

Genome Analysis

OpTiles: an R package for adaptive tiling and methylation variability profiling

Giorgia Migliaccio¹, Lena Möbus^{1,2}, Giusy del Giudice^{1,3}, Jack Morikka^{1,2}, Antonio Federico^{1,3}, Angela Serra^{1,3}, Dario Greco^{1,3,*} 

¹Finnish Hub for Development and Validation of Integrated Approaches (FHAIVE), Faculty of Medicine and Health Technology, Tampere University, 33100 Tampere, Finland

²Tampere Institute for Advanced Study, Tampere University, 33520 Tampere, Finland

³Division of Pharmaceutical Biosciences, Faculty of Pharmacy, University of Helsinki, 00790 Helsinki, Finland

*Corresponding author. FHAIVE, Faculty of Medicine and Health Technology, Tampere University, Arvo Ylpön katu 34, 33520, Tampere, Finland; Division of Pharmaceutical Biosciences, Faculty of Pharmacy, University of Helsinki, 00790 Helsinki, Finland. E-mail: dario.greco@tuni.fi

Associate Editor: Peter Robinson

Abstract

Summary: OpTiles is an R package that dynamically defines tiling windows based on the distribution of sequenced CpGs, addressing the limitations of traditional fixed-tiling approaches in targeted methylation datasets. By integrating CpG density with intra-region methylation variability, it provides a reliability metric and extended functionality for annotating, prioritizing, and interpreting complex methylation data.

Availability and implementation: OpTiles is implemented in R and source code is freely available at <https://github.com/fhaive/OpTiles>. Data are available on Zenodo at <https://doi.org/10.5281/zenodo.16961292>.

1 Introduction

The advent of next-generation sequencing technologies has significantly advanced epigenomic research, enabling genome-wide profiling of DNA methylation at single-CpG resolution (Lister *et al.* 2009, Plongthongkum *et al.* 2014). However, despite this high-resolution capability, it remains standard practice to aggregate neighboring CpG sites into larger regions to capture broader methylation patterns and extract biologically meaningful signals (Akalın *et al.* 2012, Hansen *et al.* 2012, Li *et al.* 2013, Gaspar and Hart 2017, Piao *et al.* 2021). Region-based strategies have been developed to define regions from CpG-level patterns. Several methods, such as DSS (Feng *et al.* 2014) and bsseq (Hansen *et al.* 2012), model methylation as a continuous signal across the genome, using smoothing or hierarchical frameworks to capture spatially coherent methylation patterns across neighboring CpGs. Others, like metilene (Jühling *et al.* 2016) and DMRcate (Peters *et al.* 2015), identify and cluster differentially methylated CpGs based on proximity and methylation status. Other approaches as DMAP (Stockwell *et al.* 2014) define regions based on the actual DNA fragments obtained during library preparation, effectively anchoring methylation summaries to experimental fragment boundaries. The approaches described here represent only a subset of the many strategies reviewed in the literature (Chen *et al.* 2016, Shafi *et al.* 2018, Piao *et al.* 2021), where a variety of methodologies have been proposed to aggregate CpG sites into regions. However, no consensus has been reached on what constitutes the optimal strategy.

As noted by Shafi and colleagues (Shafi *et al.* 2018), one of the most widely cited approaches is implemented in methylKit (Akalın *et al.* 2012), which uses a tiling-window strategy to divide the genome into fixed-size segments and summarize the methylation signals of CpGs within each of them. While this approach offers systematic coverage it assumes a uniform distribution of CpGs and consistent sequencing coverage which limits its suitability for targeted approaches like Reduced Representation Bisulfite Sequencing (RRBS), where the CpG distribution is biased toward CpG-rich regions within the genome. As a result, fixed tiles may span poorly covered sites, potentially diluting signals and introducing noise. Filtering tiles by CpG density is therefore common practice, but when region boundaries are not aligned with the distribution of sequenced data, this filtering can inadvertently discard regions that still hold valuable information. Furthermore, the summarization approach used in methylKit (Akalın *et al.* 2012) reduces each region to a single average methylation value, without accounting for intra-region variability. Some methods treat CpG methylation consistency as a requirement for defining regions (Hansen *et al.* 2012, Shi *et al.* 2021, Balaramane *et al.* 2024), instead in the most straightforward genome-tiling approach methylation consistency within regions is overall ignored. Although colocalized CpGs are generally expected to exhibit coordinated methylation patterns (Eckhardt *et al.* 2006, Affinito *et al.* 2020), heterogeneity in methylation states is frequently observed (Elliott *et al.* 2015). Such variability can be influenced by technical factors, such as PCR bias, sequencing coverage,

