

# 1 **methylscaper: an R/Shiny app for joint visualization of DNA methylation and** 2 **nucleosome occupancy in single-molecule and single-cell data**

3  
4 Parker Knight<sup>1</sup>, Marie-Pierre L. Gauthier<sup>2</sup>, Carolina E. Pardo<sup>2</sup>, Russell P. Darst<sup>2</sup>, Alberto Riva<sup>3</sup>,  
5 Michael P. Kladde<sup>2</sup>, Rhonda Bacher<sup>1\*</sup>

6  
7 <sup>1</sup>Department of Biostatistics, University of Florida, Gainesville, FL

8 <sup>2</sup>Department of Biochemistry and Molecular Biology, University of Florida, Gainesville, FL

9 <sup>3</sup>Bioinformatics Core, Interdisciplinary Center for Biotechnology Research, University of Florida,  
10 Gainesville, Florida

11 \*To whom correspondence should be addressed.

## 12 13 **Abstract**

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

27 **Availability and implementation:** Methylscaper, is available on GitHub at  
28 <https://github.com/rhondabacher/methylscaper>.

29 **Contact:** rbacher@ufl.edu  
30  
31

## 32 **Introduction**

33 Abnormal epigenetic changes are a key hallmark of cancer. Alterations in DNA methylation,  
34 including the co-occurrence of both hyper- and hypo-methylation of different regions of the  
35 genome, have been detected in nearly all cancer types(1–3). Additionally, both cancer- and  
36 tissue-specific differences exist in nucleosome positioning and occupancy, as well as  
37 transcription factor binding activity, which determine chromatin accessibility (4). However,  
38 profiling endogenous methylation and accessibility states separately ignores their complementary  
39 nature in regulating gene expression and, by definition, queries different sets of molecules (5).  
40 To address this, assays such as MAPit-BGS(6) and NOME-seq(7) that simultaneously capture  
41 nucleosome occupancy and methylation states at single-molecule resolution have been developed.  
42 In both cases, chromatin accessibility is first probed by the methyltransferase M.CviPI(8), which  
43 methylates unprotected GC sites. Next, accessibility at GC sites and CG endogenous methylation  
44 are profiled by bisulfite(9) or bisulfite-free enzymatic conversion(10). After sequencing, the  
45 methylation signals of all cytosines are translated bioinformatically. Long-read sequencing is  
46 particularly advantageous to phase the co-occurrence of epigenetic features, e.g., multiple  
47 nucleosomes. Recently, an extension of NOME-seq, nanoNOME, made use of long-read  
48 nanopore sequencing and resolved long-range patterns along individual DNA molecules(11).  
49 Methods for simultaneously profiling accessibility and methylation have also been extended to  
50 single cells via the scNOME-seq(12) and scNMT-seq(13) techniques.

51 For MAPit and nanoNOME, the long reads derive from contiguous single DNA  
52 molecules, while single-cell methods use short reads that are reconstructed into contiguous DNA  
53 molecules from individual cells. Both types of methods allow for discerning the heterogeneous  
54 nature of cellular DNA methylation and chromatin structure. Bioinformatic software programs,  
55 such as Bismark(14), are used to align the data; however, many analytical pipelines and  
56 downstream visualization tools fail to highlight the epigenetic variation in a useful way.  
57 Previously developed methods utilize the output from Bismark but are limited to a relatively  
58 small number of reads or provide summary plots rather than site-level data(15,16). Two other  
59 such visualization tools are the NOMEPlot(17) and MethylViewer(18) applications, which were  
60 designed to simultaneously visualize CG methylation/GC accessibility patterns. Despite their  
61 integrated pipelines, the commonly used ‘lollipop’ plots are not intuitive in highlighting the joint  
62 occupancy and methylation states along a continuous DNA strand, especially when considering

63 hundreds or thousands of molecules. The previously developed MethylTracker(19) plots visually  
64 intuitive methylation/accessibility patterns by connecting consecutively methylated or  
65 unmethylated sites with contrasting colors, however it is computationally inefficient and unable  
66 to effectively organize hundreds of reads.

