Bioinformatics, 37(24), 2021, 4857–4859
doi: 10.1093/bioinformatics/btab438
Advance Access Publication Date: 14 June 2021
Applications Note

Genome analysis

Methylscaper: an R/Shiny app for joint visualization of DNA methylation and nucleosome occupancy in single-molecule and single-cell data

Parker Knight[1], Marie-Pierre L. Gauthier[2], Carolina E. Pardo[2], Russell P. Darst[2], Kevin Kapadia[3], Hadley Browder[3], Eliza Morton[3], Alberto Riva [4], Michael P. Kladde[2] and Rhonda Bacher [1,*]

*To whom correspondence should be addressed.
Associate Editor: Peter Robinson

Received on November 13, 2020; revised on April 19, 2021; editorial decision on June 5, 2021; accepted on June 11, 2021

Abstract

Summary: Differential DNA methylation and chromatin accessibility are associated with disease development, particularly cancer. Methods that allow profiling of these epigenetic mechanisms in the same reaction and at the single-molecule or single-cell level continue to emerge. However, a challenge lies in jointly visualizing and analyzing the heterogeneous nature of the data and extracting regulatory insight. Here, we present methylscaper, a visualization framework for simultaneous analysis of DNA methylation and chromatin accessibility landscapes. Methylscaper implements a weighted principal component analysis that orders DNA molecules, each providing a record of the chromatin state of one epiallele, and reveals patterns of nucleosome positioning, transcription factor occupancy, and DNA methylation. We demonstrate methylscaper's utility on a long-read, single-molecule methyltransferase accessibility protocol for individual templates (MAPit-BGS) dataset and a single-cell nucleosome, methylation, and transcription sequencing (scNMT-seq) dataset. In comparison to other procedures, methylscaper is able to readily identify chromatin features that are biologically relevant to transcriptional status while scaling to larger datasets.

Availability and implementation: Methylscaper, is implemented in R (version $>$ 4.1) and available on Bioconductor: https://bioconductor.org/packages/methylscaper/, GitHub: https://github.com/rhondabacher/methylscaper/, and Web: https://methylscaper.com.

Contact: rbacher@ufl.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Abnormal epigenetic changes are a key hallmark of cancer. Alterations in DNA methylation, including the co-occurrence of both hyper- and hypo-methylation of different regions of the genome, have been detected in nearly all cancer types (Kushwaha *et al.*, 2016; Orjuela *et al.*, 2020; Pérez *et al.*, 2018). In addition, both cancer- and tissue-specific differences exist in nucleosome positioning and occupancy, as well as transcription factor binding activity, which together determine chromatin accessibility (Corces *et al.*, 2018). However, profiling endogenous methylation and accessibility states separately ignores their complementary nature in regulating gene expression and, by definition, queries different sets of molecules (Portela *et al.*, 2013). To address this, assays such as MAPit-BGS (Pondugula and Kladde, 2008) and NOMe-seq (Kelly *et al.*, 2012) have been developed to simultaneously capture nucleosome occupancy and methylation states at single-molecule resolution. In both cases, chromatin accessibility is first probed by the methyltransferase M.CviPI (Xu, 1998), which methylates unprotected GC sites. Next, accessibility at GC sites and CG endogenous methylation are profiled by bisulfite (Darst *et al.*, 2010) or bisulfite-free enzymatic conversion (Schutsky *et al.*, 2018). After sequencing, the methylation signals of all cytosines are translated bioinformatically. Long-read sequencing is particularly advantageous to phase the co-occurrence of epigenetic features, e.g. multiple nucleosomes. Recently, an extension of NOMe-seq, nanoNOMe, made use of long-read nanopore sequencing and resolved long-range patterns along with individual DNA molecules (Lee *et al.*, 2018). Methods for simultaneously profiling accessibility and methylation have also been extended to single cells via the scNOMe-seq (Pott, 2017) and scNMT-seq (Clark *et al.*, 2018) techniques.