different read lengths or stochastic erosion (Scherer *et al.* 2020). PCR amplification may preferentially enrich molecules with particular base compositions or methylation states, while incomplete bisulfite conversion, uneven sequencing coverage, or read-mapping errors can introduce spurious heterogeneity across CpGs (Scherer *et al.* 2020). However, methylation heterogeneity may also reflect variation in cell-type composition, allele or strand specific methylation, imprinted regions, or even transcriptional heterogeneity (Olova *et al.* 2018, Shi *et al.* 2021). By incorporating variability metrics into fixed-tiles, OpTiles adds this missing layer of context by enabling the identification of regions with high intra-region variability and guiding decisions on whether to summarize a genomic locus as a region or retain the finer resolution of individual analysis.

In this study, we introduce OpTiles, an R package built as a modular extension to the methylKit workflow (Akalin *et al.* 2012) which leverages its well-established tiling-window framework to enhance methylation region definition. While designed to integrate seamlessly with methylKit, OpTiles' core functions, such as optimization of tiles definition, intra-region variability assessment and annotations functions, require only genomic coordinates and CpG-level beta values as input, enabling the use outside the methylKit environment. Its core functions are modular and can operate independently, while auxiliary functions are also provided to convert data into methylKit format for running parts of its standard workflow (e.g. loading data or differential methylation analysis), facilitating direct integration into existing pipelines. The manual (on <https://github.com/fhaive/OpTiles>) explains in detail how the function works and what inputs are needed. Rather than applying uniform, fixed-size tiling across the genome, OpTiles refines the tiles *post hoc* by repositioning them based on the distribution of sequenced CpGs, ensuring tiles to be anchored in sequenced CpGs. Additionally, OpTiles provides functions to assess intra-region methylation variability, an often overlooked dimension, that can help users identify high variable regions, either technically unstable or biologically heterogeneous, supporting more informed filtering and prioritization decisions. Beyond tile optimization and variability assessment, OpTiles maps user-defined regions to genomic annotations and provides overlap metrics that consider both region size and CpG content. Together, these functions support a more informed and flexible interpretation of DNA methylation landscapes, complementing existing pipelines without altering their core analytical logic.

2 Optimization of tiles definition

OpTiles refines the definition of tiling window with a data-driven adaptive binning: starting from the tiling window genomic locations, OpTiles computes the number of CpGs within each tile (with `map_cpg_to_regions` function) and identifies pairs of consecutive tiles defined as directly adjacent, non-overlapping pairs. For each pair of candidate tiles, OpTiles maps all CpG positions and calculates two distances (Fig. 1A): (i) AD is the distance between the first CpG of the first tile and the last CpG of the second; (ii) BC is the distance between the last CpG of the first tile and the first CpG of the adjacent one. If CpGs are tightly clustered ($AD < \text{tile size}$), the tiles are merged into a single region spanning all CpGs from both tiles. If $AD > \text{tile size}$ but $BC < \text{tile size}$, indicating a narrow gap between CpG clusters, OpTiles scans the span

to generate fixed size windows and selects the one with the highest CpG count, retaining the most downstream in case of a tie. If neither AD nor BC meet these criteria, the tiles remain unchanged. This optimization is performed using the `merging_consecutive_regions` function, which takes as input the original CpGs genomic location matrix, the location of the tiles, and the regions length.

