Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Applications Note

Applications Note

PyMethylProcess - highly parallelized preprocessing for DNA methylation array data

Joshua J. Levy[1,*], Alexander J. Titus[2], Lucas A. Salas[2] and Brock C. Christensen[2,3]

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

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

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.

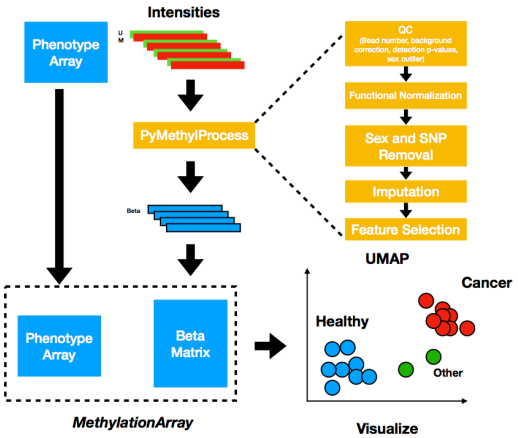
Availability and Implementation: Project Name: PyMethylProcess
Project Home Page: https://github.com/Christensen-Lab-Dartmouth/PyMethylProcess. Available on PyPI as *pymethylprocess.* Available on DockerHub via *joshualevy44/pymethylprocess*.
Help Documentation: https://christensen-lab-dartmouth.github.io/PyMethylProcess/
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 [1-3]. 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 [4-6]. Popular R packages used for DNA methylation data processing and analysis include *minfi* [7], *ENmix* [8], or *meffil* [9]. 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 become more accessible to both machine learning researchers without an epigenetic background and epigenetic researchers with a limited machine learning background [10-11].

PyMethylProcess is a pip-installable command line interface built using Python 3.6 that interfaces with *minfi, ENmix,* and *meffil* in R via rpy2 [12] 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 [13-15]. 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 [16], and raw, quantile, noob, and functional normalization using *minfi, meffil, ENmix* (*PreprocessIDAT*), imputation (*ImputerObject)*, feature selection, final preparation and storage for machine learning applications (*MethylationArray[s]* store beta and phenotype data), and a basic machine learning API class *MachineLearning* that trains any scikit-learn-like model on *MethylationArray* objects. These datatypes are used to handle complex preprocessing calculations, while being abstracted away via a convenient command-line interface. Additional commands, such as the removal of non-autosomal sites, SNP removal (either via QC methods or post-QC by subsetting CpGs that are not in a list of CpGs supplied by *meffil* for the respective array platform), and reference-based cell-type estimation (constrained projection/quadratic programming) [17-19], and class methods are available in the help documentation. In addition, a visualization module generates interactive 3-D representations of the data using UMAP and Plotly [20] for further inspection. The beta values and phenotype data from the *MethylationArray* objects can easily be written to csv format using *write_csv*, and the command line interface can speed up the time it takes for researchers to set-up their standardized methylation data for downstream analyses.

1. Figure 1. Flow Diagram for *PyMethylProcess*



The pipeline differs from other python frameworks such as pyMAP [21] and GLINT [22] in functionality. pyMAP only operates on the 450k framework, relying on user specific csv annotation files and preprocessed Genome Studio txt files as its input. pyMAP only performs graphical exploration for candidate CpGs, CpG Island feature subsetting, and export to BED files for downstream analyses. Similarly, GLINT requires a txt phenotype file and either a preprocessed beta values txt files or a R data.frame methylation object (in a RData file) as its inputs. GLINT stores beta values and covariate information as a binary "glint" file. GLINT was designed for EWAS analysis, including reference-based and reference-free estimations, imputed genetic structure, and statistical models (linear, logistic and linear mixed effects models). However, it relies on preprocessed data, with some limited quality control options and as so it could benefit from preprocessed data generated in our pipeline. In addition, GLINT is not designed to export this information to perform user customized downstream machine learning analyses.