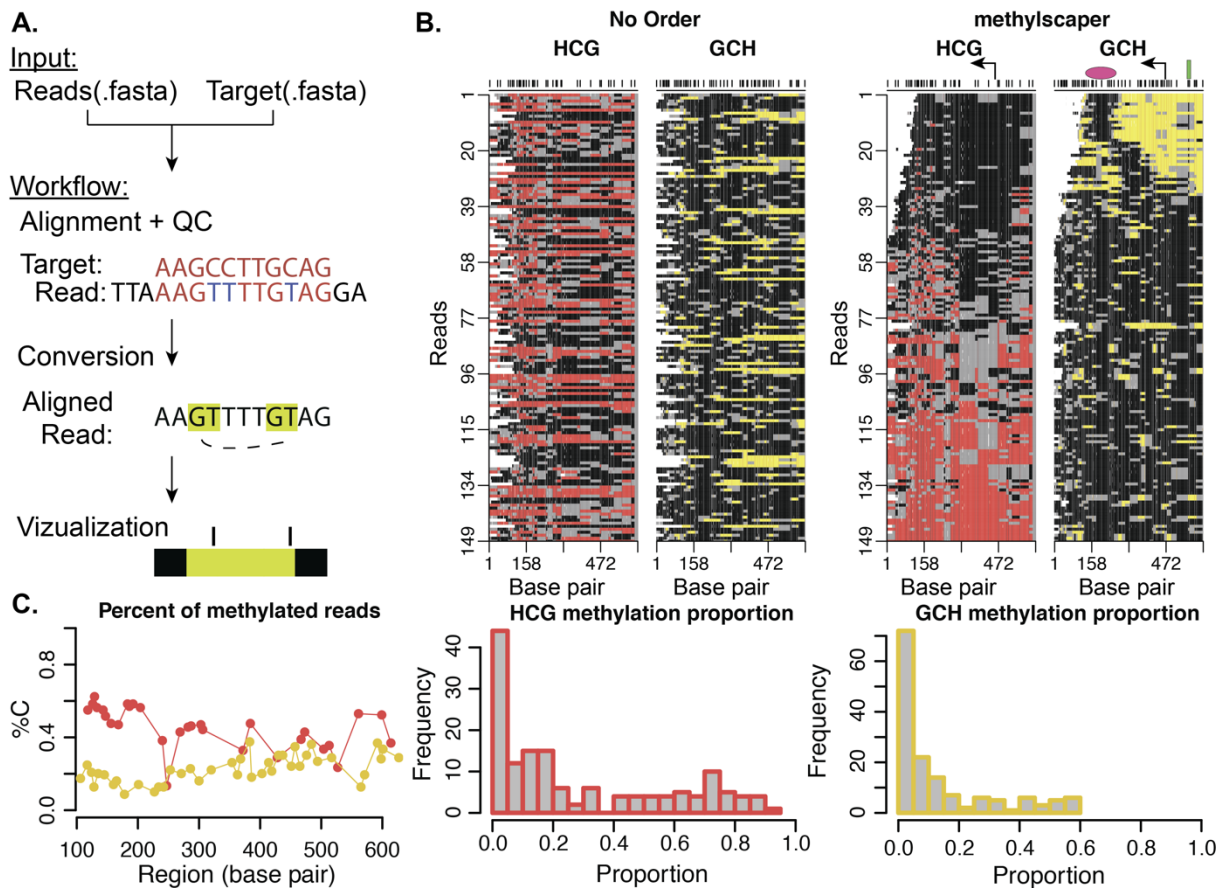
67 Here, we describe methylscaper, a bioinformatic and statistical software package that  
68 processes raw sequencing reads and generates visualizations of the DNA methylation and  
69 chromatin accessibility patterns. Additionally, output from Bismark for single-cell joint profiling  
70 experiments may also be used as input. Ordering the reads is a key step for visualization and our  
71 pipeline implements a two-stage weighted principal component analysis (PCA) framework that  
72 is feature and site specific. Weighting allows the user to emphasize specific genomic regions or  
73 features of interest. Compared to alternative procedures, our ordering is also efficient for large-  
74 scale datasets. Methylscaper is an interactive visualization platform available as an R/Shiny  
75 application and may also be used directly via the R package. We evaluate methylscaper on an  
76 epigenetic DNA resilencing MAPit-BGS dataset and demonstrate its superior ability to elucidate  
77 epigenetic patterns. We further demonstrate methylscaper on a single-cell dataset generated  
78 using scNMT-seq and identify regions of cell-to-cell nucleosome sliding.

