Applications Note

PyMethylProcess - highly parallelized preprocessing for DNA methylation array data

Joshua J. Levy^{1,*}, Alexander J. Titus², Lucas A. Salas² and Brock C. Christensen^{2,3}

¹Program in Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth, Lebanon, NH, ²Department of Epidemiology, Geisel School of Medicine at Dartmouth, Lebanon, NH, ³Department of Epidemiology and Department of Molecular and Systems Biology, Lebanon, NH

*To whom correspondence should be addressed.

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: The ability to perform high-throughput preprocessing of methylation array data is essential in large scale methylation studies. While R is a convenient language for methylation analyses, performing highly parallelized preprocessing using Python can accelerate data preparation for downstream methylation analyses, including large scale production-ready machine learning pipelines. Here, we present a methylation data preprocessing pipeline called PyMethylProcess that is highly reproducible, scalable, and that can be quickly set-up and deployed through Docker and PIP.

Availability and Implementation: Project Name: PyMethylProcess

Project Home Page: <https://github.com/Christensen-Lab-Dartmouth/PyMethylProcess>. Available on PyPI as *pymethylprocess*.

Available on DockerHub via *joshualevy44/pymethylprocess*.

Help Documentation: <https://christensen-lab-dartmouth.github.io/PyMethylProcess/>

Operating Systems: Linux, MacOS, Windows (Docker)

Programming Language: Python, R

Other Requirements: Python 3.6, R 3.5.1, Docker (optional)

License: MIT

Contact: joshua.j.levy.gr@dartmouth.edu

1. Implementation

DNA methylation plays a critical role in cell fate determination during development and regulation of transcription throughout life (Ji *et al.*, 2010; Khavari *et al.*, 2010; Rönnblad *et al.*, 2014). Studies that measure DNA methylation in large numbers of human biospecimens typically use the Illumina Infinium BeadArray platforms known as 27K, 450K and 850K/EPIC arrays (Bibikova *et al.*, 2011, 2009; Moran *et al.*, 2016). Popular R packages used for DNA methylation data processing and analysis include *minfi* (Aryee *et al.*, 2014), *ENmix* (Xu *et al.*, 2016), or *meffil* (Min *et al.*, 2018). However, a straightforward and tractable approach to perform data quality control and functional normalization in bulk to prepare the data for use in the object-oriented environment is lacking. Here, we introduce a simple and easy-to-use command line interface that makes methylation analyses more object oriented and ready for use in downstream analyses such as immune cell proportion estimation, and python-oriented Epigenome-Wide Association Studies (EWAS) studies. In addition to more traditional methylation-based analyses, popular machine

learning libraries such as scikit-learn, keras, and tensorflow (Abadi *et al.*, 2016; Pedregosa *et al.*, 2011) become more accessible to both machine learning researchers without an epigenetic background and epigenetic researchers with a limited machine learning background.

PyMethylProcess is a pip-installable command line interface built using Python 3.6 that interfaces with *minfi*, *ENmix*, and *meffil* in R via rpy2 (Gautier, 2010) to allow users to preprocess and set-up their methylation array data for downstream analyses included machine learning, presenting unique methylation datatypes that are built for the use of python classification, clustering, dimensionality reduction and regression algorithms such as UMAP, random forest, neural networks, k-nearest neighbors, and HDBSCAN (Campello *et al.*, 2013; Ho, 1995; McInnes *et al.*, 2018). Eight Python classes have been introduced to handle the following tasks: package installation (*PackageInstaller* installs any R/bioconductor package via python), data acquisition from TCGA and GEO (*TCGADownloader*) and formatting (*PreProcessPhenoData*), parallelized quality control (QC), principal component selection via kneedle (Satopaa *et al.*, 2011), and raw, quantile, noob, and functional normalization using *minfi*, *meffil*, *ENmix*

PyMethylProcess

Conflicts of Interest: none declared.