2. Results

Here we demonstrate some of *PyMethylProcess*'s capabilities, performing preprocessing of seven different datasets (GSE87571, GSE81961, GSE69138, GSE42861, GSE112179, GSE109381, TCGA Pancancer [23-30]) from the HumanMethylation450 BeadArray (450K array), and the HumanMethylationEPIC BeadArray (850K array), platform, derived from GEO, most of which benchmarked for loading, QC and normalization time. After preprocessing, each of the below data sets were split into 70% training, 10% validation, and 20% test sets for downstream prediction.

1. Table 1. Benchmark and Preprocessing Results on Runs of 450K and 850K Studies

PyMethylProcess Pipeline Benchmarks

DataSets	Brief Description	Sample Size	Preprocessing Pipeline	# CpGs After Normalization	Principal Components	# Outlier Samples	# CPUs	Memory (Gb)	Runtime (Minutes)	# Sites Removed	Percentage Imputed (%)	Imputation Method
GSE87571	Johannson Aging	732	Minfi: Noob Normalization	482669	NA	13	1	NA	150.0	11233	0.706	K-NN: 15 neighbors
GSE81961	Crohn's Disease	40	Meffil: Functional Normalization	480329	4.0	0	35	10	4.1	11587	0.096	K-NN: 5 neighbors
GSE69138	Stroke	185	Meffil: Functional Normalization	474021	13.0	2	35	NA	11.5	11305	0.166	K-NN: 5 neighbors
GSE42861	Smoking and Arthritis	689	Meffil: Functional Normalization	477482	13.0	8	30[1]	110[2]	38.5[3]	11448	0.190	K-NN: 5 neighbors
GSE112179	Schizophrenia and Bipolar	100	Meffil: Functional Normalization	853772	10.0	2	30	NA	40.0	19278	0.240	K-NN: 10 neighbors
GSE109381	Brain Cancer Subclasses	3897	Meffil: Functional Normalization[4]	320023	10.4[5]	135	30	60	135.0	6791	0.054	Mean
TCGA Pancancer	33 Pan-Cancer subtypes	8891	Meffil: Functional Normalization[4]	378588	14.6[6]	515	30	60	255.0	7869	0.112	Mean

[1] Quality control used 30 CPUs, Normalization Used 14 CPUs
[2] Only for normalization step
[3] 13.5 Minutes for QC, 25 minutes for Normalization
[4] Subclasses Processed in Parallel
[6] Averaged Across Disease Subtypes

3. Benefits and Future Direction

The *PyMethylProcess* streamlines the process of preprocessing Methylation Array data while making Methylation data highly accessible and standardized for the Python machine learning community. *PyMethylProcess* is open source software, and additional development based on the community needs may be added. The authors of this paper would like to solicit issues and pull requests from the greater bioinformatics community on GitHub.

Future development will expand functionality to other preprocessing pipelines such as Watermelon and BigMelon [31] and feature importance evaluations using measures such as the Gini index will be included. In addition, this command-line tool is available for use via Docker (*joshualevy44/pymethylprocess*) [32], making this analysis easily sharable, standardized, and operating system agnostic. The entire workflow will be wrapped using Common Workflow Language (CWL) [33], making the entire analysis executable at the click of a button, highly reproducible, and easily sharable.

The *PyMethylProcess* is installable from PyPI via the name *pymethylprocess*, and source code ***can be found on GitHub at:***

https://github.com/Christensen-Lab-Dartmouth/PyMethylProcess

Acknowledgements

JJL conceived of idea and implementation, programmed and tested pipeline on datasets, wrote text; AJT tested pipeline; LAS provided technical support to streamline and debug important aspects of the pipeline; BCC provided research direction support; all authors participated in editing the text.

Funding

This work was supported by NIH grants R01CA216265, R01DE022772, and P20GM104416 to BCC, a Dartmouth College Neukom Institute for Computational Science CompX award to BCC, and training fellowship support for AJT from T32LM012204.

Conflict 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

Khavari D.A. , Sen G.L., Rinn J.L. (2010) DNA methylation and epigenetic control of cellular differentiation. *Cell Cycle* , 9, 3880–3883.

