Bioinformatics, 36(22-23), 2020, 5524–5525
doi: 10.1093/bioinformatics/btaa1048
Advance Access Publication Date: 25 December 2020
Applications Note

Genome analysis

Methrix: an R/Bioconductor package for systematic aggregation and analysis of bisulfite sequencing data

Anand Mayakonda [1,2], Maximilian Schönung [2,3], Joschka Hey[1,2,4], Rajbir Nath Batra[1], Clarissa Feuerstein-Akgoz[1,2], Kristin Köhler[5], Daniel B. Lipka[3], Rocio Sotillo[6], Christoph Plass[1,7], Pavlo Lutsik [1] and Reka Toth[1,6,*]

[1]Division of Cancer Epigenomics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany,, [2]Faculty of Biosciences, Heidelberg University, 69117 Heidelberg, Germany, [3]Section Translational Cancer Epigenomics, Division of Translational Medical Oncology, German Cancer Research Center (DKFZ) & National Center for Tumor Diseases (NCT), 69120 Heidelberg, Germany, [4]German-Israeli Helmholtz Research School in Cancer Biology, partner site Heidelberg, Germany, [5]Bioinformatics Bachelor Program, Free University Berlin, 14195 Berlin, Germany, [6]Division of Molecular Thoracic Oncology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany and [7]German Cancer Research Consortium (DKTK), partner site Heidelberg, Germany

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia
Received on June 8, 2020; revised on October 20, 2020; editorial decision on December 6, 2020; accepted on December 8, 2020

Abstract

Motivation: Whole-genome bisulfite sequencing (WGBS) measures DNA methylation at base pair resolution resulting in large bedGraph like coverage files. Current options for processing such files are hindered by discrepancies in file format specification, speed, and memory requirements.

Results: We developed *methrix*, an R package, which provides a toolset for systematic analysis of large datasets. Core functionality of the package includes a comprehensive bedGraph or similar tab-separated text file reader—which summarizes methylation calls based on annotated reference indices, infers and collapses strands and handles uncovered reference CpG sites while facilitating a flexible input file format specification. Additional optimized functions for quality control filtering, subsetting and visualization allow user-friendly and effective processing of WGBS results. Easy integration with tools for differentially methylated region (DMR) calling and annotation further eases the analysis of genome-wide methylation data. Overall, *methrix* enriches established WGBS workflows by bringing together computational efficiency and versatile functionality.

Availability and implementation: *Methrix* is implemented as an R package, made available under MIT license at https://github.com/CompEpigen/methrix and can be installed from the Bioconductor repository.

Contact: r.toth@dkfz-heidelberg.de

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

1 Introduction

DNA methylation is an epigenetic modification associated with transcriptional regulation and cellular identity. Next-generation sequencing assays, such as whole-genome bisulfite sequencing (WGBS), measure DNA methylation at base pair resolution. Several processing tools have been developed resulting in large bedGraph-like files with slightly different formats. Downstream processing involves aggregation of these files into coverage and methylation matrices, the size of which rapidly increase along with the sample size. Current options, such as *bsseq*, *methylKit* and *RnBeads* implemented in R environment for processing such output are hindered by either file format restrictions and limited functionality (Hansen *et al.*, 2012) or speed and memory requirements (Akalin *et al.*, 2012; Assenov *et al.*, 2014). The subsequent steps such as filtering and summarizing for the seamless integration with other tools, if not optimized, can be time and memory consuming. To address these limitations, we present *methrix*, an R package which combines computational performance, input format flexibility and user-friendly functions for WGBS data processing.

2 Package overview

Methrix exploits the *data.table* R package for faster data import, aggregation, and binary search operations, which are critical for querying large genomic datasets. The resulting *methrix* object is an extension of Bioconductor *SummarizedExperiment* container and

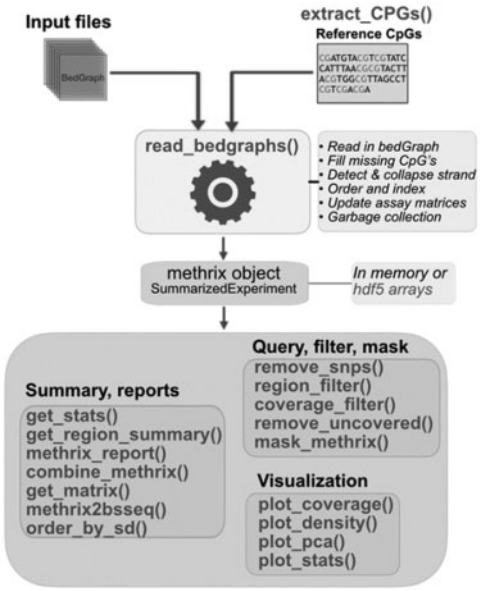

Fig. 1. Package overview. Data processing using the *methrix* package begins with the main function *read_bedgraphs()* which accepts bedGraph-like coverage files along with genome-wide reference CpG loci as input. *Methrix* aggregates all input files into a single *methrix* object which can be passed to several downstream functions for quality filtering, summary reports and visualization

support in-memory as well as on-disk HDF5-backed arrays. A complete overview of the package structure and usage is depicted in Figure 1.

