Intro

**What is epigenetics?**

Epigenetics is the study of heritable and reversible changes of the genome that do not change the underlying sequence. These changes are fundamental for controlling gene expression in all organisms, and can occur through intrinsic methods (e.g., X-chromosome inactivation in females) or external causes (e.g., exposure to ultra-violet radiation). The molecular marks of epigenetics span several levels from overall chromatin structure, through histone modifications, down to changes of individual bases, with methylation.

**Describe epigenetic marks**

Methylation is the most studied epigenetic modification. Adenine and cytosine both readily accept methylation on N4and C4/C5, respectively. Methylated adenine is weakly studied but widely known in bacteria and lower eukaryotes, but is not considered to have any effect or significant presence in mammalian DNA1,2. Comparatively, cytosine modifications are abundant and well-studied among all domains of life. 5-methyl-cytosine (5mC) is the most common epigenetic mark in humans, with about 28 million sites present in the genome3. These sites, known as CpG sites for the cytosine-phosphate-guanine bridge, are not randomly distributed but instead highly localized around regulatory motifs: 70% of promoters show high density of CpG sites4. This density is called a CpG island (CGI). The bulky methyl groups at CpG sites provide a physical block to binding of transcription factors or even RNA polymerase itself5. As such, methylation is a major factor in gene expression in mammals, so mechanisms for adding and removing methylation must be strictly regulated.

These epigenetic marks can be *de novo* added by DNA Methyltransferase proteins (DNMT3a/b), or removed by ten-eleven translocation methylcytosine dioxygenases (TET family), so these modifications are considered reversible. These are necessary for normal development of tissues, as Example of adding and removing. Furthermore, these modifications can also be maintained through cell division by DNMT1, hence they are considered heritable. Example of heritable

These mechanisms must be strictly controlled, as dysregulation can significantly affect health and disease. Example of sickness by epigenetics

Implications to cancer

Malignant cell growth is also strongly linked to epigenetic abnormalities. Abnormal overexpression of DNMT proteins has shown significant effects on multiple cancers (e.g. DNMT3A in 25% of acute myeloid leukemia cases and DNMT1 in 12% of uterine cancer cases <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7465608/#:~:text=found%20that%20the%20coding%20exons,characteristic%20change%20of%20tumor%20cells>.). Conversely, underexpression of TET proteins has been shown in some cancer types (e.g. up to 58% of chronic myelomonocytic leukemia cases). Thus, epimutations can both increase and decrease gene expression, which is problematic with pro-tumor and tumor-suppressing genes, respectively. Example related to later work. Alterations of the epigenomic landscape can cause widespread tumor and disease, and understanding this process is critical for well being.

Whole genome bisulfite sequencing (WGBS) is the *de facto* technique for studying the epigenomic landscape. In this technique, sodium bisulfite converts unmethylated cytosines to uracil (and later thymine) which allows differentiation by comparing pre- and post-treatment sequencing, then aligning to the reference methylome. The resulting single nucleotide polymorphisms (SNPs) allow generation of the epigenomic landscape. A coverage of at least 30x is recommended with this technique, as during treatment, single strand nicks are randomly introduced, so up to 95% of the CpG sites lost during sequencing cite.

This causes problems with low population samples. Recently, new techniques focusing on single cell data have emerged to address these problems. Single cell bisulfite sequencing (sc-BS) and XXX

Existing methods

Importance of single cell methods

However, WGBS suffers some drawbacks due to its bulk processing. Low population samples cannot give suitable coverage and is difficult to differentiate between heterogenous cells. Single cell bisulfite sequencing (sc-BS) fills this gap.

Furthermore, the data obtained from sc-BS presents its own problems. While low coverage can be accommodated by many tools, sparsity of the data is a challenge. There are inherent losses with bisulfite treatment as described above. Existing WGBS tools are unable of handle a high proportion of NA values, and there is a high memory requirement if stored in a dense matrix, as most also require. Hence, we have developed a new tool to address these problems and allow a pipeline similar to traditional WGBS tools.

To address these problems, we developed scMethrix, a methylation-specific single cell data storage and manipulation tool. This R package is well-suited for the specific needs of single cell data and for integration into the Bioconductor ecosystem.

https://www.nature.com/articles/nri.2017.125

Ly6Chi classical inflammatory monocytes (CD14hiCD16− in humans) and Ly6Clow non-classical patrolling monocytes (CD14+CD16hi in humans)[35](https://www.nature.com/articles/nri.2017.125#ref-CR35). Ly6Clow monocytes are derived from Ly6Chi monocytes in either the blood or the bone marrow (BM)[36](https://www.nature.com/articles/nri.2017.125#ref-CR36). During certain diseases or injuries involving breach of the [blood–brain barrier](https://www.nature.com/articles/nri.2017.125#Glos1) (BBB), Ly6Chi monocytes may infiltrate the brain parenchyma and differentiate into microglia-like cells, which are intermingled with the resident microglia, to exacerbate or alleviate disease progression[29](https://www.nature.com/articles/nri.2017.125#ref-CR29),[35](https://www.nature.com/articles/nri.2017.125#ref-CR35).