To illustrate the impact of our merging strategy, we analyzed an RRBS dataset from THP-1-derived macrophages (data available on Zenodo <https://doi.org/10.5281/zenodo.16961292>). THP-1 cells (ATCC TIB-202, USA) were cultured in RPMI-1640 (Gibco, USA) supplemented with 10% FBS (Gibco, USA) (culture media). Cells were cultured in 75-cm² flasks at a density $< 1 \times 10^6$ cells/ml. Cells were differentiated in 12 well plates, with 127 000 cells/cm², in the culture media supplemented with 30.9 ng/ml of phorbol 12-myristate 13-acetate (PMA) (Sigma-Aldrich, USA) for 48 h, followed by a 72-h recovery period in standard culture medium. The preprocessing of the RRBS data has been done using Bismark suite (Krueger and Andrews 2011) and methylKit pipeline (Akalin *et al.* 2012) (script available on <https://github.com/fhaive/OpTiles>), and regions produced by the standard methylKit tiling were compared with those from our optimized approach. As shown in Fig. 1B, the optimized tiles display a higher CpG density compared to the standard tiles, accompanied by a slight increase in intra-region variability, which is likely attributable to the greater number of CpGs included per region. This improvement is particularly useful given that many pipelines apply minimum-CpG filters to remove low-information regions (Akalin *et al.* 2012, Li *et al.* 2013, Dolzhenko and Smith 2014, Piao *et al.* 2021). For example, applying a > 5 CpG filter to our dataset removed 45.6% of standard tiles but only 22.3% of optimized tiles (Table 1, available as supplementary data at *Bioinformatics* online). By increasing CpG density per region, OpTiles preserves potentially informative regions that would otherwise be lost due to sparse coverage or uneven CpG distribution.

3 Intra-region variability

OpTiles evaluates intra-region methylation variability by calculating the standard deviation of methylation values within each region across individual samples or biological replicates, using the `compute_beta_sd_regions` function. The input data consists in CpG-level methylation values together with their corresponding mapped genomic regions and specify the sample group across which the variability should be computed. In general, regional methylation summaries are typically computed as averages of CpG methylation levels within a region (Akalin *et al.* 2012), some methods weight this average by read coverage to account for uneven sequencing depth (Hansen *et al.* 2012, Feng *et al.* 2014). We believe that, beyond calculating a region average methylation value, examining the variability of individual CpG sites can provide insight into both the consistency of the regional signal and biologically relevant heterogeneity.

Figure 1C–E illustrates how flagging this variability can provide an additional layer of insight. In Fig. 1C, a single fully non-methylated CpG site drives the high variability in this otherwise highly methylated region. This could result from genuine lack of methylation or technical factors. In this case, the site carries a known single-nucleotide polymorphism (SNP) in which the guanine following the cytosine is altered,