2.1 Input

Methrix accepts bedGraph-like coverage files as input, which are imported via the *read_bedgraphs()* function and handles different input formats, omitted reference CpGs and optionally infers/collapses strand. Pre-configurated settings for importing the outputs from the most commonly used methylation callers are also available namely, Bismark (Krueger and Andrews, 2011), MethylDackel, methylCtools, BisSNP (Liu *et al.*, 2012), and BSseeker2 (Guo *et al.*, 2013).

2.2 Analysis modules

Functions implemented in *methrix* can be divided into three main modules: filtering, summarization and reports and visualization (Fig. 1).

Filtering: Functions in this module are primarily related to quality control (QC) and filtering steps. In addition to coverage-based filtering, single nucleotide polymorphism filtering (*remove_snps()*) and coverage masking functions (*mask_methrix()*) are implemented.

Summarization and reports: *methrix_report()* function generates a sharable interactive HTML-report which provides an extensive summary of the object. Functions for subsetting (*subset_methrix()*), filtering (*region_filter()*), combining (*combine_methrix()*), and summarizing (*get_region_summary()*) *methrix* objects support the further processing steps whereas, *write_bedgraphs()* and *write_bigwigs()* functions can export the *methrix* object to standardized file formats for further analysis with IGV or UCSC genome browsers.

Visualization: Several functions have been implemented to provide an overview of WGBS data, including beta-value/coverage distribution (*plot_density(), plot_coverage()*), QC plots (*plot_stats()*), and principal component analysis (PCA) (*methrix_pca()*).

In addition to above features, a *methrix* object can easily be converted to the frequently used *bsseq* object (*methrix2bsseq()*) thereby allowing users to interact with functions from *bsseq* dependent packages such as *DSS* (Feng *et al.*, 2014) or *dmrseq* (Korthauer *et al.*, 2019) for differential methylation calling. *Methrix* also supports processing large cohorts (>100) of samples with batch processing and parallelization in many of its functions (Supplementary Fig. S1). We benchmarked and compared the functionalities/performance of *methrix* with *bsseq*, *methylKit* and *RnBeads* (Supplementary Table S1 and Supplementary Fig. S2). As the results show, while *methrix* does additional operations, such as integrating genome information and calculating genome-wide statistics as part of the import function, it still performs comparable to other packages. Additionally, *methrix* outperformed *methylKit*, and *RnBeads* in filtering operations.

To further support researchers in analyzing their WGBS data in a comprehensive manner, we provide a best practice workflow for WGBS data analysis using *methrix* as Supplementary Data as well as part of a detailed documentation available at https://compepigen.github.io/methrix_docs/. This workflow includes data import, QC filtering steps, dimension reduction techniques such as PCA, identification and visualization of differentially methylated regions.

3 Conclusion

Methrix is an R package that offers an elegant solution to import and aggregate bedGraph-like coverage files from WGBS studies. It combines computational performance with rich downstream functionality and thereby overcomes several limitations of existing software packages while being flexible and user-friendly. The implementation allows efficient WGBS data analysis and seamless integration with other commonly used packages and tools.

Funding

The work was funded in part by the Helmholtz Foundation awarded to (C.P.), the German Network for Bioinformatics partner project de.NBI-Epi, subproject Heidelberg [to P.L.], the German Ministry of Education and Research [031L0162 to P.L. and R.T.] and German Cancer Aid grant CO-CLL [70113869 to P.L. and C.P.]. R.T. is partially supported by Deutsche Zentrum für Lungenforschung (DZL) awarded to (R.S.). A.M. is funded by DFG FOR 2674 [336840530]. The research leading to these results has received funding from the European Union Seventh Framework Program (FP7/2007–2013) under grant agreement no 311876: Pathway-27 (www.pathway27.eu).

Conflict of Interest: none declared.

References

Akalin,A. *et al.* (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.*, **13**, R87.

Assenov,Y. *et al.* (2014) Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods*, **11**, 1138–1140.

Feng,H. *et al.* (2014) A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res.*, **42**, e69.

Guo,W. *et al.* (2013) BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics*, **14**, 774.

Hansen,K.D. *et al.* (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.*, **13**, R83.

Korthauer,K. *et al.* (2019) Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. *Biostatistics*, **20**, 367–383.

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

Liu,Y. *et al.* (2012) Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol.*, **13**, R61.