Journal of Biotechnology 261 (2017) 105–115

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Strategies for analyzing bisulfite sequencing data

Katarzyna Wreczycka[a,d,1], Alexander Gosdschan[a,1], Dilmurat Yusuf[a], Björn Grüning[b], Yassen Assenov[c], Altuna Akalin[a,*]

[a] Bioinformatics Platform, Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Berlin, Germany
[b] Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany
[c] Division of Epigenomics and Cancer Risk Factors at the German Cancer Research Center (DKFZ), Heidelberg, Germany
[d] Berlin Institute of Health (BIH), 10178 Berlin, Germany

ARTICLE INFO

Keywords:
Methylation
Bisulfite-sequencing
Differential methylation
Methylation segmentation
Galaxy

ABSTRACT

DNA methylation is one of the main epigenetic modifications in the eukaryotic genome; it has been shown to play a role in cell-type specific regulation of gene expression, and therefore cell-type identity. Bisulfite sequencing is the gold-standard for measuring methylation over the genomes of interest. Here, we review several techniques used for the analysis of high-throughput bisulfite sequencing. We introduce specialized short-read alignment techniques as well as pre/post-alignment quality check methods to ensure data quality. Furthermore, we discuss subsequent analysis steps after alignment. We introduce various differential methylation methods and compare their performance using simulated and real bisulfite sequencing datasets. We also discuss the methods used to segment methylomes in order to pinpoint regulatory regions. We introduce annotation methods that can be used for further classification of regions returned by segmentation and differential methylation methods. Finally, we review software packages that implement strategies to efficiently deal with large bisulfite sequencing datasets locally and we discuss online analysis workflows that do not require any prior programming skills. The analysis strategies described in this review will guide researchers at any level to the best practices of bisulfite sequencing analysis.

1. Introduction

Cytosine methylation (5-methylcytosine, 5mC) is one of the main covalent base modifications in eukaryotic genomes. It is involved in epigenetic regulation of gene expression in a cell-type specific manner. It is reversible and can remain stable through cell division. The classical understanding of DNA methylation is that it silences gene expression when occurs at a CpG rich promoter region (Bock et al., 2012). It occurs predominantly on CpG dinucleotides and seldom on non-CpG bases in metazoan genomes. The non-CpG methylation has been mainly observed in human embryonic stem and neuronal cells (Lister et al., 2009) (Lister et al., 2013). There are roughly 28 million CpGs in the human genome, 60–80% are generally methylated. Less than 10% of CpGs occur in CG-dense regions that are termed CpG islands in the human genome (Smith and Meissner, 2013). It has been demonstrated that DNA methylation is also not uniformly distributed over the genome, but rather is associated with CpG density. In vertebrate genomes, cytosine bases are usually unmethylated in CpG-rich regions such as CpG islands and tend to be methylated in CpG-deficient regions. Vertebrate genomes are largely CpG deficient except at CpG islands. Conversely, invertebrates such as *Drosophila melanogaster* and *Caenorhabditis elegans* do not exhibit cytosine methylation and consequently do not have CpG rich and poor regions but rather a steady CpG frequency over the genome (Deaton and Bird, 2011). DNA methylation is established by DNA methyltransferases DNMT3A and DNMT3 B in combination with DNMT3L and maintained through/after cell division by the methyltransferase DNMT1 and associated proteins. DNMT3a and DNMT3b are in charge of the de novo methylation during early development. Loss of 5mC can be achieved passively by dilution during replication or exclusion of DNMT1 from the nucleus. Recent discoveries of ten-eleven translocation (TET) family of proteins and their ability to convert 5-methylcytosine (5mC) into 5-hydroxymethylcytosine (5hmC) in vertebrates provide a path for catalysed active DNA demethylation (Tahiliani et al., 2009). Iterative oxidations of 5hmC catalysed by TET result in 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). 5caC mark is excised from DNA by G/T mismatch-specific thymine-DNA glycosylase (TDG), which as a result returns cytosine residue back to its unmodified state (He et al., 2011). Apart from these, mainly bacteria but possibly

* Corresponding author.
E-mail address: altuna.akalin@mdc-berlin.de (A. Akalin).
[1] Equal contributions.

http://dx.doi.org/10.1016/j.jbiotec.2017.08.007
Received 17 February 2017; Received in revised form 7 August 2017; Accepted 8 August 2017
Available online 16 August 2017

