**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

[Declaration 1](#_Toc87641245)

[Table of Contents 2](#_Toc87641246)

[List of figures 3](#_Toc87641247)

[List of tables 3](#_Toc87641248)

[List of abbreviations 3](#_Toc87641249)

[1. Abstract 4](#_Toc87641250)

[2. Introduction 4](#_Toc87641251)

[2.1. Epigenetics 4](#_Toc87641252)

[2.1.1. Methylation 4](#_Toc87641253)

[2.1.2. Epimutatons 4](#_Toc87641254)

[2.2. Implications to disease 5](#_Toc87641255)

[2.2.1. Cancer 5](#_Toc87641256)

[2.2.2. Glioma 6](#_Toc87641257)

[2.2.3. Tumor-associated macrophages 6](#_Toc87641258)

[2.3. Quantifying methylation 8](#_Toc87641259)

[2.3.1. Whole genome bisulfite sequencing 8](#_Toc87641260)

[2.3.2. Single cell bisulfite sequencing 8](#_Toc87641261)

[2.4. Project aim 9](#_Toc87641262)

[3. Methods 9](#_Toc87641263)

[3.1. scMethrix 9](#_Toc87641264)

[3.1.1. Overview 9](#_Toc87641265)

[3.1.2. Object structure 10](#_Toc87641266)

[3.1.3. Data handling 10](#_Toc87641267)

[3.1.4. Analysis 11](#_Toc87641268)

[3.1.5. Benchmarking 15](#_Toc87641269)

[3.2. Data collection 15](#_Toc87641270)

[3.2.1. Raw data 15](#_Toc87641271)

[4. Results 15](#_Toc87641272)

[5. Discussion 15](#_Toc87641273)

[6. Future Outlook 15](#_Toc87641274)

[7. Conclusion 15](#_Toc87641275)

[8. Data Availability 16](#_Toc87641276)

[9. Supplementary Data 16](#_Toc87641277)

[10. References 16](#_Toc87641278)

# List of figures

[**Figure 1. Package Overview of scMethrix** 10](#_Toc87454340)

[**Figure 2. Structure of the scMethrix class.** 11](#_Toc87454341)

[**Figure 3. Workflow for analyzing single-cell data with scMethrix.** 13](#_Toc87454342)

[**Supplementary Figure 1. Runtime of single-threaded and parallel processing of the workflow in Supplementary Method 1.** 15](#_Toc87454816)

# List of tables

**No table of figures entries found.**

# List of abbreviations

|  |  |
| --- | --- |
|  |  |
| 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 |
| scBS-seq | 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

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.

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

### Epimutations

Example of adding and removing.

Example of heritable

Example of sickness by epigenetics

## Implications to disease

Methylation state can significantly 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 types12. Only 21 CpG sites overlapping 15 genes were necessary to generate a DNAme signature for pediatric- versus adult-onset lupus.

### 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 cases13). 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 monocytes13. Up to 50% of the tumor bulk in glioblastoma is made up of these tumor-associated macrophages (TAMs)14. While microglia are the resident macrophages, 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 into the brain, and differentiate into microglia-like cells15,16. 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 microglia differentiated from infiltrating monocytes are able to phagocytose the amyloid plaques that contribute to disease onset, whereas native microglia are unable to do so17. The mechanism behind this is not clear, but interleukin-1β (IL-1β) has been implicated a key signal for plaque phagocytosis through transgenic overexpression18. This activation effect could have relevance for cancer progression, as glioblastoma has shown to aberrantly express IL-1β19. If similar mechanisms exist, infiltrating monocytes may have an increased tumor fighting response to glioblastoma compared to the native microglia, which 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 types: the classically activated, pro-inflammatory, anti-tumor M1 macrophages and the alternatively activated, anti-inflammatory, pro-tumor M2 macrophages20. 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-types21. For instance, in cytokine-deficient medium, M1 macrophages can transition to M2 type and start expressing CD11b(+)CD209(+) markers22. Thus, immunophenotypic markers are poorly reliable in classifying macrophages *in situ*. Transcriptome can similarly be used to identify M-type21, but their inherent plasticity leaves a complex web of transcriptional and pathway differences that makes identification difficult. Epigenetic phenotype has shown to be a strong method in differentiating closely related cells as well as determine cell lineage via epigenetic memory, as described above23.

