**Development of a single cell methylation summarization tool and application to lineage analysis of glioma-related macrophages**

**Masters Thesis**

**Presented to the Faculty of Biosciences of the Ruprecht-Karls-Universität Heidelberg**

**Andrew Lindsay**

**2021**

# Declaration

**This Thesis was written at the Deutsches Krebsforschungszentrum in the period from 02/06/2021 to 02/12/2021 under the supervision of Dr. Reka Toth**

**1st Examiner:** Prof. Dr. Christoph Plass

**Institute:** German Cancer Research Center (DKFZ), Division of Cancer Epigenomics

**2nd Examiner:** Ret. PD Dr. Odilia Popanda

**Institute:** German Cancer Research Center (DKFZ), Division of Cancer Epigenomics

**I herewith declare that I wrote this Masters Thesis independently, under supervision, and that I used no other sources and aids than those indicated throughout the thesis.**

|  |  |  |
| --- | --- | --- |
| 02/12/2021 |  | **A pair of glasses  Description automatically generated with medium confidence** |
| **Date** |  | **Signature** |

# Table of Contents

[1. Table of Contents 2](#_Toc87189540)

[2. List of Abbreviations and Acronyms 3](#_Toc87189541)

[3. List of figures 4](#_Toc87189542)

[4. Abstract 4](#_Toc87189543)

[5. Introduction 4](#_Toc87189544)

[5.1. Epigenetics 4](#_Toc87189545)

[5.1.1. Background 4](#_Toc87189546)

[5.1.2. Methylation 4](#_Toc87189547)

[5.2. Implications to cancer 5](#_Toc87189548)

[5.2.1. Glioma 6](#_Toc87189549)

[5.2.2. Tumor-associated macrophages 6](#_Toc87189550)

[5.3. Existing methods 8](#_Toc87189551)

[5.3.1. Whole genome bisulfite sequencing 8](#_Toc87189552)

[5.3.2. Single cell bisulfite sequencing 8](#_Toc87189553)

[5.4. Project aim 9](#_Toc87189554)

[6. Methods 9](#_Toc87189555)

[6.1. scMethrix 9](#_Toc87189556)

[6.1.1. Overview 9](#_Toc87189557)

[6.1.2. Object structure 10](#_Toc87189558)

[6.1.3. Data handling 10](#_Toc87189559)

[6.1.4. Analysis modules 10](#_Toc87189560)

[6.1.5. Benchmarking 11](#_Toc87189561)

[6.2. Data collection 11](#_Toc87189562)

[7. Data Availability 12](#_Toc87189563)

# List of tables

# List of figures

[Figure 1. Package Overview. 10](#_Toc87189903)

# List of Abbreviations and Acronyms

|  |  |
| --- | --- |
|  |  |
| 5mC | 5-methylcytosine |
| BED, bedgraph | Browser extensible data format |
| CpG, CG | cytosine - phosphate - guanine |
| CGI | CpG Island |
| DNAme | DNA methylation |
| DNMT | DNA methyltransferase |
| GBM | glioblastoma |
| GEO | gene expression omnibus |
| HDF, HDF5 | hierarchical data format |
| iPCA | iterative principal component analysis |
| IDH | isocitrate dehydrogenase |
| IL | interleukin |
| kNN | k-nearest neighbour |
| NA | not applicable |
| NK | natural killer |
| PCA | principal component analysis |
| R | R programming language |
| RF | random forest |
| RRBS | reduced representation bisulfite sequencing |
| sc-BS | single cell bisulfite sequencing |
| SNP | single nucleotide polymorphism |
| TAM | tumor-associated macrophage |
| TET | ten-eleven translocation |
| tSNE | t-distributed stochastic neighbor embedding |
| UMAP | uniform manifold approximation and projection |
| WGBS | whole genome bisulfite sequencing |

# Abstract

# Introduction

## Epigenetics

### Background

Epigenetics is the study of heritable and reversible changes of the genome that do not change the underlying DNA sequence1. These changes are fundamental for controlling gene expression in all organisms, and can occur through intrinsic methods (e.g., X-chromosome inactivation in females2) or external causes (e.g., exposure to ultra-violet radiation3). The molecular marks of epigenetics span several levels from overall chromatin structure, such as histone modifications or nucleosome positioning, down to changes of individual bases, primarily with methylation1.