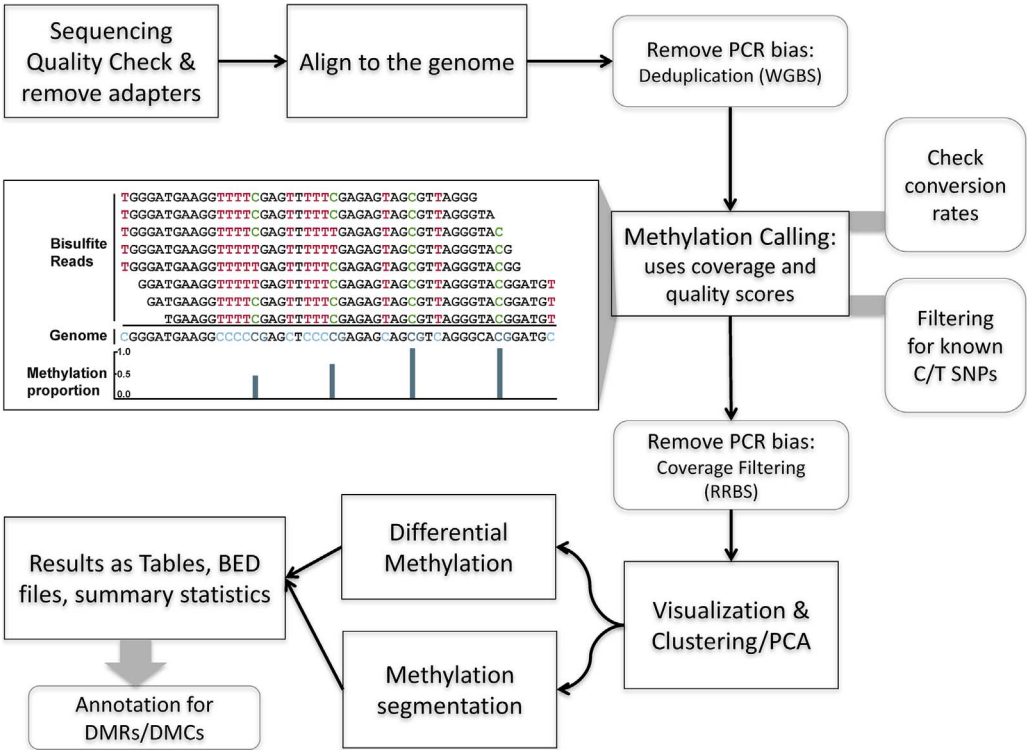

Fig. 1. Workflow for analysis of DNA methylation using data from bisulfite sequencing experiments.

higher eukaryotes contain base modifications on bases other than cytosine, such as methylated adenine or guanine (Clark et al., 2011).

One of the most reliable and popular ways to measure DNA methylation is bisulfite sequencing. This method, and related ones, allow measurement of DNA methylation at the single nucleotide resolution. In this review, we describe strategies for analyzing data from bisulfite sequencing experiments. First, we introduce high-throughput sequencing techniques based on bisulfite 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.

2. Bisulfite sequencing for detection of methylation and other base modifications

Techniques for profiling genome-wide DNA methylation fall into four categories: methods based on restriction enzymes sensitive to DNA methylation (such as MRE-seq), methylcytosine-specific antibodies (such as methylated DNA immunoprecipitation using MeDIP-seq (Weber et al., 2005)), methyl-CpG-binding domains to enrich for methylated DNA at sites of interest (Brinkman et al., 2010), and those based on sodium bisulfite treatment. However, the first three methods allow methylation detection over measured regions ranging in size from 100 to 1000 bp. Methods that use sodium bisulfite treatment, which converts unmethylated cytosines to thymine (via uracil) while methylated cytosines remain protected, measure DNA methylation at single nucleotide resolution (Baubec and Akalin, 2016). For the remainder of this section, we will focus on bisulfite-conversion based sequencing techniques.

Whole genome bisulfite sequencing (WGBS) is considered the 'gold standard' for assaying DNA methylation because it provides global coverage at single-base resolution. Briefly, it combines bisulfite conversion of DNA molecules with high-throughput sequencing. To perform WGBS, the genomic DNA is first randomly fragmented to the desired size (200 bp). The fragmented DNA is converted into a sequencing library by ligation to adaptors that contain 5mCs. The sequence library is then treated with bisulfite. This treatment effectively converts unmethylated cytosines to uracil. After amplifying the library treated with bisulfite by PCR, it is sequenced using high-throughput sequencing. After the PCR, uracils will be represented as thymines. A precise recall of cytosine methylation requires not only sufficient sequencing depth, but also strongly depends on the quality of bisulfite conversion and library amplification. The benefit of this shotgun approach is that it typically reaches coverage of over 90% of the CpGs in the human genome in unbiased representation. It allows identification of non-CG methylation as well as identification of partially methylated domains (PMDs, (Lister et al., 2009)), and regions at distal regulatory elements with low methylation (LMRs, (Stadler et al., 2011)) and DNA methylation valleys (DMVs) in embryonic stem cells (Xie et al., 2013). Despite its advantages, WGBS remains the most expensive technique and standard library prep requires relatively large quantities of DNA (100ng–5 ug); as such, it is usually not applied to large numbers of samples (Stirzaker et al., 2014). To achieve high sensitivity in detecting methylation differences between samples, high sequencing depth is required which leads to significant increase in sequencing cost.

Reduced representation bisulfite sequencing (RRBS) is another technique that can also profile DNA methylation at single-base resolution. It combines digestion of genomic DNA with restriction enzymes and sequencing with bisulfite treatment in order to enrich for areas with high CpG content. Thus, it relies first on digestion of genomic DNA with restriction enzymes, such as MspI which recognises 5′-CCGG-3′ sequences and cleaves the phosphodiester bonds upstream of CpG dinucleotide. It can sequence only CpG dense regions and does not interrogate CpG-deficient regions such as functional enhancers, intronic regions, intergenic regions or in general lowly methylated regions (LMRs) of the genome. It has limited coverage of the genome in CpG-poor regions and examines about 4% to 17% of the approximately 28 million CpG dinucleotides distributed throughout the human genome depending on the sequencing depth and which variant of RRBS is used (Meissner et al., 2005; Rampal et al., 2014).

Targeted Bisulfite sequencing also uses a combination of bisulfite sequencing with high-throughput sequencing, but it needs a prior