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

## Quantifying methylation

DNAme has several advantages over other types of genetic analysis: (1) It is cell type specific, unlike normal DNA. (2) It is more robust against transient variation than RNA expression. (3) A CpG site is either methylated or unmethylated, so there is no ambiguity compared to expression counts. (4) Special handling is not required, as is necessary with RNA. (5) Assays for DNAme are similar to normal DNA sequencing and can be easily integrated into a clinical setting. There are many different methods of DNAme quantification25, but two show the majority of use: array-based profiling and whole genome bisulfite sequencing.

### Array-based

The major disadvantage to this technique is the limited number of probes available. Only about 850K CpG sites can be profiled via the Illumina MethylationEPIC array, compared to the approximately 28M CpG sites in the genome. However, it does cover 99% of human genes, 95% of CpG islands, reliable between runs, and is generally consistent with the previous Illumina’s previous 450k and 27k probe arrays26. For complex regulatory, this may not be sufficient though, as some single CpG sites can drastically affect expression, so a technique with a wider genomic range may be necessary.

### Whole genome bisulfite sequencing

Whole genome bisulfite sequencing (WGBS) is the gold standard technique for studying the epigenomic landscape since its development in 199225. In this technique, sodium bisulfite deaminates unmethylated cytosines into uracil, whereas the 5mC bases are bisulfite conversion. Subsequent polymerase chain reaction (PCR) converts the newly generated uracils into thymines, and by comparing pre- and post-treatment sequencing, the resulting single nucleotide polymorphisms (SNPs) allow identification of the 5mC bases. This is highly reliable (>99% accuracy) and reproducible27. However, the treatment is very harsh; bisulfite treatment randomly induces single strand nicks, so up to 95% of the CpG sites lost during sequencing28, so a coverage of at least 30x is recommended.

Unfortunately, normal WGBS suffers some drawbacks due to its bulk processing. First, low population samples cannot give suitable coverage. If the cells cannot be grown, like some bacteria which are viable but not culturable, or if it may be dangerous to allow cell growth, such as cancer cells in a patient, this technique will not work. Secondly, it can be difficult to differentiate between heterogenous cells. Rare or aberrant cell types can have significant effect on mediating immune responses or ever the pathogenesis of cancer29. This cell might differ in only a few CpG sites, and thus lost as noise during WGBS. Single CpG hypermethylation has shown to have significant effect in breast cancer, appearing in tumor suppression genes at 10x the rate in breast patients versus controls30. Single cell DNAme quantification would allow this level of analysis.

### Single cell bisulfite sequencing

Whole genome single cell bisulfite sequencing (scBS-seq) has more recently been developed (2014)31, and serves to address the major drawbacks of bulk WGBS. In contrast, scBS-seq typically involves lysing of single cells, treatment with bisulfite, and the resulting DNA fragments are primed with tagged adapters. The cells can be then combined, and the PCR amplification will use indexed primers corresponding to the tagged adapters to generate multiple single-cell libraries at once. This technique is suitable for low population samples as well as detecting rare or aberrant cell types and can allow for deeper analysis where cell-to-cell differences are necessary.

Lineage of cells can be weakly studied using bulk WGBS. Distinct cell states can be observed, but does not have the resolution necessary for the mapping the trajectory of intermediate cell types32. Cells retain a strong epigenetic memory, as described above, and can be used to determine lineage with better precision than other single-cell based techniques (e.g., scRNA-seq).

While scBS-seq can allow for high resolution analysis, it some drawbacks. Data sparsity is a significant challenge, as a vast majority of CpGs will be undetected in each cell.