List of Abbreviations

EWAS - Epigenome-Wide Association Studies

QC - Quality Control

CWL - Common Workflow Language

450K - HumanMethylation450 BeadArray

850K - HumanMethylationEPIC BeadArray

References

- Abadi, M. *et al.* (2016) TensorFlow: A System for Large-Scale Machine Learning., pp. 265–283.
- Amstutz, P. *et al.* (2016) Common Workflow Language, v1.0.
- Aryee, M.J. *et al.* (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinforma. Oxf. Engl.*, **30**, 1363–1369.
- Bibikova, M. *et al.* (2009) Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics*, **1**, 177–200.
- Bibikova, M. *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288–295.
- Boettiger, C. (2015) An Introduction to Docker for Reproducible Research. *SIGOPS Oper Syst Rev*, **49**, 71–79.
- Campello, R.J.G.B. *et al.* (2013) Density-Based Clustering Based on Hierarchical Density Estimates. In, Pei, J. *et al.* (eds), *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 160–172.
- Capper, D. *et al.* (2018) DNA methylation-based classification of central nervous system tumours. *Nature*, **555**, 469–474.
- Gautier, L. (2010) An intuitive Python interface for Bioconductor libraries demonstrates the utility of language translators. *BMC Bioinformatics*, **11**, S11.
- Gorrie-Stone, T.J. *et al.* Bigmelon: tools for analysing large DNA methylation datasets. *Bioinformatics*.
- Ho, T.K. (1995) Random Decision Forests. In, *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95. IEEE Computer Society, Washington, DC, USA, pp. 278–.
- Houseman, E.A. *et al.* (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, **13**, 86.
- Jaffe, A.E. and Irizarry, R.A. (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.*, **15**, R31.
- Ji, H. *et al.* (2010) A comprehensive methylome map of lineage commitment from hematopoietic progenitors. *Nature*, **467**, 338–342.
- Johansson, Å. *et al.* (2013) Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. *PLOS ONE*, **8**, e67378.
- Khavari, D.A. *et al.* (2010) DNA methylation and epigenetic control of cellular differentiation. *Cell Cycle Georget. Tex.*, **9**, 3880–3883.
- Li Yim, A.Y.F. *et al.* (2016) Peripheral blood methylation profiling of female Crohn's disease patients. *Clin. Epigenetics*, **8**, 65.
- Mahpour, A. (2016) pyMAP: a Python package for small and large scale analysis of Illumina 450k methylation platform. *bioRxiv*, 078048.
- McInnes, L. *et al.* (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat.*
- Min, J.L. *et al.* (2018) Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinforma. Oxf. Engl.*, **34**, 3983–3989.
- Modern Analytic Apps for the Enterprise *Plotly*.
- Moran, S. *et al.* (2016) Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, **8**, 389–399.
- Pai, S. *et al.* (2018) Differential DNA modification of an enhancer at the IGF2 locus affects dopamine synthesis in patients with major psychosis. *bioRxiv*, 296756.
- Pedregosa, F. *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
- Pidsley, R. *et al.* (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, **14**, 293.
- Rahmani, E. *et al.* (2017) GLINT: a user-friendly toolset for the analysis of high-throughput DNA-methylation array data. *Bioinformatics*, **33**, 1870–1872.
- Rönnerblad, M. *et al.* (2014) Analysis of the DNA methylome and transcriptome in granulopoiesis reveals timed changes and dynamic enhancer methylation. *Blood*, **123**, e79–89.
- Salas, L.A. *et al.* (2018) An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biol.*, **19**.

Salas, L.A. *et al.* (2017) Integrative epigenetic and genetic pan-cancer somatic alteration portraits. *Epigenetics*, **12**, 561–574.

Satopaa, V. *et al.* (2011) Finding a ‘Kneedle’ in a Haystack: Detecting Knee Points in System Behavior. In, *2011 31st International Conference on Distributed Computing Systems Workshops.*, pp. 166–171.

Soriano-Tárraga, C. *et al.* (2018) Biological Age is a predictor of mortality in Ischemic Stroke. *Sci. Rep.*, **8**, 4148.

Xu, Z. *et al.* (2016) ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Res.*, **44**, e20.