For MAPit-BGS and nanoNOMe, the long reads derive from contiguous single DNA molecules, while single-cell methods use short reads that are reconstructed into contiguous DNA molecules from individual cells. Both types of methods allow for discerning the heterogeneous nature of cellular DNA methylation and chromatin

structure. Bioinformatic software programs, such as Bismark (Krueger and Andrews, 2011), are first used to align the data; however, many analytical pipelines and downstream visualization tools fail to highlight the epigenetic variation in a useful way. Previously developed methods utilize the output from Bismark but are limited to a relatively small number of reads or provide summary plots rather than site-level data (Huang *et al.*, 2018; Wong *et al.*, 2016). Two other such visualization tools are the NOMePlot (Requena *et al.*, 2019) and MethylViewer (Pardo *et al.*, 2011) applications, which were designed to simultaneously visualize CG methylation/GC accessibility patterns. Despite their integrated pipelines, the commonly used 'lollipop' plots are not intuitive in highlighting the joint occupancy and methylation states along a continuous DNA strand, especially when considering hundreds or thousands of molecules. The previously developed MethylTracker (Darst *et al.*, 2012) plots visually intuitive methylation/accessibility patterns by connecting consecutively methylated or unmethylated sites with contrasting colors, however, it is computationally inefficient and unable to effectively organize hundreds or thousands of reads.

Here, we describe methylscaper, a bioinformatic and statistical software package that generates visualizations of the DNA methylation and chromatin accessibility patterns. For single-molecule joint profiling data, or those using targeted sequencing approaches, methylscaper begins by processing raw sequencing reads. For single-cell data, or those using genome-wide approaches, output from Bismark or similar alignment programs is used as the initial input. Ordering the molecules is a key step for visualization, and our pipeline implements a two-stage weighted principal component analysis (PCA) framework that is feature- and site-specific. Weighting allows the user to emphasize specific genomic regions or features of interest. Compared to alternative procedures, our ordering is also efficient for large-scale datasets. Methylscaper is an interactive visualization platform available as a R/Shiny application and its functions may also be used directly via the R package. We evaluate methylscaper on an epigenetic DNA resilencing MAPit-BGS dataset and demonstrate its superior ability to elucidate epigenetic patterns and identify regions of cell-to-cell nucleosome sliding. We further demonstrate methylscaper on a single-cell dataset generated using scNMT-seq and identify a site of nucleosome positioning.

2 Materials and methods

Methylscaper first processes the data, followed by visualization and statistical analysis of methylated and accessible chromatin regions. For targeted sequencing datasets, the initial preprocessing steps include pairwise alignment of each sequence, quality control and filtering of poorly aligned sequences, and finally, conversion of the aligned sequences to plots of methylation and occupancy states (Fig. 1A). For genome-wide datasets, methylscaper begins with processed output from alignment programs such as Bismark. Additional details on the bioinformatic processing are available in Supplementary Materials. Regions of methylation or accessibility are identified by connecting consecutive sites having the same methylation state (Fig. 1B). A patch of endogenous methylation is plotted in red if ≥2 consecutive HCG sites show methylation (H = A, T or C, where a sequenced C [or G on the complementary strand] denotes methylation). Similarly, consecutive GCH methylation indicates accessibility, plotted in yellow. By contrast, consecutively unmethylated GCH or HCG [a sequenced T (or A on the complementary strand) in the sequence denotes an unmethylated status] are colored black. Patches of either color interrupted by a single GCH or HCG site of the opposite methylation state are emphasized as gray borders.

Ordering the single-cells or molecules displays population heterogeneity and allows identification of patterns of endogenous methylation as well as transcription factor and nucleosome occupancy. To do so, methylscaper constructs a matrix containing both endogenous (HCG) and introduced (GCH) methylation states for the set of molecules. A numerical key is used to represent patches of methylation. Weighted PCA is performed on the entire matrix, with molecules assigned a weight based on the number of methylation patches between two fixed base pairs chosen by the user. This allows

