Available online at www.sciencedirect.com
SciVerse ScienceDirect
Journal of Genetics and Genomics 40 (2013) 249−260
JGG

REVIEW

An Integrated Workflow for DNA Methylation Analysis

Pingchuan Li, Feray Demirci, Gayathri Mahalingam, Caghan Demirci, Mayumi Nakano, Blake C. Meyers*

Department of Plant & Soil Sciences, Delaware Biotechnology Institute, University of Delaware, Newark, DE 19711, USA

Received 31 January 2013; revised 25 March 2013; accepted 25 March 2013
Available online 30 March 2013

ABSTRACT

The analysis of cytosine methylation provides a new way to assess and describe epigenetic regulation at a whole-genome level in many eukaryotes. DNA methylation has a demonstrated role in the genome stability and protection, regulation of gene expression and many other aspects of genome function and maintenance. BS-seq is a relatively unbiased method for profiling the DNA methylation, with a resolution capable of measuring methylation at individual cytosines. Here we describe, as an example, a workflow to handle DNA methylation analysis, from BS-seq library preparation to the data visualization. We describe some applications for the analysis and interpretation of these data. Our laboratory provides public access to plant DNA methylation data *via* visualization tools available at our "Next-Gen Sequence" websites (http://mpss.udel.edu), along with small RNA, RNA-seq and other data types.

KEYWORDS: DNA methylation; BS-seq; Epigenetics

INTRODUCTION

The modified nucleotide 5-methylcytosine ("5mC") is a common DNA modification in many eukaryotic organisms. This nucleotide is a derivative of cytosine with a methyl group added at the 5′ position by a methyltransferase enzyme, and the modified nucleotide is also known as the 5th base (the others, of course, are A/T/C/G) (Lister and Ecker, 2009). The prevalence and possible biological functions for 5mC was described as early as 1964, at which time the modified nucleotide was thought to play a role in DNA protection or diversification (Srinivasan and Borek, 1964). More recently, diverse functions in gene and chromatin regulation, mainly as a negative regulator, have been described in plants and animals, and widely reviewed (Law and Jacobsen, 2010; He et al., 2011; Jones, 2012).

DNA methylation can occur in three different contexts, CG, CHG and CHH (where H = A, C or T). These modifications are performed by three DNA methyltransferases in plants. DNA METHYLTRANSFERASE 1 (MET1) is the plant ortholog of mammalian DNMT1 (Finnegan and Dennis, 1993; Finnegan et al., 1996; Ronemus et al., 1996), responsible for the maintenance of CG methylation (Vongs et al., 1993; Jones et al., 2001). CHROMOMETHYLASE 3 (CMT3) is a plant-specific DNA methyltransferase, responsible for the maintenance of CHG methylation (Lindroth et al., 2001) *via* recognition of histone H3 lysine 9 (H3K9) dimethylation (Jackson et al., 2002; Du et al., 2012). DOMAINS REARRANGED METHYLTRANSFERASES 1 and 2 (DMR1, DMR2) are the plant orthologs of DNMT3A/3B in mammals (Okano et al., 1999), have partially overlapping functions so are often considered together, and in plants they function to generate 5mC at CHH sites. They function with CMT3 in the so-called "RNA-directed DNA methylation" (RdDM) pathway, directed by small RNAs (Cao et al., 2003). The RdDM pathway is a complex regulatory pathway for *de novo*

* Corresponding author. Tel: +1 302 831 3418, fax: +1 302 831 4841.
E-mail address: meyers@dbi.udel.edu (B.C. Meyers).

http://dx.doi.org/10.1016/j.jgg.2013.03.010

DNA methylation, with numerous genetically- and biochemically-defined components (Law and Jacobsen, 2010). In addition to these three cytosine methyltransferases, a chromatin-remodeling helicase DDM1 (DECREASE IN DNA METHYLATION 1) is a major component for the maintenance of methylation in all three cytosine contexts (Jeddeloh et al., 1999). Thus there are numerous factors which provide exquisite regulatory control of DNA methylation in plants and other eukaryotes. A nearly exhaustive analysis of mutants in the genes involved DNA methylation and RdDM has revealed novel correlations between the three contexts of cytosine methylation as well as the interplay between DNA methylation and histone modification (Stroud et al., 2013). Given the importance of the DNA methylation to gene regulation, an integrated platform for systematic studies of DNA methylation is now an important starting point for laboratories working in this field; such a workflow needs to cover the beginning steps of library preparation and validation, and the later steps such as data visualization and an analytical pipeline.