### Methylation

Methylation is the most studied epigenetic modification. The nucleotide bases adenine and cytosine both readily accept methylation on N4and C4/C5, respectively4. The bulky methyl groups provide a physical block to binding of transcription factors and RNA polymerase5. As such, methylation is a major factor in gene expression in mammals. 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 DNA6,7. 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 genome8. 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 sites9. This density is called a CpG island (CGI). Due to the high presence, and their significance for gene regulation, mechanisms for adding and removing methylation to CpG sites must be strictly regulated.

These epigenetic marks can be *de novo* added by DNA Methyltransferase proteins (DNMT3a/b)10 or removed by ten-eleven translocation methylcytosine dioxygenases (TET family)11, so these modifications are considered reversible. These are necessary for normal development of tissues, from Dnmt3a having a key factor in embryonic stem cell differentiation10 to TET1/TET2 maintaining pluripotency of stem cells11.

Example of adding and removing.

Furthermore, these modifications can also be maintained through cell division by DNMT1, hence they are considered heritable10.

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 cases12). 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.

Epigenetics can also drastically change the phenotype of cells and induce a pro- or anti-cancer state.

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.

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

### Tumor-associated macrophages

Glioma has shown strong correlation with macrophages such as native microglia and infiltrating monocytes12. Up to 50% of the tumor bulk in glioblastoma is made up of these tumor-associated macrophages (TAMs)13. In response to brain lesions caused by glioma or other diseases, circulating lymphocyte antigen 6hi C-C chemokine receptor type 2+ (Ly‑6ChiCCR2+) monocytes are preferentially recruited and differentiate into microglia14,15. 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 onset16. The mechanisms for this are not clear, but interleukin-1β (IL-1β) has been implicated a key signal for plaque phagocytosis through transgenic overexpression17. This activation effect could have relevance for cancer progression, as glioblastoma has shown to aberrantly express IL-1β18, so increased recruitment of circulating monocytes could drastically affect patient outcome. Hence, identifying the types of macrophages involved and their functional capacities could lead to improved treatment.

Traditionally, macrophages have been divided into two polarized types: the classically activated pro-inflammatory M1 macrophages and the alternatively activated anti-inflammatory M2 macrophages19. However, *in situ*, macrophages show high plasticity with molecular characteristics and functions of both. Phenotypic markers exist to differentiate them (e.g., CD11b(+)CD209(+) for M2), but external stimulus can cause reprogramming between M-types. For instance, in cytokine-deficient medium, M1 macrophages can transition/polarize to M2 type and start expressing CD11b(+)CD209(+) markers20. Hence, immunophenotypic markers are poorly reliable in classifying macrophages *in situ*. Transcriptome can similarly be used to identify M-type19, but their inherent plasticity leaves a complex web of transcriptional and pathway differences that make lineage analysis difficult. Epigenetic phenotype has shown to be a strong method in differentiating closely related cells and providing 21.

Microglia are no exception.

Previous studies have tried to deplete macrophage

These tumor-associated macrophages (TAMs) have previously shown promise as a target for treatment22.

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 types23. 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 monocytes24.

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

## Existing methods

### Whole genome bisulfite sequencing

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

### Single cell bisulfite sequencing

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.

## Project aim

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, *methrix25*, 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

## scMethrix

### Overview

We introduce *scMethrix*, an R-based data storage and manipulation tool that is ideal for single cell methylation data. A complete overview of the package structure is show in Figure 1. A condensed workflow is provided in the Supple, as well as a comprehensive vignette that outlines the specifics of each function via a sample workflow with open-source single cell methylation data.

|  |
| --- |
| Text  Description automatically generated |
| *Figure 1.* Package Overview of scMethrix.  Single cell bisulfite data stored in BedGraph-based files can be imported via read\_beds() to be stored within a summarized experiment object. Numerous functions are available for quality control, filtering, and visualization. Multiple export formats are supported for further downstream analysis. |

### Object structure

The scMethrix experiment object extends the *SingleCellExperiment* container26 and can interface with many other packages in the Bioconductor ecosystem. It largely uses the *data.table*27 and *DelayedMatrixStats*28 R packages for quick and efficient operations.

As this is an S4 object29, multiple components are encapsulated into a single instance. This includes storing genomic regions of features, data from inputted or transformed assays, additional sample or feature information, as well as reduced dimensionality data.

### Data handling

**Data input:** scMethrix imports BedGraph-based file formats via the *read\_beds()* function. Pre-configured settings are available for many commonly used methylation calling tools, including Bismark30, MethylDackel31, methylCtools32, BisSNP33, and BS-Seeker234. Other calling tools can be supported via manual input of column indexes. Reference CpGs can optionally be input for alignment or to exclude CpG sites before importing. Providing coverage data is optional, but necessary for certain analysis functions (e.g., *mask\_by\_coverage()*) and export formats (e.g., *export\_bsseq()*).