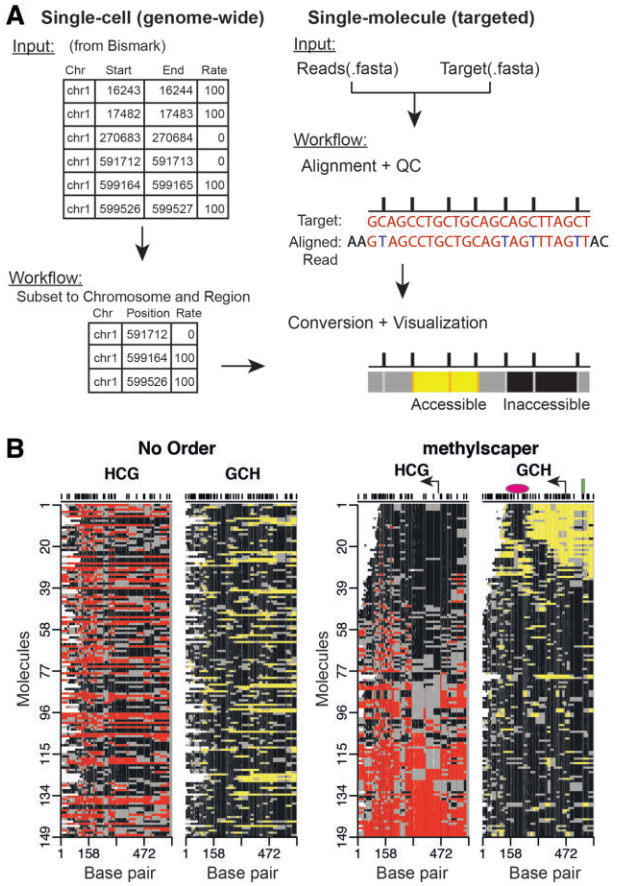

Fig. 1. An overview of methylscaper. (**A**) Flowchart of the bioinformatic preprocessing pipeline. (**B**) methylscaper plots of the MAPit-BGS data, generated with two different orderings. The data in the left plot is not ordered; the data on the right was ordered with methylscaper's weighted principal component algorithm. A pink oval was added to indicate the ~150-bp +1 nucleosome downstream of the transcription start site; a green rectangle was added to indicate a sequence-specific DNA-binding factor; and a bent arrow was added to indicate the TSS

the weighting to focus on either type of methylation and to emphasize specific genomic regions (Supplementary Fig. S1). The first weighted principal component is used to determine the global order; as shown in Supplementary Figure S2, where the first component is highly correlated with methylation and accessibility.

Following the determination of the global ordering, users can perform an optional second-stage refinement step in which a contiguous subset of the molecules is reordered using the PCA procedure to increase the resolution of patterns (Supplementary Fig. S3). Additional experiment-wide statistics are also calculated from the molecules that are then comparable across datasets or treatments (Supplementary Fig. S4).

3 Results

We applied methylscaper to a dataset with 149 single-molecule reads generated using MAPit-BGS. This dataset is from an epigenetic study of methylation resilencing in the *EMP2AIP1* promoter region following withdrawal of the DNA methyltransferase inhibitor 5-aza-2′-deoxycytidine using cell line RKO. A comparison of the visualization without any ordering versus our weighted PCA is shown in Fig. 1B. Without any ordering of the molecules (left panel), drawing biological conclusions is precluded. Using methylscaper (right panel), it becomes evident that endogenous HCG methylation (red-gray) inversely correlates with GCH accessibility (yellow-gray). The quality of ordering also allows visualization of the +1 nucleosome sliding and occupying different positions in each cell–the ~150 bp

footprints (black areas) that move in register with expansion/shortening of the accessible nucleosome-free region at the transcription start site (TSS). In molecules 25–35, two-phased nucleosomes are observed, punctuated by an accessible linker. Finally, protection of two GCH sites upstream of the TSS and within the nucleosome-free region detects binding of a sequence-specific transcription factor.

We also compared visualization with methylscaper to existing tools. In previous manuscripts using MAPit-BGS, hierarchical clustering alone was used to order the molecules. However, we have found this method fails with increasing complexity of patterns and number of molecules and often breaks the molecules into distinct blocks that have locally optimal orderings but are out of order with respect to a global structure (Supplementary Fig. S5). When patterns in the data are heterogeneous and many molecules are available, this leads to unorganized and potentially uninformative visualizations. Line plots, also commonly used to visualize methylation and accessibility status [e.g. as implemented in the aaRon R package (Statham *et al.*, 2015) or the NOMePlot software(Requena *et al.*, 2019)] either present the status of a single molecule at a time or of a moving population average of statuses across all molecules (Supplementary Fig. S6). This type of plot is insufficient when visualizing a large number of molecules, as using population averages often leads to a loss of critical information when methylation status or nucleosome occupancy is highly variable in heterogeneous cell populations. Commonly used lollipop plots also become unclear when a large number of molecules are available (Supplementary Fig. S6).