Ji H. , Ehrlich L.I.R., Seita J., Murakami P., Doi A., Lindau P., Lee H., Aryee M.J., Irizarry R.A., Kim K. et al. (2010) Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* , 467, 338–342.

Rönnerblad M. , Andersson R., Olofsson T., Douagi I., Karimi M., Lehmann S.S., Hoof I., de Hoon M., ItohM., Nagao-Sato S. et al. (2014) Analysis of the DNA methylome and transcriptome in granulopoiesis reveals timed changes and dynamic enhancer methylation. *Blood* , 123, e79–e89.

Bibikova M., Le J, Barnes B., Saedinia-Melnyk S., Zhou L., Shen R., Gunderson K.L. (2009) Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics* , 1, 177–200.

Bibikova M. , Barnes B., Tsan C., Ho V., Klotzle B., Le J.M., Delano D., Zhang L., Schroth G.P., GundersonK.L. et al. (2011) High density DNA methylation array with single CpG site resolution. *Genomics* , 98, 288–295.

Moran S. , Arribas C., Esteller M. (2016) Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* , 8, 389–399.

Aryee M.J. , Jaffe A.E., Corrada-Bravo H., Ladd-Acosta C., Feinberg A.P., Hansen K.D., Irizarry R.A.(2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* , 30, 1363–1369.

Xu, Z., Niu, L., Li, L., & Taylor, J. A. (2016). ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Research*, *44*(3), e20. https://doi.org/10.1093/nar/gkv907

Min, J. L., Hemani, G., Davey Smith, G., Relton, C., & Suderman, M. (2018). Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics (Oxford, England)*, *34*(23), 3983–3989. https://doi.org/10.1093/bioinformatics/bty476

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … Zheng, X. (2016). TensorFlow: A System for Large-Scale Machine Learning (pp. 265–283). Presented at the 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16). Retrieved from https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

rpy2. (n.d.). Retrieved March 6, 2019, from http://rpy.sourceforge.net/

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv:1802.03426 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1802.03426

Ho, Tin Kam (1995). *Random Decision Forests* (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. Retrieved 5 March 2019.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 160–172). Springer Berlin Heidelberg.

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (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). https://doi.org/10.1109/ICDCSW.2011.20

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., … Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, *13*, 86. https://doi.org/10.1186/1471-2105-13-86

Jaffe, A. E., & Irizarry, R. A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, *15*(2), R31. https://doi.org/10.1186/gb-2014-15-2-r31

Salas, L. A., Koestler, D. C., Butler, R. A., Hansen, H. M., Wiencke, J. K., Kelsey, K. T., & Christensen, B. C. (2018). An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biology*, *19*. https://doi.org/10.1186/s13059-018-1448-7

Perkel, J. M. (2018). Data visualization tools drive interactivity and reproducibility in online publishing. *Nature*, *554*, 133. https://doi.org/10.1038/d41586-018-01322-9

Mahpour, A. (2016). pyMAP: a Python package for small and large scale analysis of Illumina 450k methylation platform. *BioRxiv*, 078048. https://doi.org/10.1101/078048

Rahmani, Elior, Reut Yedidim, Liat Shenhav, Regev Schweiger, Omer Weissbrod, Noah Zaitlen, and Eran Halperin. "GLINT: a user-friendly toolset for the analysis of high-throughput DNA-methylation array data." Bioinformatics 2017; 33 (12): 1870-1872.