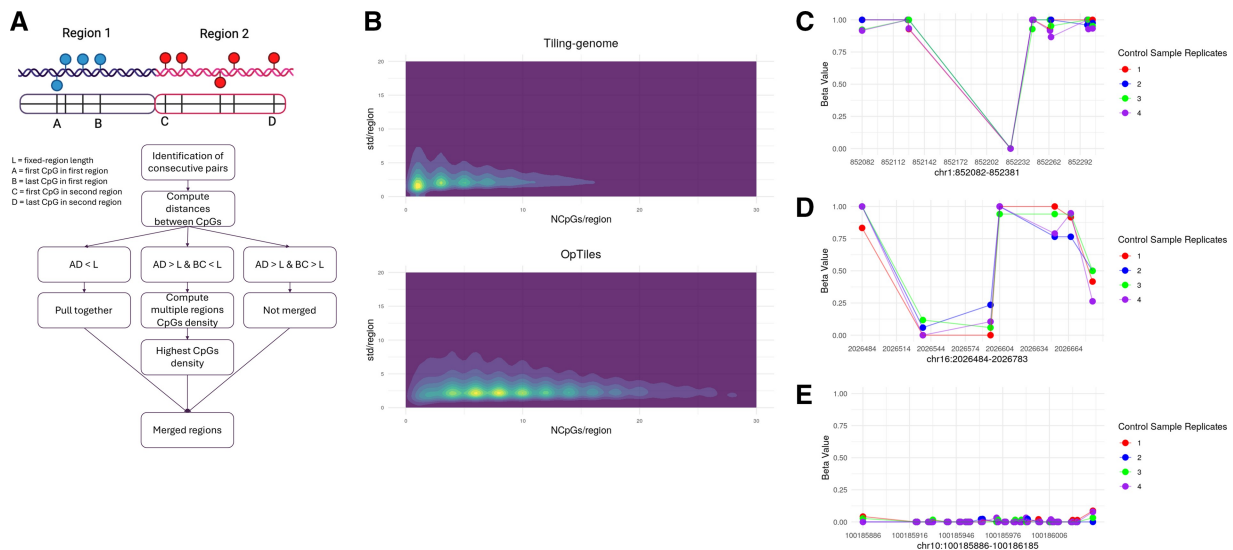


Figure 1. (A) Schematic overview of the workflow for optimizing tile definitions. Created with <https://BioRender.com>. (B) Comparison of average CpG density per region before and after applying the optimization strategy. The x-axis shows the average CpG density per region, while the y-axis represents the intra-region variability. (C–E) Intra-region methylation variability of three selected genomic regions.

thereby eliminating the CpG site and preventing methylation. If polymorphisms are not filtered at the outset, this approach allows case-by-case identification of variants that need to be removed. While removing all sites overlapping with common polymorphisms a common practice, it may not be ideal, especially when dealing with single cell types where SNP lists might be incomplete or not readily available (LaBarre *et al.* 2019). By flagging such regions with inconsistent methylation, the user remains informed about potential bias without the need of completely removing all species-level variation. Such regions can yield skewed mean methylation values, and the variability flag highlights them as candidates for closer examination or possible exclusion from downstream analyses.

In contrast, Fig. 1D illustrates heterogeneous methylation across the CpGs within the region, indicating a pattern that might reflect regulatory complexity. Notably, this pattern maps to a proximal enhancer-like signature, suggesting that such heterogeneity may be associated with regulatory activity, transcription factor binding variability, or context-specific gene expression. Recent work highlights that methylation heterogeneity can arise from the dynamic interplay between DNA methyltransferases (DNMT) and ten-eleven translocation (TET) mediated modification cycles, with intermediate patterns reflecting either gradual transitions or regulatory states stabilized by transcription factor binding (Shi *et al.* 2021). Such heterogeneity has been linked to context-specific gene expression and may provide more sensitive indicators of regulatory activity than average methylation levels alone (Lin *et al.* 2023). The flag highlights potentially meaningful heterogeneity that warrants further investigation; this variability cannot be generalized across the entire region and instead requires higher-resolution analysis to fully understand its functional implications. In Fig. 1E, the CpGs within the region display consistently similar methylation levels, indicating minimal intra-region variability. This region overlaps a CpG island (CpG43) located near the transcription start site (TSS) of the *ERLIN1* gene. The uniform demethylation pattern is consistent with the well-established tendency of CpG islands in promoter regions to remain unmethylated (Bird

2002, Deaton and Bird 2011). In such cases, summarizing the region with a single average methylation value is both appropriate and representative of its biological state.

This additional layer of information improves the interpretability of regional methylation patterns, especially in complex or heterogeneous datasets. By integrating intra-region variability with CpG density into a single scoring framework we define the InfoScore, calculated as the ratio between CpG density (number of CpGs divided by region length) and the standard deviation of methylation values within the region. With this score, OpTiles helps prioritizing regions that are both well-supported by data and internally consistent, while simultaneously flagging those with variability that could indicate underlying biological or technical factors. This enables more informed prioritization without prematurely excluding regions that may hold biological relevance.