Next, we applied our results to a single-cell dataset generated using the scNMT-seq protocol that jointly profiles methylation and accessibility chromatin states in single cells (Clark *et al.*, 2018). As shown in Clark et al., we also observe high levels of open chromatin near the TSS for *Eef1g*, though we find evidence of a + 1 nucleosome approximately +250 bp downstream of the TSS (Supplementary Fig. S7).

Data Availability

The MAPit-BGS reads and reference sequence are available with this article as Supplementary data and available in the methylscaper R package on Github (https://github.com/rhondabacher/methylscaper/). The scNMT-seq dataset was downloaded from GSE109262.

Funding

This work was supported by the University of Florida Health Cancer Center; the National Institutes of Health [R01 CA155390 to M.P.K.] and the Defense Threat Reduction Agency [HDTRA1-16-1-0048].

Conflict of Interest: none declared.

References

Clark,S.J. *et al.* (2018) scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.*, **9**, 781.

Corces,M.R. *et al.*; The Cancer Genome Atlas Analysis Network. (2018) The chromatin accessibility landscape of primary human cancers. *Science*, **362**, eaav1898.

Darst,R.P. *et al.* (2010) Bisulfite Sequencing of DNA. *Curr. Protoc. Mol. Biol.*, **91**, 7.9.1–7.9.17.

.Darst, R.P. et al. (2012) DNA methyltransferase accessibility protocol for individual templates by deep sequencing. In: Wu,C and David, A (eds.) *Methods in Enzymology*. Elsevier, Academic Press, 513, pp. 185–204.

Huang,X. *et al.* (2018) ViewBS: a powerful toolkit for visualization of high-throughput bisulfite sequencing data. *Bioinformatics*, **34**, 708–709.

Kelly,T.K. *et al.* (2012) Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.*, **22**, 2497–2506.

Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.

Kushwaha,G. *et al.* (2016) Hypomethylation coordinates antagonistically with hypermethylation in cancer development: a case study of leukemia. *Hum. Genomics*, **10**, 18.

Lee,I. *et al.* (2018) Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Genomics*, **36**, 338–345.

Orjuela,S. *et al.* (2020) The DNA hypermethylation phenotype of colorectal cancer liver metastases resembles that of the primary colorectal cancers. *BMC Cancer*, **20**, 290.

Pardo,C.E. *et al.* (2011) MethylViewer: computational analysis and editing for bisulfite sequencing and methyltransferase accessibility protocol for individual templates (MAPit) projects. *Nucleic Acids Res.*, **39**, e5.

Pérez,R.F. *et al.* (2018) Distinct chromatin signatures of DNA hypomethylation in aging and cancer. *Aging Cell*, **17**, e12744.

Pondugula,S. and Kladde,M.P. (2008) Single-molecule analysis of chromatin: changing the view of genomes one molecule at a time. *J. Cell. Biochem.*, **105**, 330–337.

Portela,A. *et al.* (2013) DNA methylation determines nucleosome occupancy in the 5′-CpG islands of tumor suppressor genes. *Oncogene*, **32**, 5421–5428.

Pott,S. (2017) Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *eLife*, **6**, e23203.

Requena,F. *et al.* (2019) NOMePlot: analysis of DNA methylation and nucleosome occupancy at the single molecule. *Sci. Rep.*, **9**, 8140.

Schutsky,E.K. *et al.* (2018) Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nat. Biotechnol.*, **36**, 1083–1090.

Statham,A.L. *et al.* (2015) Genome-wide nucleosome occupancy and DNA methylation profiling of four human cell lines. *Genomics Data*, **3**, 94–96.

Wong,N.C. *et al.* (2016) MethPat: a tool for the analysis and visualisation of complex methylation patterns obtained by massively parallel sequencing. *BMC Bioinformatics*, **17**, 98.

Xu,M. *et al.* (1998) Cloning, characterization and expression of the gene coding for a cytosine-5-DNA methyltransferase recognizing GpC. *Nucleic Acids Res.*, **26**, 3961–3966.