**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](#_Toc88849726)

[Table of Contents 2](#_Toc88849727)

[List of figures 4](#_Toc88849728)

[List of tables 4](#_Toc88849729)

[List of abbreviations 4](#_Toc88849730)

[1. Abstract 6](#_Toc88849731)

[2. Introduction 6](#_Toc88849732)

[2.1. Epigenetics 6](#_Toc88849733)

[2.1.1. Methylation 6](#_Toc88849734)

[2.1.2. Gene regulation 6](#_Toc88849735)

[2.1.3. Differentiation 7](#_Toc88849736)

[2.1.4. Epimutations 7](#_Toc88849737)

[2.2. Implications to disease 8](#_Toc88849738)

[2.2.1. Cancer 8](#_Toc88849739)

[2.2.2. Glioma 9](#_Toc88849740)

[2.2.3. Tumor-associated macrophages 9](#_Toc88849741)

[2.2.4. Potential for treatments 10](#_Toc88849742)

[2.3. Quantifying methylation 10](#_Toc88849743)

[2.3.1. Methylation arrays 10](#_Toc88849744)

[2.3.2. Whole genome bisulfite sequencing 11](#_Toc88849745)

[2.3.3. Single cell bisulfite sequencing 11](#_Toc88849746)

[2.4. Project aim 12](#_Toc88849747)

[3. Methods 12](#_Toc88849748)

[3.1. scMethrix 12](#_Toc88849749)

[3.1.1. Overview 12](#_Toc88849750)

[3.1.2. Data handling 13](#_Toc88849751)

[3.1.3. Analysis 14](#_Toc88849752)

[3.1.4. Additional Features 18](#_Toc88849753)

[3.1.5. Performance 18](#_Toc88849754)

[3.1.6. Validation 18](#_Toc88849755)

[3.2. Glioma tumor-associated macrophages 19](#_Toc88849756)

[3.2.1. Raw data 19](#_Toc88849757)

[3.2.2. Data processing 19](#_Toc88849758)

[3.2.3. Cell deconvolution 20](#_Toc88849759)

[3.2.4. Glioma TAMs 21](#_Toc88849760)

[3.2.5. Lineage reconstruction 21](#_Toc88849761)

[3.2.6. 22](#_Toc88849762)

[4. Results 22](#_Toc88849763)

[4.1. scMethrix 22](#_Toc88849764)

[4.2. Glioma tumor-associated macrophages 22](#_Toc88849765)

[4.2.1. Array-based 22](#_Toc88849766)

[5. Discussion 22](#_Toc88849767)

[5.1. scMethrix 22](#_Toc88849768)

[5.1.1. Validation 22](#_Toc88849769)

[6. Future Outlook 23](#_Toc88849770)

[7. Conclusion 23](#_Toc88849771)

[8. Data Availability 23](#_Toc88849772)

[9. Supplementary Data 23](#_Toc88849773)

# 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)

# List of tables

**No table of figures entries found.**

# List of supplementary information

[Supplementary Figure 1. Runtime of select functions in single- and multi-threaded configuration. 23](#_Toc89088614)

[Supplementary Table 1. Cell types obtained from the Gene Expression Omnibus (GEO) accessions included in the study. 23](#_Toc89088654)

[Supplementary Table 2. Data source and list of included cells for generating methylation signatures matricies and feature sets 24](#_Toc89088655)

[Supplementary Table 3. Gene IDs for glioma signatures genes as determine by 24](#_Toc89088656)

# List of abbreviations

|  |  |
| --- | --- |
|  |  |
| 5mC | 5-methylcytosine |
| BED, BedGraph | Browser extensible data format |
| CpG, CG | cytosine - phosphate - guanine |
| CGI | CpG Island |
| CTCF | CCCTC-binding factor |
| DNAme | DNA methylation |
| DNMT | DNA methyltransferase protein |
| GBM | glioblastoma |
| GEO | gene expression omnibus |
| HDF, HDF5 | hierarchical data format |
| iPCA | iterative principal component analysis |
| IDH | isocitrate dehydrogenase |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| kNN | k-nearest neighbour |
| mCpG | methylated CpG |
| MDMi | monocyte-derived microglia-like |
| 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 |
| TCGA | The Cancer Genome Atlas program |
| TET | ten-eleven translocation protein |
| TF | transcription factor |
| TSG | tumor-suppressing genes |
| TSS | transcription start site |
| 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 most organisms. Methylated adenine is weakly studied but widely known in bacteria and lower eukaryotes6,7, but is not considered to have any effect or significant presence in mammalian DNA8,9, where instead cytosine modifications are more common. 5-methyl-cytosine (5mC) is the dominant epigenetic mark in humans, with about 28 million sites present in the genome10. 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 sites11. This density is called a CpG island (CGI)5. 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 added *de novo* by DNA Methyltransferase proteins (DNMT3a/b)12,13 or removed by ten-eleven translocation methylcytosine dioxygenases (TET family)14, so these modifications are considered reversible. Recruitment of DNMT3A/B is not well understood, but is regulated by mechanisms including chromatin modifications, transcription levels, and the presence of non-coding RNAs or DNA-binding factors12. DNMT3A/B also aids in maintaining methylation state15, as demethylation can occur passively during errors in replication, or actively by numerous demethylation pathways16, including via TET proteins. As with DNMT3, there are many mechanisms for TET recruitment or inhibition, including distinct metabolites, cofactors, and post-translational modifications14. TET does not directly remove 5mC, but instead catalyzes hydroxylation into 5-hydroxymethylcytosine, which can later be removed by the base excision repair pathway. Free of the obstructing methyl group, transcription can proceed4.