Grouping of homogenous cells is a common technique, and this can be achieved by methylation-specific clustering methods, such as *Epiclomal*33, and allow discovery of hidden subpopulations where bulk WGBS is unable to do so.

Finally, this is a new technique to the mainstream, and few tools adequately support it. Existing bulk WGBS tools are unable of handle sparsity and the resulting high proportion of missing values. Statistical analysis is complicated by missing data and may require specialized algorithms. As well, there is high memory requirement if stored in a dense matrix, as most also require. As more scBS-seq experiments are performed, new software tools will be needed to adequately handle single cell data. Therefore, 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, *methrix*34, 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, whether they be native microglia or differentiated from 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 Supplemental Method 1, as well as a comprehensive vignette that outlines the specifics of each function via a sample workflow using publicly available single cell methylation data (see [Data Availability](#_Data_Availability) section).

|  |
| --- |
| Text  Description automatically generated |
| Figure 1. Overview of the scMethrix package.  Single cell bisulfite data stored in BedGraph-based files can be imported via read\_beds() to be stored as a summarized experiment within the scMethrix 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*35 container and will interface with many other packages in the Bioconductor ecosystem. It largely uses the *data.table*36 and *DelayedMatrixStats*37 R packages for quick and efficient operations.

As this is an S4 object38, multiple components are encapsulated into a single instance. This includes genomic regions of features, data from inputted or transformed assays, additional sample or feature information, and reduced dimensionality data (Figure 2).

The *scMethrix* class has functions to collapse samples (e.g., homogenous cells from the same patient, or clustering of broad cell types) and bin CpGs (e.g., into CGIs or promoter regions). Subsequent operations are unaffected, as the general structure of the object does not change. For the purposes of this section, a CpG or sample can refer to one or more features or cells, respectively.

### Data handling

**Data input:** scMethrix imports BedGraph-based file formats via the *read\_beds()* function. Samples are contained in columns and CpGs are represented in rows. Data points with an NA value can be included. Pre-configured settings are available for many commonly used methylation calling tools, including Bismark39, MethylDackel40, methylCtools41, BisSNP42, and BS-Seeker243. Other calling tools can be supported via manual input of column indexes. Reference CpGs can optionally be inputted for alignment or exclusion of CpG sites before importing, or the CpG sites can be generated from the input files themselves. Providing coverage data is optional, as scBS-seq data typically has one or two reads per CpG per chromatid per cell, depending on strandedness. However, the coverage matrix is necessary for certain analysis functions (e.g., *mask\_by\_coverage()*) and export formats (e.g., *export\_bsseq()*). Experiment objects can also be built using subsets of files and later merged.

**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 package44), and can be converted between each other, depending on system resources. Most external functions cannot interface with HDF5 data, so it may be cast as a matrix before processing. Genomic coordinates are handled via the *GenomicRanges* R package45, and can be stored as stranded or unstranded, with the option to collapse the strands 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*(). The experiment object can be serialized and stored on-disk for ease of transfer or later analysis.

**Data exporting:** Multiple common sequencing file types can be output with this package, including full compatibility with BedGraph, metilene, 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. The *scMethrix* object can also be used directly with any package that supports *SingleCellExperiment* objects (e.g., scater46 or scran47, though like Seurat, these are designed primarily for RNA-seq data).

### Analysis

**Workflow:** A interactive HTML summary report can be generated for an initial verification of successful file import (see Supplemental Method 1). Then, numerous quality control and analysis modules are available to (1) verify data quality, (2) select relevant features, (3) reduce data, and (4) generate and visualize relationships between samples (Figure 3).

**Quality control:** As described above, sparsity is a key challenge for single cell methylome data. Low coverage CpG sites may be unreliable and can be masked by sample count or by coverage (if provided). Low variance of CpG beta values can indicate homogeneous methylation, which may not be useful for downstream analysis, and can be masked as well. The *mask\_scMethrix()* function is flexible for many different types of statistical filtering by either samples or CpGs. Masked sites can be removed via *remove\_uncovered()*.