4 Additional functionalities

Beyond tile optimization and intra-region variability assessment, OpTiles includes a suite of functions to map user-defined regions to genomic annotations. This is achieved through integration with biomaRt database (Durinck *et al.* 2009), using the `biomart_annotation` function, which requires a specified database and dataset, and accepts optional filters such as chromosomes, attributes, attribute page, gene biotype(s), and promoter distance. The function enables annotation of methylation regions relative to known genomic features (e.g. promoters, genes, or CpG islands if annotation file is provided, as shown in the manual available on <https://github.com/fhaive/OpTiles>). A key aspect of this mapping functionality is the introduction of two overlap metrics that provide more nuanced control over region-feature associations. First, OpTiles computes the percentage of base-pair overlap between a given region and the target genomic feature. Second, it quantifies how many CpGs within the region fall inside the overlapping window. These metrics offer complementary views of region relevance and can be used as customizable filters to refine annotation outputs. This allows

users to prioritize overlaps that are not only spatially extensive but also rich in informative CpGs, thereby supporting more tailored and interpretable downstream analyses.

5 Conclusion

OpTiles extends the standard tiling-windows approach by refining region definitions, supporting data-driven prioritization, and enhancing interpretation of heterogeneous or complex datasets. To improve regions definition in DNA methylation analysis, OpTiles implements a data-adaptive framework that considers the non-uniform nature of CpG distribution and sequencing coverage. By repositioning tiling windows to better reflect the distribution of sequenced CpGs, OpTiles preserves region size while increasing the retention of informative CpGs that might otherwise be excluded by genome fixed-window strategies. Beyond optimizing tile boundaries, OpTiles incorporates intra-region variability as an additional metric to assess the reliability and biological relevance of regional methylation summaries. Variability across CpGs within a region can arise from technical noise, cell-line specific SNPs, other sources of (epi)genetic variation, all of which are important to recognize. Rather than being treated as variation, OpTiles quantifies it alongside CpG density to offer a composite score, the InfoScore, designed to support informed region selection. This score helps to prioritize regions where methylation values are likely to be more stable and representative, and others that are instead more variable and inconsistent in methylation value that might need more careful evaluation. In parallel, the annotation and overlap-scoring functions provided by OpTiles offer researchers a finer-grained toolkit for contextualizing methylation patterns. This is especially valuable in studies aiming to link methylation changes to regulatory regions, where the extent and density of overlap carry different but complementary implications. Rather than replacing existing workflows, OpTiles is designed to complement and enhance them by introducing a data-driven, flexible approach to region definition and interpretation. Overall, it complements existing pipelines by refining region definition, enhancing data quality, and improving the biological relevance of epigenomic analyses.

Author contributions

Giorgia Migliaccio (Conceptualization [equal], Data curation [lead], Formal analysis [lead], Methodology [equal], Software [lead], Validation [equal], Visualization [lead], Writing—original draft [lead], Writing—review & editing [equal]), Lena Möbus (Conceptualization [equal], Formal analysis [equal], Methodology [equal], Supervision [lead], Writing—original draft [equal], Writing—review & editing [equal]), Giusy del Giudice (Conceptualization [equal], Supervision [equal], Writing—review & editing [equal]), Jack Morikka (Conceptualization [equal], Supervision [equal], Writing—review & editing [equal]), Antonio Federico (Supervision [equal], Writing—review & editing [equal]), Angela Serra (Supervision [equal], Writing—review & editing [equal]), and Dario Greco (Conceptualization [equal], Funding acquisition [lead], Project administration [lead], Resources [lead], Supervision [equal], Writing—review & editing [equal])

Supplementary data

Supplementary data is available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by the European Research Council (ERC) program, Consolidator project “ARCHIMEDES” [grant agreement number 101043848]. L.M. and J.M. were supported by the Tampere Institute for Advanced Study (IAS). A.F. was supported by the Faculty of Pharmacy, University of Helsinki.

