MethylStar: A fast and robust pipeline for high-throughput analysis of bulk or single-cell WGBS data

Yadollah Shahryary[1], Rashmi R. Hazarika[2], Frank Johannes[1,2*]

*Correspondence: frank@johanneslab.org

Author details

[1] Technical University of Munich, Department of Plant Sciences, Liesel-Beckmann-Str. 2, 85354 Freising, Germany.

[2] Technical University of Munich, Institute for Advanced Study (IAS), Lichtenbergstr. 2a, 85748 Garching, Germany.

Abstract

Summary: Whole-Genome Bisulfite Sequencing (WGBS) is a Next Generation Sequencing (NGS) technique for measuring DNA methylation at base resolution. Recent drops in sequencing costs are beginning to enable high-throughput surveys of DNA methylation in large samples of individuals and/or single cells. These surveys can generate hundreds or even thousands of whole genome bisulfite sequencing (WGBS) datasets in a single study. The computational analysis of this data poses major challenges and creates unnecessary bottlenecks for biological interpretation. To offer an efficient analysis solution for such emerging data, we have developed MethylStar, a fast, stable and flexible computational pipeline. MethylStar offers easy installation through a dockerized container with all preloaded dependencies and also features a user-friendly interface designed for experts/non-experts. We show that MethylStar outperforms existing tools/pipelines for bulk and single-cell WGBS analysis.

Availability and implementation: MethylStar is distributed under GPL-3.0 license and source code is publicly available for download from github `https://github.com/jlab-code/MethylStar`. Installation through a docker image is available from `http://jlabdata.org/methylstar.tar.gz`

Introduction: As a result of recent drops in sequencing costs, an increasing number of laboratories and international consortia are adopting WGBS as the method of choice to survey DNA methylation in large population samples or in collections of cell lines and tissue types (IHEC, SYSCID, BLUEPRINT, EpiDiverse, NIH ROADMAP, Arabidopsis 1001 Epigenomes, Genomes and physical Maps), either in bulk or at the single-cell level ([Luo et al., 2017]; [Zhu et al., 2018]). Such surveys can easily generate hundreds or even thousands of WGBS datasets in a single study. A major computational challenge is the fast and reliable analysis of these large amounts of data. Although a number of WGBS pipelines exist, including gemBS ([Merkel et al., 2018]), nf-core/methylseq `https://github.com/nf-core/methylseq`, Bicycle ([Graña et al., 2017]), Methylpy ([Schultz et al., 2015]), they are usually used as standard processing tools and have not been optimized for high-throughput analysis. Moreover, these pipelines have been geared mainly towards human genome applications and may therefore show sub-optimal performance in the analysis of plant genomes, which can be substantially larger and more complex. To address these shortcomings, we have developed MethylStar, a fast and robust computational pipeline for high-throughput analysis of bulk or single-cell WGBS experiments.

Software features

MethylStar efficiently integrates all the steps of WGBS analysis. At its core, the pipeline uses established NGS tools including Trimmomatic ([Bolger et al., 2014]) for read processing, fastQC `https://www.bioinformatics.babraham.ac.uk/projects/fastqc` for quality control, Bismark ([Krueger and Andrews, 2011]) for alignment, and (optionally) METHimpute ([Taudt et al., 2018]) for methylation state calling.

Installation

MethylStar can be easily installed via a Docker image. This includes all the softwares, libraries, packages within the container, and thus solves any dependency issues. Advanced users can edit the existing docker container and build their own image.

Pipeline architecture and parallel support

The pipeline architecture comprises three main layers (Fig. 1A). The first layer is the user-interface implemented in Python. It is a simple command-based interface for configuring software settings, and is aimed at both experts and non-experts. The second layer consists of shell scripts, which handle low-level processes, efficiently coordinate the major software components and manage computational resources. The final layer is implemented in R, and is used to generate output files and other downstream analysis steps. MethylStar features a "Quick Run option", which allows the user to run all pipeline steps in one go. Alternatively, the "Advanced option" allows the user to manually run individual steps of the pipeline (Fig. 1A). All steps have been parallelized using GNU Parallel. The user can either set the number of parallel jobs manually, or can opt to use the inbuilt parallel option where the number of parallel is automatically detected based on available system resources.

Data processing and downstream functionalities

MethylStar integrates processing of raw fastq reads for both single- and paired-end data with options for adapter trimming (Trimmomatic), quality control (fastQC) and removal of PCR duplicates (Bismark software suite). Read alignment and cytosine context extraction

is performed with the Bismark software suite. Finally, cytosine-level methylation calls are obtained with METHimpute. All the different data processing steps have been optimized for speed and performance, and can run on local machines as well as on larger compute nodes. In addition to cytosine-level methylation calls, MethylStar offers functionalities for generating output files that are compatible with a number of publicly available DMR-callers such as Methylkit ([Akalin et al., 2012]), DMRcaller ([Catoni et al., 2018]). For visualization, the user can upload the final methylomes to a Genome Browser such as JBrowse ([Skinner et al., 2009]). All outputs are provided in standard data formats for downstream analysis.

