. http://dx.doi.org/10.1093/bioinformatics/btu775.

4. Assenov, Y., Müller, F., Lutsik, P., Walter, J., Lengauer, T., Bock, C. (2014). Comprehensive analysis of DNA methylation data with RnBeads. Nat. Methods 11, pp. 1138–1140. http://dx.doi.org/10.1038/nmeth.3115.

5. Ball, M.P., Li, J.B., Gao, Y., Lee, J.-H., LeProust, E.M., Park, I.-H., et. al. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat. Biotechnol. 27, pp. 361–368. http://dx.doi.org/10.1038/nbt.1533.

6. Baubec, T., Akalin, A. (2016). Genome-wide analysis of DNA methylation patterns by highthroughput sequencing. Field Guidelines Genetic Exp. Designs in High-Throughput Sequen. pp. 197–221. http://dx.doi.org/10.1007/978-3-319-31350-4_9.

7. Bird, A., 2002. DNA methylation patterns and epigenetic memory. Genes Dev. 16, pp. 6–21. http://dx.doi.org/10.1101/gad.947102.

8. Bock, C., Beerman, I., Lien, W.-H., Smith, Z.D., Gu, H., Boyle, P., et. al. (2012). DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. Mol. Cell 47, pp. 633–647. http://dx.doi.org/10.1016/j.molcel.2012.06.019.

9. Bock, C., 2012. Analysing and interpreting DNA methylation data. Nat. Rev. Genet. 13, 705–719. http://dx.doi.org/10.1038/nrg3273. Bonev, B., Cavalli, G. (2016). Organization and function of the 3D genome. Nat. Rev. Genet. 17, pp. 772. http://dx.doi.org/10.1038/nrg.2016.147.

10. Booth, M.J., Branco, M.R., Ficz, G., Oxley, D., Krueger, F., Reik, W., et. al. (2012). Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at singlebase resolution. Science 336, pp. 934–937. http://dx.doi.org/10.1126/science.1220671.

11. Booth, M.J., Marsico, G., Bachman, M., Beraldi, D., Balasubramanian, S. (2014). Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. Nat. Chem. 6, 435–440. http://dx.doi.org/10.1038/nchem.1893.

12. Brinkman, A.B., Simmer, F., Ma, K., Kaan, A., Zhu, J., Stunnenberg, H.G. (2010). Wholegenome DNA methylation profiling using methylCap-seq. Methods 52, pp. 232–236. http://dx.doi.org/10.1016/j.ymeth.2010.06.012.