79

## 80 **Methods**

81 Methylscaper first performs pre-processing to align the raw sequence files, followed by  
82 visualization and statistical analysis of methylated and accessible chromatin regions. The initial  
83 pre-processing steps include pairwise alignment of each sequence, quality control and filtering of  
84 poorly aligned sequences, and finally, conversion of the aligned sequences to methylation and  
85 occupancy states (Figure 1A). Additional details on the bioinformatic processing are available in  
86 Supplementary Materials. Regions of methylation or accessibility are identified by connecting  
87 consecutive sites having the same methylation state (Figure 2B). A patch of endogenous  
88 methylation is plotted in red if  $\geq 2$  consecutive HCG sites show methylation (C in the sequence).  
89 Similarly, consecutive GCH (H=A,T, or C; details in Supplementary Materials) methylation  
90 indicates accessibility, plotted in yellow. By contrast, consecutively unmethylated GCH or HCG  
91 (T in the sequence) are colored black. Patches of any color are interrupted by single GCH or  
92 HCG of the opposite methylation state, which are emphasized as gray borders.

93



94  
 95 Figure 1: An overview of methylscaper. A. Flowchart of the bioinformatic pre-processing  
 96 pipeline. B. methylscaper plots of the MAPit-BGS data, generated with two different orderings.  
 97 The data in the left plot is not ordered; the data on the right was ordered with methylscaper's  
 98 weighted principal component algorithm. A pink oval was added to indicate the ~150-bp +1  
 99 nucleosome downstream of the transcription start site; a green rectangle was added to indicate a  
 100 sequence-specific DNA-binding factor. C. Summary plots generated by methylscaper. Left: The  
 101 percentage of reads methylated at each base pair. Center: A histogram of the proportion of HCG  
 102 sites that are methylated in each read. Right: A histogram of the proportion of GCH sites that are  
 103 methylated in each read.

104  
 105 Methylscaper then orders the single-molecule reads to visualize heterogeneity, allowing  
 106 identification of patterns of endogenous methylation, as well as transcription factor and  
 107 nucleosome occupancy. Using a numerical key to represent patches of methylation patterns,  
 108 methylscaper constructs a matrix containing both endogenous (HCG) and induced (GCH)

109 methylation states for the set of reads. Weighted-PCA is performed on the entire matrix, with  
110 reads assigned a weight based on the number of methylation patches between two fixed base  
111 pairs chosen by the user. This allows the weighting to focus on either type of methylation and to  
112 emphasize specific genomic regions (Supplementary Figure 1). The first weighted principal  
113 component is used to determine the global read order; as shown in Supplementary Figure 2, the  
114 first component is highly correlated with methylation and accessibility.

115       Following the determination of the global ordering, users can perform an optional  
116 second-stage refinement step, in which a contiguous subset of the reads is reordered using the  
117 PCA procedure to increase the resolution of patterns (Supplementary Figure 3). Additional  
118 statistics are also calculated from the reads that are then comparable across datasets or treatments.  
119 In Figure 1C, methylscaper calculates experiment-wide statistics, such as the proportion of  
120 methylated sites at each base across all reads, as well as read-specific statistics including the  
121 proportion of each read that is GCH or HCG methylated.