Johansson, Å., Enroth, S., & Gyllensten, U. (2013). Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. *PLOS ONE*, *8*(6), e67378. https://doi.org/10.1371/journal.pone.0067378

Peripheral blood methylation profiling of female Crohn's disease patients | Clinical Epigenetics | Full Text. (n.d.). Retrieved March 6, 2019, from https://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-016-0230-5

Soriano-Tárraga, C., Giralt-Steinhauer, E., Mola-Caminal, M., Ois, A., Rodríguez-Campello, A., Cuadrado-Godia, E., … Jiménez-Conde, J. (2018). Biological Age is a predictor of mortality in Ischemic Stroke. *Scientific Reports*, *8*(1), 4148. https://doi.org/10.1038/s41598-018-22579-0

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., … Feinberg, A. P. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, *31*(2), 142–147. https://doi.org/10.1038/nbt.2487

Pai, S., Li, P., Killinger, B., Marshall, L., Jia, P., Liao, J., … Labrie, V. (2018). Differential DNA modification of an enhancer at the IGF2 locus affects dopamine synthesis in patients with major psychosis. *BioRxiv*, 296756. https://doi.org/10.1101/296756

Capper, D., Jones, D. T. W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., … Pfister, S. M. (2018). DNA methylation-based classification of central nervous system tumours. *Nature*, *555*(7697), 469–474. https://doi.org/10.1038/nature26000

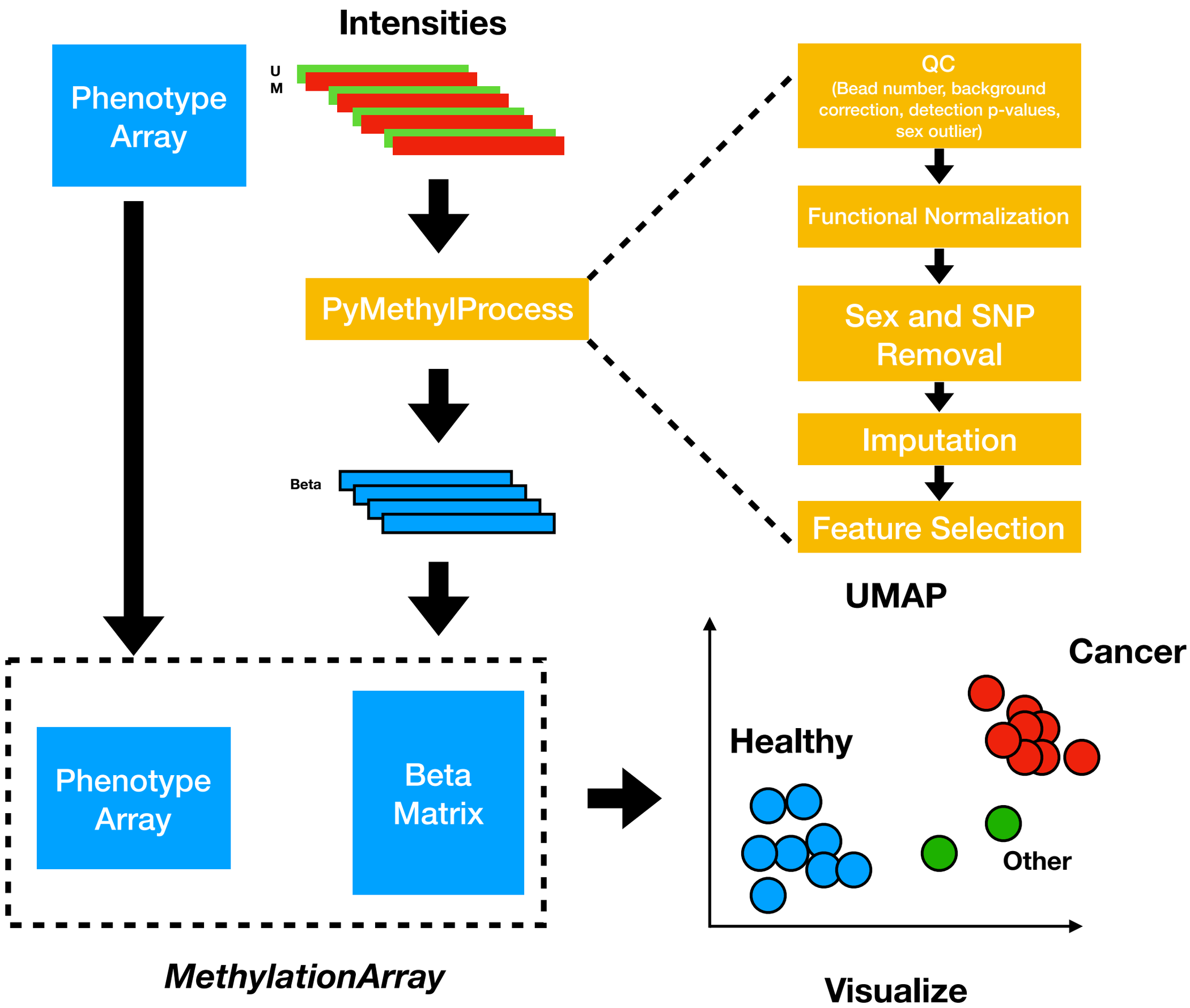
Salas, L. A., Johnson, K. C., Koestler, D. C., O'Sullivan, D. E., & Christensen, B. C. (2017). Integrative epigenetic and genetic pan-cancer somatic alteration portraits. *Epigenetics*, *12*(7), 561–574. https://doi.org/10.1080/15592294.2017.1319043

Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., & Schalkwyk, L. C. (2013). A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, *14*, 293. https://doi.org/10.1186/1471-2164-14-293

Gorrie-Stone, T. J., Smart, M. C., Saffari, A., Malki, K., Hannon, E., Burrage, J., … Schalkwyk, L. C. (n.d.). Bigmelon: tools for analysing large DNA methylation datasets. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty713

Boettiger, C. (2015). An Introduction to Docker for Reproducible Research. *SIGOPS Oper. Syst. Rev.*, *49*(1), 71–79. https://doi.org/10.1145/2723872.2723882

Amstutz, P., Crusoe, M. R., Tijanić, N., Chapman, B., Chilton, J., Heuer, M., … Stojanovic, L. (2016). Common Workflow Language, v1.0. https://doi.org/10.6084/m9.figshare.3115156.v2


Intensities
Phenotype Array
U
M
PyMethylProcess
Beta
QC
(Bead number, background correction, detection p-values, sex outlier)
Functional Normalization
Sex and SNP Removal
Imputation
Feature Selection
UMAP
Phenotype Array
Beta Matrix
MethylationArray
Cancer
Healthy
Other
Visualize

PyMethylProcess Pipeline Benchmarks

DataSets	Brief Description	Sample Size	Preprocessing Pipeline	# CpGs After Normalization	Principal Components	# Outlier Samples	# CPUs	Memory (Gb)	Runtime (Minutes)	# Sites Removed	Percentage Imputed (%)	Imputation Method
GSE87571	Johannson Aging	732	Minfi: Noob Normalization	482669	NA	13	1	NA	150.0	11233	0.706	K-NN: 15 neighbors
GSE81961	Crohn's Disease	40	Meffil: Functional Normalization	480329	4.0	0	35	10	4.1	11587	0.096	K-NN: 5 neighbors
GSE69138	Stroke	185	Meffil: Functional Normalization	474021	13.0	2	35	NA	11.5	11305	0.166	K-NN: 5 neighbors
GSE42861	Smoking and Arthritis	689	Meffil: Functional Normalization	477482	13.0	8	30[1]	110[2]	38.5[3]	11448	0.190	K-NN: 5 neighbors
GSE112179	Schizophrenia and Bipolar	100	Meffil: Functional Normalization	853772	10.0	2	30	NA	40.0	19278	0.240	K-NN: 10 neighbors
GSE109381	Brain Cancer Subclasses	3897	Meffil: Functional Normalization[4]	320023	10.4[5]	135	30	60	135.0	6791	0.054	Mean
TCGA Pancancer	33 Pan-Cancer subtypes	8891	Meffil: Functional Normalization[4]	378588	14.6[5]	515	30	60	255.0	7869	0.112	Mean

[1] Quality control used 30 CPUs, Normalization Used 14 CPUs
[2] Only for normalization step
[3] 13.5 Minutes for QC, 25 minutes for Normalization
[4] Subclasses Processed in Parallel
[5] Averaged Across Disease Subtypes