**Data storage:** Experimental data can be stored either in-memory or stored on the hard-disk as an hierarchical data format object (HDF5; via *HDF5Array* R package35), and can be converted between each other, depending on system resources. Most external functions cannot interface with HDF5 data, so it will be cast as an R matrix before processing. Genomic coordinates are handled via the *GenomicRanges* R package36, and can be stored as stranded or unstranded, with the option to collapse the stranded during *read\_beds()*. 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 using *assays*(), as well as dimensionality reduction data using *reducedDim*().

**Data exporting:** Multiple common sequencing file types can be output with this package, including full compatibility with BedGraph, bigWig, and BSSeq. There is partial compatibility with Seurat, but as it is designed more for coverage data from RNA-seq experiments, there is limited functionality available for downstream analysis of *scMethrix* objects.

### Analysis modules

**Filtering:** Both quality control and region- or sample-based filtering is possible. Masking can be done by coverage, sample count, or CpG site 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*37), iterative principal component analysis (iPCA; via *missMDA*38), and random forest (RF; via *missForest*39). 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. The number of CpGs binned in each region will be stored in *rowData()*.

**Collapsing:** Similar samples can be collapsed via a specified *colData()* column. Typically, this column would be given by the user during *read\_beds()* from some type of external sorting (e.g., from a cell identification assay) or after clustering (see below). Like binning, mean or sum is used by default for calculations, but an arbitrary function can also be used.

**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*40). Using this metric, clusters can be generated via hierarchy or partitions (via base R), as well as model-based clustering (via *mclust*41). Other clustering algorithms and distance metrics can be used via arbitrary function.

**Dimensionality reduction:** For plotting and clustering visualization, dimensionality must be reduced. Using the *dim\_red\_scMethrix*() function, reduction can be done via PCA (via base R), uniform manifold approximation and projection (UMAP; via *umap*42), and t-distributed stochastic neighbor embedding (tSNE; via *tsne*43). The number of CpGs to use can be either by highest variance or randomly chosen. This data is stored in the experiment object for later plotting under *reduced\_dims*().

**Visualization:** Using *ggplot2*44, dataset characteristics can be plotted, including β-value/coverage (*plot\_violin()*, *plot\_density()*, *plot\_coverage()*) and quality control (*plot\_sparsity()*, *plot\_stats()*) as well as dimensionality reduction (*plot\_dim\_red()*). Shape or color of data points can be specified using *colData()* derived from clustering or previous user input.

### Benchmarking

**Parallelization:** Many of the functions used in scMethrix can utilize multiple cores. This can provide a substantial increase in performance at the cost of memory usage. Through the *parallel* and *doParallel* R packages45,46, both Windows and Unix-based systems are supported.

## Data collection

Raw data files were obtained from the Gene Expression Omnibus (GEO). Datasets contained glioma-related cell types, including monocyte, neutrophil, natural killer (NK) cells, eosinophils (GSE3506947, GSE8882448, GSE16684449), glia and neuron (GSE6635150), microglia-like (GSE12148351), glioma-related immune cells (GSE15150652), glioblastoma (IDH-WT and IDH-mut; GSE15150652), and glioma (). See Supplementary Table S1 for full breakdown. Additional processed data for GSE15150652 was obtained directly from the authors. For datasets where cell types are identified only by protein markers, archetypal cell type was used instead (e.g., CD45+ representing NK cells). As well, most studies were investigating other diseases (e.g., Alzheimer’s disease50), so only data from control subjects was used for study.

Methylation array data (e.g., Illumina 450k and EPIC) were converted to genomic positions via the Minfi R package53. All genome assemblies were translated to hg38 by the Rtracklayer::liftOver R package54. For data sourced from mouse models (GSE12148351), only the 19 420 probes conserved with humans were kept55, and it has been show that CpG methylation for glial cells is highly conserved between mice and humans56. Array data and bisulfite sequencing show correlation coefficients ranging from 0.95 to 0.9757, so are suitable for comparison.

# Data Availability

The datasets presented in this study can be found in online open source repositories. The names of the repository/repositories and accession number(s) can be found in the Methods section or in Supplementary Table S1. Scripts for data import and analysis are available on Github (<https://github.com/knacko/monobrainDNAme>).