122

## 123 **Results**

124       We applied methylscaper to a dataset with 149 single-molecule reads generated using  
125 MAPit-BGS. This dataset is from an epigenetic study of methylation resilencing in the  
126 *EMP2AIP1* promoter region following withdrawal of the DNA methyltransferase inhibitor 5-aza-  
127 2'-deoxycytidine using cell line RKO. A comparison of the visualization without any ordering  
128 versus our weighted PCA is shown in Figure 1B. Without any ordering of the read structure (left  
129 panel), drawing biological conclusions is precluded. Using methylscaper (right panel), it  
130 becomes evident that endogenous HCG methylation (red-gray) inversely correlates with GCH  
131 accessibility (yellow-gray). The quality of ordering also allows visualization of the +1  
132 nucleosome sliding across cells—the ~150 bp footprints (black areas) that move in register with  
133 expansion/shortening of the accessible nucleosome-free region at the transcription start site  
134 (TSS). In reads 25-35, two phased nucleosomes are observed, punctuated by an accessible linker.  
135 Finally, protection of two GCH sites upstream of the TSS and within the nucleosome-free region  
136 detects binding of a sequence-specific transcription factor.

137       We also compared visualization with methylscaper to existing tools. In previous  
138 manuscripts using MAPit-BGS, hierarchical clustering alone was used to order the reads, but we  
139 have found this method fails with increasing complexity of patterns and number of reads and

140 often breaks the reads into distinct blocks that have locally optimal orderings, but are out of  
141 order with respect to a global structure (Supplementary Figure 4). When patterns in the data are  
142 heterogeneous and many reads are available, this leads to unorganized and potentially  
143 uninformative visualizations. Line plots, also commonly used to visualize methylation and  
144 accessibility status (for example, as implemented in the aaRon R package(20) or the NOMePlot  
145 software), either present the status of a single read at a time or of a moving population average of  
146 statuses across all reads (Supplementary Figure 5). This type of plot is insufficient when  
147 visualizing a large number of reads, as using population averages often leads to a loss of critical  
148 information when methylation status or nucleosome occupancy is highly variable in  
149 heterogeneous cell populations. Commonly used lollipop plots also become unclear when a large  
150 number of reads are available (Supplementary Figure 5).

151 Next, we applied our results to a single-cell dataset generated using the scNMT-seq  
152 protocol that jointly profiles methylation and accessibility chromatin states in single cells(13). As  
153 shown in Clark et al., we also observe high levels of open chromatin near the TSS for *Eef1g*,  
154 though we find evidence of a +1 nucleosome approximately +250 bp downstream of the TSS  
155 (Supplementary Figure 6).

156

## 157 **References**

158

- 159 1. Kushwaha G, Dozmorov M, Wren JD, Qiu J, Shi H, Xu D. Hypomethylation coordinates  
160 antagonistically with hypermethylation in cancer development: a case study of leukemia.  
161 *Hum Genomics*. 2016 Jul;10(S2):18.
- 162 2. Pérez RF, Tejedor JR, Bayón GF, Fernández AF, Fraga MF. Distinct chromatin signatures of  
163 DNA hypomethylation in aging and cancer. *Aging Cell*. 2018 Jun;17(3):e12744.
- 164 3. Orjuela S, Menigatti M, Schraml P, Kambakamba P, Robinson MD, Marra G. The DNA  
165 hypermethylation phenotype of colorectal cancer liver metastases resembles that of the  
166 primary colorectal cancers. *BMC Cancer*. 2020 Dec;20(1):290.
- 167 4. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin  
168 accessibility landscape of primary human cancers. *Science*. 2018 Oct  
169 26;362(6413):eaav1898.
- 170 5. Portela A, Liz J, Nogales V, Setién F, Villanueva A, Esteller M. DNA methylation determines  
171 nucleosome occupancy in the 5'-CpG islands of tumor suppressor genes. *Oncogene*. 2013  
172 Nov;32(47):5421–8.