Benchmarking

To demonstrate MethylStar's performance we analyzed bulk WGBS data from a selection of 200 *Arabidopsis thaliana* ecotypes (paired-end, 295GB, $\sim 8.63X$ depth, 85.66% genome coverage, GSE54292), 75 maize strains (paired-end, 209GB, $\sim 0.36X$ depth, $\sim 22.12\%$ genome coverage, GSE39232) and 88 Human H1 cell lines (single-end, 82GB, $\sim 0.12X$ depth, $\sim 10.62\%$ genome coverage, GSM429321). MethylStar was compared with three popular pipelines: Methylpy, nf-core/methylseq and gemBS. All pipelines were run with default parameters on a computing cluster with a total of 88 cores (CPU 2.2 GHz with 378 GB RAM). Speed performance was assessed for a series of batch sizes (*A. thaliana*: 50, 100, 150, 200 samples; human H1 cell line: 22, 44, 66, 88 samples; maize: 15, 30, 45, 60, 75 samples) and was restricted to a fixed number of jobs (=32), see Fig. 1B-C. Although gemBS achieved the fastest processing times for the *A. thaliana* samples, MethylStar clearly outperformed the other pipelines when applied to the more complex genomes of human and maize, which are computationally more expansive and resource-demanding (Fig. 1B). For instance, for 88 human WGBS samples (82GB of data), MethylStar showed a 75.61% reduction in processing time relative to gemBS, the second fastest pipeline (909 mins vs. 3727 mins). Extrapolating from these numbers, we expect that for 1000 human WGBS samples, MethylStar could save about ~ 22.24 days of run time (4x faster). To demonstrate that MethylStar can also be applied to single-cell WGBS data, we analyzed DNA methylation of 200 single cells from human early embryo tissue (paired-end, 845GB, ~ 0.38 depth, $\sim 9.97\%$ genome coverage, GSE81233) split into batches of 100 and 200, see Fig. 1C. MethylStar's processing times increased linearly with batch size

(i.e. number of cells). For 200 cells, MethylStar required only 4227 mins, thus making it an efficient analysis solution for deep single-cell WGBS experiments. Comparisons with the other pipelines were unfortunately not available in this setting, as their default implementation is incompatible with single-cell WGBS data.

Conclusion

MethylStar is a fast, stable and flexible pipeline for the high-throughput analysis of bulk or single-cell WGBS data. Its easy installation and user-friendly interface should make it a useful resource for the wider epigenomics community.

Funding

This work was supported by the SFB/Sonderforschungsbereich924 of the Deutsche Forschungsgemeinschaft (DFG) and the Technical University of Munich-Institute for Advanced Study funded by the German Excellent Initiative and the European Seventh Framework Programme under grant agreement no. 29176.

References

Akalin et al., 2012. Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., and Mason, C. E. (2012). methylkit: a comprehensive r package for the analysis of genome-wide dna methylation profiles. *Genome biology*, 13(10):R87.

Bolger et al., 2014. Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Catoni et al., 2018. Catoni, M., Tsang, J. M., Greco, A. P., and Zabet, N. R. (2018). Dmrcaller: a versatile r/bioconductor package for detection and visualization of differentially methylated regions in cpg and non-cpg contexts. *Nucleic acids research*, 46(19):e114–e114.

Graña et al., 2017. Graña, O., López-Fernández, H., Fdez-Riverola, F., González Pisano, D., and Glez-Peña, D. (2017). Bicycle: a bioinformatics pipeline to analyze bisulfite sequencing data. *Bioinformatics*, 34(8):1414–1415.

Krueger and Andrews, 2011. Krueger, F. and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics (Oxford, England)*, 27(11):1571–1572.

Luo et al., 2017. Luo, C., Keown, C. L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J. R., Sandoval, J. P., Bui, B., Sejnowski, T. J., Harkins, T. T., Mukamel, E. A., Behrens, M. M., and Ecker, J. R. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science (New York, N.Y.)*, 357(6351):600–604.

Merkel et al., 2018. Merkel, A., Fernández-Callejo, M., Casals, E., Marco-Sola, S., Schuyler, R., Gut, I. G., and Heath, S. C. (2018). gemBS: high throughput processing for DNA methylation data from bisulfite sequencing. *Bioinformatics*, 35(5):737–742.

Schultz et al., 2015. Schultz, M. D., He, Y., Whitaker, J. W., Hariharan, M., Mukamel, E. A., Leung, D., Rajagopal, N., Nery, J. R., Urich, M. A., Chen, H., Lin, S., Lin, Y., Jung, I., Schmitt, A. D., Selvaraj, S., Ren, B., Sejnowski, T. J., Wang, W., and Ecker, J. R. (2015). Human body epigenome maps reveal noncanonical dna methylation variation. *Nature*, 523(7559):212–216.

Skinner et al., 2009. Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J., and Holmes, I. H. (2009). Jbrowse: a next-generation genome browser. *Genome research*, 19(9):1630–1638.

Taudt et al., 2018. Taudt, A., Roquis, D., Vidalis, A., Wardenaar, R., Johannes, F., and Colomé-Tatché, M. (2018). Methimpute: imputation-guided construction of complete methylomes from wgbs data. *BMC Genomics*, 19(1):444.