# Supplemental Information

|  |
| --- |
|  |
| Supplemental Method 1. Condensed workflow for scMethrix  This workflow uses data from GEO accession GSE |

References1. Wolffe, A. P. & Matzke, M. A. Epigenetics: Regulation Through Repression. *Science* **286**, 481–486 (1999).

2. Panning, B. X-chromosome inactivation: the molecular basis of silencing. *Journal of Biology* **7**, 30 (2008).

3. de Oliveira, N. F. P., de Souza, B. F. & de Castro Coêlho, M. UV Radiation and Its Relation to DNA Methylation in Epidermal Cells: A Review. *Epigenomes* **4**, 23 (2020).

4. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38**, 23–38 (2013).

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

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

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

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

9. 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).

10. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247–257 (1999).

11. Yang, J., Bashkenova, N., Zang, R., Huang, X. & Wang, J. The roles of TET family proteins in development and stem cells. *Development* **147**, dev183129 (2020).

12. 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).

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

14. 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).

15. 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).

16. 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).

17. 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).

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

19. Sica, A. & Mantovani, A. Macrophage plasticity and polarization: in vivo veritas. *J Clin Invest* **122**, 787–795 (2012).

20. Tarique, A. A. *et al.* Phenotypic, functional, and plasticity features of classical and alternatively activated human macrophages. *Am J Respir Cell Mol Biol* **53**, 676–688 (2015).

21. Singh, O., Pratt, D. & Aldape, K. Immune cell deconvolution of bulk DNA methylation data reveals an association with methylation class, key somatic alterations, and cell state in glial/glioneuronal tumors. *Acta Neuropathologica Communications* **9**, 148 (2021).

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

23. 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).

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

25. Methrix: an R/Bioconductor package for systematic aggregation and analysis of bisulfite sequencing data | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/36/22-23/5524/6042753.

26. Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat Methods* **17**, 137–145 (2020).

27. Dowle, M. *et al.* *data.table: Extension of ‘data.frame’*. (2021).

28. Hickey, P., Pagès, H. & Lun, A. *DelayedMatrixStats: Functions that Apply to Rows and Columns of ‘DelayedMatrix’ Objects*. (Bioconductor version: Release (3.14), 2021). doi:10.18129/B9.bioc.DelayedMatrixStats.

29. Chambers, J. M. Object-Oriented Programming, Functional Programming and R. *Statist. Sci.* **29**, (2014).

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

31. Ryan, D. *MethylDackel - A (mostly) universal methylation extractor for BS-seq experiments.* (2021).

32. Hovestadt, V. *et al.* Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature* **510**, 537–541 (2014).

33. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data | Genome Biology | Full Text. https://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-7-r61.

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

35. Pagès, H. *HDF5Array: HDF5 backend for DelayedArray objects*. (Bioconductor version: Release (3.14), 2021). doi:10.18129/B9.bioc.HDF5Array.

36. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology* **9**, e1003118 (2013).

37. 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.

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

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

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

41. 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).

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

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

44. CRAN - Package ggplot2. https://cran.r-project.org/web/packages/ggplot2/index.html.

45. Team, R. C. & others. R: A language and environment for statistical computing. (2013).

46. Wallig, M., Corporation, M., Weston, S. & Tenenbaum, D. *doParallel: Foreach Parallel Adaptor for the ‘parallel’ Package*. (2020).

47. Reinius, L. E. *et al.* Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One* **7**, e41361 (2012).

48. Kennedy, D. W. *et al.* Critical evaluation of linear regression models for cell-subtype specific methylation signal from mixed blood cell DNA. *PLoS One* **13**, e0208915 (2018).

49. Hannon, E. *et al.* Assessing the co-variability of DNA methylation across peripheral cells and tissues: Implications for the interpretation of findings in epigenetic epidemiology. *PLoS Genet* **17**, e1009443 (2021).

50. Gasparoni, G. *et al.* DNA methylation analysis on purified neurons and glia dissects age and Alzheimer’s disease-specific changes in the human cortex. *Epigenetics Chromatin* **11**, 41 (2018).

51. Lund, H. *et al.* Competitive repopulation of an empty microglial niche yields functionally distinct subsets of microglia-like cells. *Nat Commun* **9**, 4845 (2018).

52. Gaiti, F. *et al.* Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE151506.

53. 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).

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

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

56. 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).

57. Zhao, M., Tuo, H., Wang, S. & Zhao, L. The Roles of Monocyte and Monocyte-Derived Macrophages in Common Brain Disorders. *BioMed Research International* **2020**, e9396021 (2020).