- 173 6. Pondugula S, Kladde MP. Single-molecule analysis of chromatin: Changing the view of  
174 genomes one molecule at a time. *J Cell Biochem.* 2008 Oct 1;105(2):330–7.
- 175 7. Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. Genome-wide mapping of nucleosome  
176 positioning and DNA methylation within individual DNA molecules. *Genome Research.* 2012  
177 Dec 1;22(12):2497–506.
- 178 8. Xu M. Cloning, characterization and expression of the gene coding for a cytosine-5-DNA  
179 methyltransferase recognizing GpC. *Nucleic Acids Research.* 1998 Sep 1;26(17):3961–6.
- 180 9. Darst RP, Pardo CE, Ai L, Brown KD, Kladde MP. Bisulfite Sequencing of DNA. *Current*  
181 *Protocols in Molecular Biology.* 2010 Jul;91(1):7.9.1-7.9.17.
- 182 10. Schutsky EK, DeNizio JE, Hu P, Liu MY, Nabel CS, Fabyanic EB, et al. Nondestructive, base-  
183 resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nat Biotechnol.*  
184 2018 Nov;36(11):1083–90.
- 185 11. Lee I, Razaghi R, Gilpatrick T, Molnar M, Sadowski N, Simpson JT, et al. Simultaneous  
186 profiling of chromatin accessibility and methylation on human cell lines with nanopore  
187 sequencing [Internet]. *Genomics*; 2018 Dec [cited 2020 Jun 17]. Available from:  
188 <http://biorxiv.org/lookup/doi/10.1101/504993>
- 189 12. Pott S. Simultaneous measurement of chromatin accessibility, DNA methylation, and  
190 nucleosome phasing in single cells. *eLife.* 2017 Jun 27;6:e23203.
- 191 13. Clark SJ, Argelaguet R, Kapourani C-A, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. scNMT-seq  
192 enables joint profiling of chromatin accessibility DNA methylation and transcription in single  
193 cells. *Nat Commun.* 2018 Dec;9(1):781.
- 194 14. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq  
195 applications. *Bioinformatics.* 2011 Jun 1;27(11):1571–2.
- 196 15. Huang X, Zhang S, Li K, Thimmapuram J, Xie S. ViewBS: a powerful toolkit for visualization of  
197 high-throughput bisulfite sequencing data. Wren J, editor. *Bioinformatics.* 2018 Feb  
198 15;34(4):708–9.
- 199 16. Wong NC, Pope BJ, Candiloro IL, Korbie D, Trau M, Wong SQ, et al. MethPat: a tool for the  
200 analysis and visualisation of complex methylation patterns obtained by massively parallel  
201 sequencing. *BMC Bioinformatics.* 2016 Dec;17(1):98.
- 202 17. Requena F, Asenjo HG, Barturen G, Martorell-Marugán J, Carmona-Sáez P, Landeira D.  
203 NOMePlot: analysis of DNA methylation and nucleosome occupancy at the single molecule.  
204 *Sci Rep.* 2019 Dec;9(1):8140.
- 205 18. Pardo CE, Carr IM, Hoffman CJ, Darst RP, Markham AF, Bonthron DT, et al. MethylViewer:  
206 computational analysis and editing for bisulfite sequencing and methyltransferase

- 207 accessibility protocol for individual templates (MAPit) projects. *Nucleic Acids Research*.  
208 2011 Jan;39(1):e5–e5.
- 209 19. Darst RP, Nabils NH, Pardo CE, Riva A, Kladd MP. DNA Methyltransferase Accessibility  
210 Protocol for Individual Templates by Deep Sequencing. In: *Methods in Enzymology*  
211 [Internet]. Elsevier; 2012 [cited 2020 Oct 4]. p. 185–204. Available from:  
212 <https://linkinghub.elsevier.com/retrieve/pii/B9780123919380000082>
- 213 20. Statham AL, Taberlay PC, Kelly TK, Jones PA, Clark SJ. Genome-wide nucleosome occupancy  
214 and DNA methylation profiling of four human cell lines. *Genomics Data*. 2015 Mar;3:94–6.
- 215