High-throughput sequencing technologies provide a mechanism to quickly and comprehensively evaluate the positions of 5mC at a single-base resolution. DNA methylation has traditionally been assayed by sodium bisulfite conversion of genomic DNA (described in more detail below), a process scaled up to a whole-genome method known as bisulfite sequencing (BS-seq) and described first by the Jacobsen lab (Cokus et al., 2008) and soon thereafter by the Ecker lab (Lister et al., 2008). As one benefit from the latest sequencing technologies, the drop in their cost, and the huge improvements in their yield, large amounts of plant BS-seq data are now being generated, presenting an issue of how to handle such data. In the decade since we established our websites for next-gen transcriptional data (Meyers et al., 2004), we have aimed to supply powerful genomic analysis tools for the public and data that include hundreds of libraries from mRNA (full-length and decay products) and small RNA, representing organisms as diverse as *Arabidopsis*, rice, soybean, maize, fungi and even chickens, but mainly focusing on plants (http://mpss.udel.edu). BS-seq data is important to integrate into these resources, as 5mC directly influences gene activity and transcript levels and there is substantial utility to comparing mRNA, small RNA and 5mC levels, yet these data are substantially greater in size and have unique challenges for mapping reads. In this article, we provide an overview of one such workflow pipeline − the one used in our laboratory, including laboratory and informatics methods − as a guide for other labs interested in these data and process of utilizing them.

A BRIEF COMPARISON OF METHODS FOR THE MEASUREMENT OF DNA METHYLATION

Recent years have brought a wealth of developments in high-throughput measurements of DNA methylation. There are essentially two major categories of methods in current use, microarray-based methods that require hybridization, and next-gen-based sequencing methods. Comparisons of these methods have been published recently (Beck, 2010; Harris et al., 2010; Laird, 2010; Bock, 2012). Array-based hybridization for 5mC measurement was developed earlier, and typically utilizes methylation-sensitive restriction enzymes like *Hpa* II or *Msp* I for genomic DNA digestion. In a typical experiment for example to examine a mutant in one of the DNA methylation pathways, mutant *versus* control will be labeled with Cy5 and Cy3, respectively, for two-color hybridization (Schumacher et al., 2006; Yan et al., 2009). Methylation levels are reflected by the fluorescence intensity of the hybridized DNA; data need strict normalization and statistics to assess the ratio of methylated to unmethylated DNA (Nouzova et al., 2004). In a small genome, like *Arabidopsis*, amplification prior to the microarray hybridization is unnecessary (Tompa et al., 2002; Tran et al., 2005), but adding the methylation-dependent endonuclease *Mcr*BC (which cleaves DNA at methylcytosines) can increase the sensitivity for densely methylated regions. In the era of next-gen sequencing, array-based methods are used less frequently, although data collection and analysis is easier with arrays than with sequence data.

Unlike array hybridization which utilizes a fluorescence signal that may be noisy, sequencing-based measurements can directly quantify methylation levels. For instance, MeDIP-seq (methylated DNA immunoprecipitation-sequencing) takes advantage of an antibody specific to methylated DNA to isolate and sequence the methylated component of the genome (Jacinto et al., 2008; Harris et al., 2010). Alternatively, methylation-sensitive restriction enzymes may be used to eliminate unmethylated fragments, enriching for methylated DNA in the precipitated DNA fragments (Ruike et al., 2010; Taiwo et al., 2012; Vining et al., 2012). Both of these sequencing methods may have biases due to uneven enrichment. In BS-seq, there is no enrichment for methylated DNA − total genomic DNA is treated, captured, and sequenced. In a genome of mixed methylated and unmethylated DNA, sodium bisulfite converts only unmethylated cytosines into uracils, and thus by comparison of sequenced, converted reads to a reference genome, methylation levels are measured as the frequency of unconverted cytosines to converted cytosines, providing single-nucleotide resolution (Cokus et al., 2008; Lister et al., 2008). BS-seq experiments have fewer steps and better resolution compared with hybridization-based experiments. And generally, methods based on hybridization or affinity enrichment have inherent issues of bias, lower genome coverage and variation due to sample handling. Thus BS-seq has become the method of choice for genome-wide analysis of DNA methylation.

Promising technologies for direct detection of methylated bases have started to appear. For example, Pacific Biosciences "SMRT" sequencing can directly detect 5mC without bisulfite conversion (Flusberg et al., 2010). However, the method still needs improvements to distinguish 5mC and 5-hydroxymethylcytosine (5hmC) in animals, as well as to improve genome coverage *via* better throughput (Huang et al., 2010; Booth et al., 2012). But in the absence of mainstream availability of direct detection methods, our lab adopted bisulfite sequencing for DNA methylation analysis.