Glioma has shown strong correlation with macrophages such as native microglia and infiltrating monocytes6. Up to 50% of the tumor bulk in glioblastoma is made up of these tumor-associated macrophages (TAMs)7. In response to brain lesions caused by glioma or other diseases, circulating Ly‑6ChiCCR2+ monocytes are preferentially recruited and differentiate into microglia8,9. These non-native microglia show functional differences to their native counterparts, and thus are an important factor for disease progression. For instance, in Alzheimer's disease, only bone marrow-derived monocytes are able to phagocytose the amyloid plaques that contribute to disease onset10. The mechanisms for this are not clear, but IL-1β has been implicated a key signal for plaque phagocytosis through transgenic overexpression11. This activation effect could have relevance for cancer progression, as glioblastoma has shown to aberrantly express IL-1β12.

Previous studies have tried to deplete macrophage

These tumor-associated macrophages (TAMs) have previously shown promise as a target for treatment13.

However, native and monocyte-derived microglia are virtually indistinguishable by known immunophenotypic markers.

Methylation has been used to determine lineage of stem cells through differentiations

Imethylation profiles in the macrophage landscape

Methylation state can contribute to disease severity. In systemic lupus erythematosus, pediatric-onset cases show a severe clinical course compared to adult-onset cases and is attributed to differences in methylation between immune cell types14. Only 21 CpG sites overlapping 15 genes where necessary to generate a DNAme signature for pediatric- versus adult-onset lupus.

After ischemic stroke, native microglia show compromised cell progression and largely adopt a pro-inflammatory phenotype compare to infiltrating monocytes15.

Phagocytosis is the defining characteristic of macrophages,

This occurs not only with amyloid plaques and Alzheimer’s, but with myelin debris in multiple sclerosis

and concluded that DNAme changes in different immune cell-lineages

The aim of this project is two-fold. First, a new tool is to be developed for handling single cell methylome data, as the challenges of data sparsity and incompatibility with established downstream methylation analysis techniques remain poorly fulfilled by existing software tools. The software package will be modelled after its bulk-WGBS predecessor, methrix, but will give additionally functionality for imputation, clustering, binning, and more. Our intention is to integrate it within the Bioconductor ecosystem to maximize compatibility and audience. Secondly, to help validate the tool, we will look at the relationship between glioma and macrophage methylation. Using epigenetic lineage, we hypothesize that methylation can be used to identify and help determine the origins of TAMs implicated in gliomagenesis, including native microglia and infiltrating monocytes. With this knowledge, it may be possible to generate prognostic outcomes based on diversity of TAMs or their characteristic methylation profiles. This may improve patient-specific treatment and aid in identifying further avenues of research for fighting this devastating disease.

Methods

Raw data files were obtained from publicly available sources, including monocyte, neutrophil, NK cells, eosinophils from whole blood samples of 44 individuals (GSE35069, GSE88824, GSE166844), glia (GSE66351), microglia-like cells from 25 mouse samples (GSE121483), glioma-related immune cells from 77 individuals (GSE151506).

Methylation array data (e.g., Illumina 450k and EPIC) were converted to genomic positions via the Minfi R package16. All genome assemblies were translated to hg38 by the Rtracklayer::liftOver R package17. For data sourced from mouse models (GSE121483), only the 19 420 probes conserved with humans were kept18, and it has been show that CpG methylation for glial cells is highly conserved between mice and humans19.

Scripts for data importing are available on Github (<https://github.com/knacko/monobrainDNAme>).

The package contains a comprehensive vignette that outlines the specifics of each function as well as a sample workflow using open source data.

This is a summary of the package functionality:

**Data storage:** Experimental data can be stored either in-memory or as an HDF5 object on a hard disk, and can be converted between each other. Most external functions cannot interface with HDF5 data, so it will be cast as a R matrix before processing. Metadata for the experiment (e.g., additional sample IDs, machines used, relevant dates, CpG annotation) can be stored as either *rowData*() for CpGs, *colData*() for samples, or *metadata*() for overall experiment information. Named assays can be stored in the object, as well as dimensionality reduction data.

**Filtering:** Both quality control and region- or sample-based filtering is possible. Masking can be done by coverage, sample count, or CpG methylation variance. Genomic regions can be subset by sample, contig, or chromosome.

**Imputation:** Due to the sparsity of single cell data, imputation may be used to fill the gaps. This package contains three methods of imputation: k-nearest-neighbour (kNN; via *impute*20), iterative principal component analysis (iPCA; via *missMDA*21), and random forest (RF; via *missForest*22). Other imputation functions can be used as well via an arbitrary input function. Imputation can be done either by whole genome or by regions, but this should be done with the assumption of quasi-independence of those regions.

**Binning:** There are multiple options to customize binning windows. Binning can be done as whole genome or by regions, and within these, windows can be specified by number of base pairs or number of CpG sites. By default, the binning will calculate the mean of the region, except for the coverage matrix, in which the sum count will be taken. The transformation used for binning is user customizable.