Data availability

The package, the manual, the example script, and analysis scripts are available at <https://github.com/fhaive/OpTiles>, which also contains the documentation and the installation instructions. The datasets used for the analysis, along with the corresponding results are available on Zenodo at <https://doi.org/10.5281/zenodo.16961292>.

References

- Affinito O, Palumbo D, Fierro A *et al.* Nucleotide distance influences co-methylation between nearby CpG sites. *Genomics* 2020;112:144–50.
- Akalin A, Kormaksson M, Li S *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* 2012;13:R87.
- Balaramane D, Spill YG, Weber M *et al.* MethyLasso: a segmentation approach to analyze DNA methylation patterns and identify differentially methylated regions from whole-genome datasets. *Nucleic Acids Res* 2024;52:e98.
- Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev* 2002;16:6–21.
- Chen D-P, Lin Y-C, Fann CSJ. Methods for identifying differentially methylated regions for sequence- and array-based data. *Brief Funct Genomics* 2016;15:485–90.
- Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Dev* 2011;25:1010–22.
- Dolzhenko E, Smith AD. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics* 2014;15:215.
- Durinck S, Spellman PT, Birney E *et al.* Mapping identifiers for the integration of genomic datasets with the R/bioconductor package biomaRt. *Nat Protoc* 2009;4:1184–91.
- Eckhardt F, Lewin J, Cortese R *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* 2006;38:1378–85.
- Elliott G, Hong C, Xing X *et al.* Intermediate DNA methylation is a conserved signature of genome regulation. *Nat Commun* 2015;6:6363.
- Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res* 2014;42:e69.
- Gaspar JM, Hart RP. DMRfinder: efficiently identifying differentially methylated regions from MethylC-Seq data. *BMC Bioinformatics* 2017;18:528.
- Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* 2012;13:R83.
- Jühling F, Kretzmer H, Bernhart SH *et al.* Metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res* 2016;26:256–62.
- Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 2011;27:1571–2.

- LaBarre BA, Goncarenco A, Petrykowska HM *et al.* MethylToSNP: identifying SNPs in illumina DNA methylation array data. *Epigenetics Chromatin* 2019;**12**:79.
- Li S, Garrett-Bakelman FE, Akalin A *et al.* An optimized algorithm for detecting and annotating regional differential methylation. *BMC Bioinformatics* 2013;**14**:S10.
- Lin P-Y, Chang Y-T, Huang Y-C *et al.* Estimating genome-wide DNA methylation heterogeneity with methylation patterns. *Epigenet Chromatin* 2023;**16**:44.
- Lister R, Pelizzola M, Downen RH *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;**462**:315–22.
- Olova N, Krueger F, Andrews S *et al.* Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol* 2018;**19**:33.
- Peters TJ, Buckley MJ, Statham AL *et al.* De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin* 2015;**8**:6.
- Piao Y, Xu W, Park KH *et al.* Comprehensive evaluation of differential methylation analysis methods for bisulfite sequencing data. *Int J Environ Res Public Health* 2021;**18**:7975.
- Plongthongkum N, Diep DH, Zhang K. Advances in the profiling of DNA modifications: cytosine methylation and Beyond. *Nat Rev Genet* 2014;**15**:647–61.
- Scherer M, Nebel A, Franke A *et al.* Quantitative comparison of within-sample heterogeneity scores for DNA methylation data. *Nucleic Acids Res* 2020;**48**:e46.
- Shafi A, Mitrea C, Nguyen T *et al.* A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Brief Bioinf* 2018;**19**:737–53.
- Shi J, Xu J, Chen YE *et al.* The concurrence of DNA methylation and demethylation is associated with transcription regulation. *Nat Commun* 2021;**12**:5285.
- Stockwell PA, Chatterjee A, Rodger EJ *et al.* DMAP: differential methylation analysis package for RRBS and WGBS data. *Bioinformatics* 2014;**30**:1814–22.