**Feature selection:** Descriptive statistics can be generated to help identify usefulfeatures and samples. Genomic regions can be subset by sample, chromosome, or region. The *GenomicRanges* datatype allows easy generation of genomic windows as well as range-based set operations to isolate the target features.

**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 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 in binning is user customizable by assay. The list of CpGs binned in each region can optionally be stored in *rowData()*.

|  |
| --- |
| figure2 |
| **Figure 2. Structure of the scMethrix class.**  Multiple containers are present in the scMethrix object to store data from single cell methylation experiments. Metadata for CpGs is stored in rowData(), and sample information is stored in colData(). The rows and columns of each assay() represent each CpG and sample, respectively. Dimensionality reduction data is stored in reducedDims(). |

**Collapsing:** 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. As well, *colData()* can capture the names of samples in each collapsed group.

**Imputation:** Due to the sparsity of single cell data, imputation is typically used to fill the gaps after binning or collapsing. This package contains three methods of imputation: k-nearest-neighbor (kNN; via *impute*48), iterative principal component analysis (iPCA; via *missMDA*49), and random forest (RF; via *missForest*50). Other imputation functions can be used as well via an arbitrary input function. Imputation can be performed either by whole genome or by regions, but this should be done with the assumption of quasi-independence of those regions.

**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*51). Using this metric, clusters can be generated via hierarchy or partitions (via base R), as well as model-based clustering (via *mclust*52). 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 by PCA (via base R), uniform manifold approximation and projection (UMAP; via *umap*53), or t-distributed stochastic neighbor embedding (tSNE; via *tsne*54). 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*().

**Transformation:** Further operations to each data point can be done directly on assays. A new assay will be generated via some arbitrary transformation function, such as the included *binarize()* function, and stored in the experiment object. Unneeded assays, excluding the score matrix, can easily be removed from the *scMethrix* object.

**Visualization:** Using *ggplot2*55, 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.

|  |
| --- |
| figure3 |
| **Figure 3. Workflow for analyzing single-cell data with scMethrix.** |

### Benchmarking

Batch processing:

**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 packages56,57, both Windows and Unix-based systems are supported. See for runtimes on Supplemental Method 1.

## Data collection

### Raw data

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 (GSE3506958, GSE8882459, GSE16684460), glia and neuron (GSE6635161), microglia-like (GSE12148362), glioma-related immune cells (GSE15150663), glioblastoma (IDH-WT and IDH-mut; GSE15150663), and glioma (GSE15203564, 65). See Supplementary Table S1 for full breakdown of cell types and sample size. Additional processed data for GSE151506 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, for studies investigating methylation and disease (e.g., GSE66351; Alzheimer’s disease), only the data from control subjects was taken.

Annotation for methylation array data was obtained66–68, then converted to genomic positions via *Minfi*69. Array data and bisulfite sequencing show correlation coefficients ranging from 0.95 to 0.9728, so are suitable for comparison. All genome assemblies were translated to hg38 by the *liftOver*70. For data sourced from mouse models (GSE121483), only the 19 420 probes conserved with humans were kept71, and it has been show that CpG methylation for glial cells is highly conserved between mice and humans72.

Array probe data for glioma-specific CpGs was taken from by a previous TCGA bulk DNA methylation study73, and converted into hg38 genomic loci.

# Results

## Reference methylomes

### Array-based

Methylation arrays showed

# Discussion

# Future Outlook

# Conclusion

# Data Availability

Extended documentation and vignettes for *scMethrix* are also available on Github (<https://github.com/CompEpigen/scMethrix/>). 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](#_Data_collection) section or in Supplementary Table S1. Scripts for data import and analysis are available on Github (<https://github.com/knacko/monobrainDNAme>).