**Clustering:** Before clustering, multiple distance metrics are available to create the distance matrix: euclidean, manhattan, canberra, binary, and minkowski (via base R), as well as pearson, spearman, and kendall (via *biodist*23). Using this metric, clusters can be generated via hierarchy or partitions (via base R), as well as model-based clustering (via *mclust*24). Other clustering algorithms and distance metrics can be used via arbitrary function.

**Dimensionality reduction:** For plotting and clustering visualization, dimensionality must be reduced. Included are functions for PCA (via base R), uniform manifold approximation and projection (UMAP; via *umap*25) and t-distributed stochastic neighbor embedding (tSNE; via *tsne*26). The number of CpGs to use can be either by highest variance or randomly chosen. This data is stored in the object for later plotting.

Visualization:

Exporting: Multiple output formats are possible with this package.

1. Bochtler, M. & Fernandes, H. DNA adenine methylation in eukaryotes: Enzymatic mark or a form of DNA damage? *BioEssays* **43**, 2000243 (2021).

2. Douvlataniotis, K., Bensberg, M., Lentini, A., Gylemo, B. & Nestor, C. E. No evidence for DNA N6-methyladenine in mammals. *Science Advances* **6**, eaay3335.

3. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* **8**, 1499–1504 (1980).

4. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103**, 1412–1417 (2006).

5. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25**, 1010–1022 (2011).

6. Zhang, N. *et al.* The Predictive Value of Monocytes in Immune Microenvironment and Prognosis of Glioma Patients Based on Machine Learning. *Front Immunol* **12**, 656541 (2021).

7. Morantz, R. A., Kepes, J. J., Batnitzky, S. & Masterson, B. J. Extraspinal ependymomas. Report of three cases. *J Neurosurg* **51**, 383–391 (1979).

8. Mildner, A. *et al.* Microglia in the adult brain arise from Ly-6ChiCCR2+ monocytes only under defined host conditions. *Nat Neurosci* **10**, 1544–1553 (2007).

9. Djukic, M. *et al.* Circulating monocytes engraft in the brain, differentiate into microglia and contribute to the pathology following meningitis in mice. *Brain* **129**, 2394–2403 (2006).

10. Simard, A. R., Soulet, D., Gowing, G., Julien, J.-P. & Rivest, S. Bone Marrow-Derived Microglia Play a Critical Role in Restricting Senile Plaque Formation in Alzheimer’s Disease. *Neuron* **49**, 489–502 (2006).

11. Shaftel, S. S. *et al.* Sustained hippocampal IL-1 beta overexpression mediates chronic neuroinflammation and ameliorates Alzheimer plaque pathology. *J Clin Invest* **117**, 1595–1604 (2007).

12. Tarassishin, L., Casper, D. & Lee, S. C. Aberrant Expression of Interleukin-1β and Inflammasome Activation in Human Malignant Gliomas. *PLoS One* **9**, e103432 (2014).

13. Grégoire, H. *et al.* Targeting Tumor Associated Macrophages to Overcome Conventional Treatment Resistance in Glioblastoma. *Front Pharmacol* **11**, 368 (2020).

14. Yeung, K. S. *et al.* Cell lineage-specific genome-wide DNA methylation analysis of patients with paediatric-onset systemic lupus erythematosus. *Epigenetics* **14**, 341–351 (2019).

15. Ritzel, R. M. *et al.* Functional differences between microglia and monocytes after ischemic stroke. *Journal of Neuroinflammation* **12**, 106 (2015).

16. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).

17. Bioconductor Package Maintainer. liftOver: Changing genomic coordinate systems with rtracklayer::liftOver. R package version 1.18.0. https://www.bioconductor.org/help/workflows/liftOver/.

18. Needhamsen, M. *et al.* Usability of human Infinium MethylationEPIC BeadChip for mouse DNA methylation studies. *BMC Bioinformatics* **18**, 486 (2017).

19. Kessler, N. J. *et al.* CpG methylation differences between neurons and glia are highly conserved from mouse to human. *Human Molecular Genetics* **25**, 223–232 (2016).

20. Hastie, T., Tibshirani, R., Narasimhan, B. & Chu, G. *impute: impute: Imputation for microarray data*. (Bioconductor version: Release (3.14), 2021). doi:10.18129/B9.bioc.impute.

21. Husson, F. & Josse, J. *missMDA: Handling Missing Values with Multivariate Data Analysis*. (2020).

22. Stekhoven, D. J. *missForest: Nonparametric Missing Value Imputation using Random Forest*. (2013).

23. Ding, B., Gentleman, R. & Carey, V. *bioDist: Different distance measures*. (Bioconductor version: Release (3.14), 2021). doi:10.18129/B9.bioc.bioDist.

24. Fraley, C., Raftery, A. E., Scrucca, L., Murphy, T. B. & Fop, M. *mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation*. (2020).

25. Konopka, T. *umap: Uniform Manifold Approximation and Projection*. (2020).

26. Donaldson, J. *tsne: T-Distributed Stochastic Neighbor Embedding for R (t-SNE)*. (2016).