Zhu et al., 2018. Zhu, P., Guo, H., Ren, Y., Hou, Y., Dong, J., Li, R., Lian, Y., Fan, X., Hu, B., Gao, Y., Wang, X., Wei, Y., Liu, P., Yan, J., Ren, X., Yuan, P., Yuan, Y., Yan, Z., Wen, L., Yan, L., Qiao, J., and Tang, F. (2018). Single-cell dna methylome sequencing of human preimplantation embryos. *Nature Genetics*, 50(1):12–19.

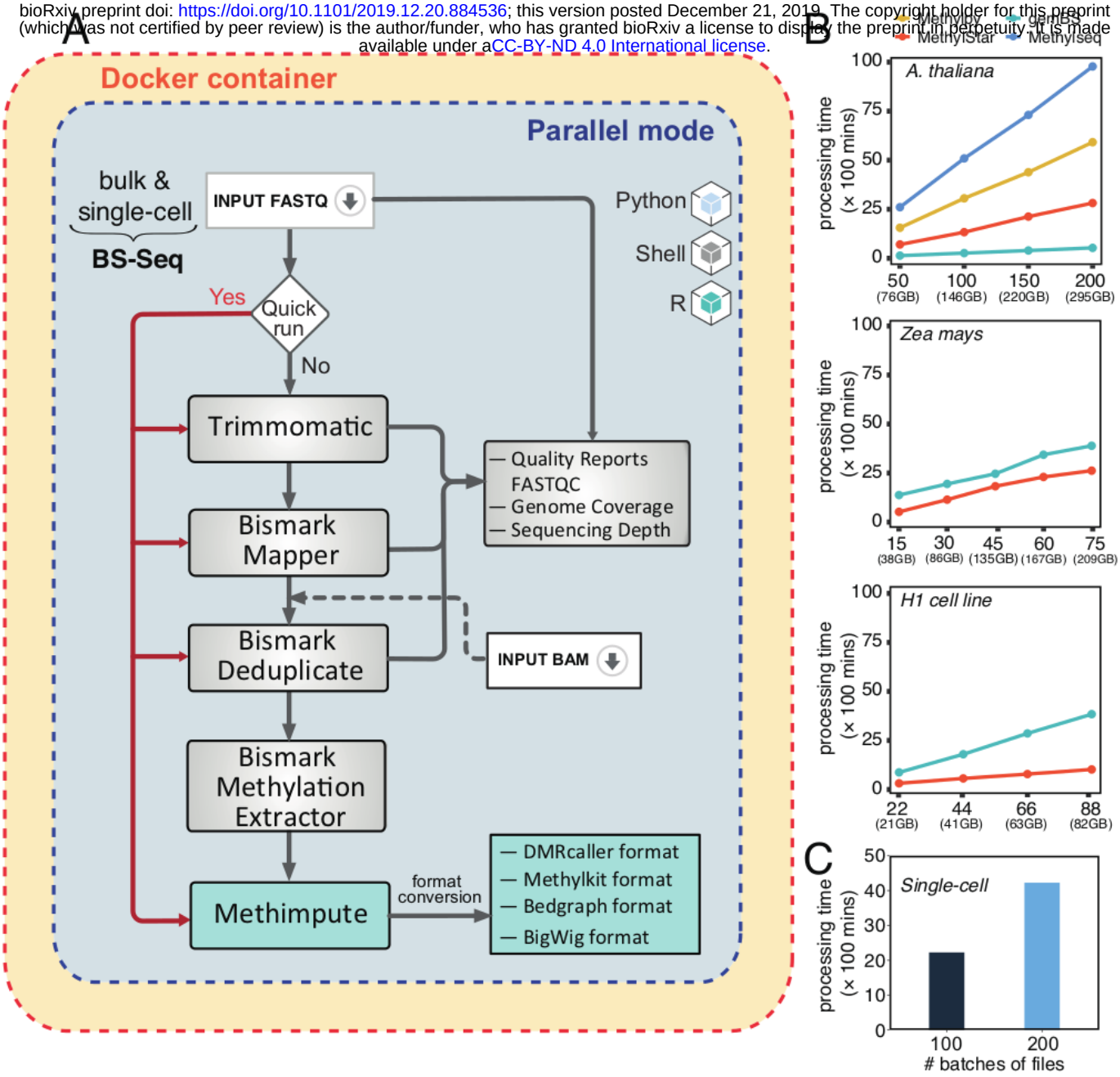

Fig. 1. (A) Basic workflow of MethylStar showing the pipeline architecture and different components. (B) Performance of MethylStar as compared with other BS-Seq analysis pipelines viz. Methylpy, nf-core/methylseq and gemBS. CPU processing time taken by METHimpute was not included in the current benchmarking process as there is no equivalent method in the other pipelines to compare with. Because of the very long run times observed for the *A. thaliana* data, methylpy and methylseq were no longer considered for benchmarking of the maize and H1 cell line samples. All pipelines were run using 32 jobs. (C) Time taken while processing batches of scBS-Seq samples.