# Supplementary Data

|  |
| --- |
|  |
| Supplementary Figure 1. Runtime of select functions in single- and multi-threaded configuration.  The workflow is described in Supplementary Method 1. The workstation was equipped with a 2.9 GHz AMD Ryzen 7 4800H processor, 2x16GB of PC4 25600 RAM, and HDF5 data was stored on a 150,000 IOPS solid state drive. |

|  |
| --- |
|  |
| **Supplementary Table 1. Cell types obtained from the Gene Expression Omnibus (GEO) accessions included in the study.** |

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

# References

1. 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. 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).

13. 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).

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

15. 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).

16. 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).

17. 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).

18. 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).

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

20. Pan, Y., Yu, Y., Wang, X. & Zhang, T. Tumor-Associated Macrophages in Tumor Immunity. *Frontiers in Immunology* **11**, 3151 (2020).

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

22. 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).

23. 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).

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

25. Kurdyukov, S. & Bullock, M. DNA Methylation Analysis: Choosing the Right Method. *Biology (Basel)* **5**, 3 (2016).

26. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology* **17**, 208 (2016).

27. Ahn, J., Heo, S., Lee, J. & Bang, D. Introduction to Single-Cell DNA Methylation Profiling Methods. *Biomolecules* **11**, 1013 (2021).

28. Zhou, L. *et al.* Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. *Sci Rep* **9**, 10383 (2019).

29. Jindal, A., Gupta, P., Jayadeva & Sengupta, D. Discovery of rare cells from voluminous single cell expression data. *Nat Commun* **9**, 4719 (2018).

30. Böck, J. *et al.* Single CpG hypermethylation, allele methylation errors, and decreased expression of multiple tumor suppressor genes in normal body cells of mutation‐negative early‐onset and high‐risk breast cancer patients. *Int J Cancer* **143**, 1416–1425 (2018).

31. Smallwood, S. A. *et al.* Single-Cell Genome-Wide Bisulfite Sequencing for Assessing Epigenetic Heterogeneity. *Nat Methods* **11**, 817–820 (2014).

32. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* **21**, 31 (2020).

33. Souza, C. P. E. de *et al.* Epiclomal: Probabilistic clustering of sparse single-cell DNA methylation data. *PLOS Computational Biology* **16**, e1008270 (2020).

34. 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.

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

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

37. 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.

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

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

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

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

42. 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.

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

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

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

46. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).

47. Lun, A., Bach, K., Kim, J. K. & Scialdone, A. *scran: Methods for Single-Cell RNA-Seq Data Analysis*. (Bioconductor version: Release (3.14), 2021). doi:10.18129/B9.bioc.scran.

48. 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.

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

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

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

52. 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).

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

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

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

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

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

58. 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).

59. 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).

60. 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).

61. 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).

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

63. 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.

64. He, C. *et al.* Patient-derived models recapitulate heterogeneity of molecular signatures and drug response in pediatric high-grade glioma. *Nat Commun* **12**, 4089 (2021).

65. Golebiewska, A. *et al.* Patient-derived organoids and orthotopic xenografts of primary and recurrent gliomas represent relevant patient avatars for precision oncology. *Acta Neuropathol* **140**, 919–949 (2020).

66. Hansen, K. & Fortin, J.-P. IlluminaHumanMethylation27kanno.ilmn12.hg19. *Bioconductor* http://bioconductor.org/packages/IlluminaHumanMethylation27kanno.ilmn12.hg19/.

67. Hansen, K. IlluminaHumanMethylation450kanno.ilmn12.hg19. *Bioconductor* http://bioconductor.org/packages/IlluminaHumanMethylation450kanno.ilmn12.hg19/.

68. Hansen, K. IlluminaHumanMethylationEPICanno.ilm10b4.hg19. *Bioconductor* http://bioconductor.org/packages/IlluminaHumanMethylationEPICanno.ilm10b4.hg19/.

69. 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).

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

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

72. 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).

73. Ceccarelli, M. *et al.* Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* **164**, 550–563 (2016).