### Gene regulation

Transcription of genes can be affected directly or indirectly by methylation1. As CpG islands are dense in promoter regions, transcription can be directly blocked via steric hinderance, as transcription factor (TF) binding sites are blocked by the bulky methyl molecules5. The major TF families E2F17, NF-κB18, and Myc19 all show methylation-sensitivity. As well, methylated CpGs (mCpGs) can indirectly repress transcription by recruiting methyl-binding proteins to compete for TF binding sites. Other chromatin remodeling proteins (e.g., histone deacetylases) are also recruited and can further impede transcription. Conversely, TFs have been found that preferentially bind to mCpGs *in vivo* (e.g., ARNT2, DIDO1, MEF2A and HOXA920), so methylation is not limited to only gene repression. However, *in vitro* experiments show contradictory evidence of pro-methylation TF binding, so it is unknown whether increased gene expression by mCpGs is a widespread effect21.

Gene products can also be affected by methylation via alternative splicing, as transcription factors also bind to exons and can similarly be blocked as in promoters4. In the human CD45 gene, CCCTC-binding factor (CTCF) binds to exon 5 and reduces the elongation rate of RNA pol II, thus favoring the inclusion of the exon into mature mRNA, but when CTCF binding motifs are methylated, exon 5 is typically excluded22. Though CD45 isoforms have shown to strongly influence leukocyte function, these mechanisms and effect of alternate exons are still unclear23. Alternative splicing can also occur via recruitment of splicing factors. Heterochromatin protein 1 can create a protein bridge with mCpGs, and is known enhancer and silencer for splicing alternative exons24.

### Differentiation

Differentiation of cells is strongly influenced by epigenetic factors. *Dnmt3*−/− embryonic stem (ES) cells that are terminally unmethylated (0.6% of normal) are still viable and can self-renew, but unable to initiate differentiation25. If differentiation is induced in hypomethylated *Dnmt1*−/− ES cells, apoptosis will occur25. Conversely, TET1−/− neural progenitor cells (NPCs) show impaired self-renewal but maintaining the differentiation capacity26. TET2−/− hematopoietic stem cells (HSCs) have increased self-renewal but bias differentiation towards myelomonocytic lineage27, and TET2−/− mice universally develop highly aggressive myeloid leukemia28.

Among non-mutant cells, there is a global trend towards hypermethylation during differentiation, including gamete-specific and pluripotency-associated genes29. However, specific loci remain hypomethylated in a cell-type-dependent manner30. Terminally differentiated cells generally maintain their methylation state29, and may even resist aberrant or artificially-induced demethylation, though with some shift in phenotype31. Depending on cell type, they may still proliferate and pass on their epigenetic identity32.

Methylation patterns are maintained through cell division, so they are considered heritable33. During DNA replication, proliferating cell nuclear antigen (PCNA) molecules associate with DNA polymerases at 3’ end of the replication fork. The PCNA molecules recruit NP95, which preferentially binds to the hemi-methylated regions resulting from semi-conservative replication. NP95 aids in localization of DNMT1 to methylate the newly replicated, unmethylated strand. This imparts an epigenetic memory into cells, and the lineage of a cell can be traced via their specific methylation patterns30,34,35. This also means that aberrant methylation can be passed to daughter cells as well.

### Epimutations

Epigenetic mutations, or epimutations, are aberrant chromatin states that can be caused by hypo- and hypermethylation of CpG sites. These can arise spontaneously, as like DNA replication, methyl transfer by DNMTs is not error-proof35. Spontaneous epimutations occur at a rate of 10-100x that of mutations during cell division, and can be caused by impairment of methylation machinery or by dilution of the classical methyl donor, S-adenosyl methionine36. While a single epimutation alone is unlikely to cause drastic phenotype changes, it can act as a traceable cell tag. This has previously been observed in continuously differentiating colon crypt cells32 as well as multiple types of leukocytes37. During replication, about 0.3-1.3 epimutations are expected to randomly occur36, allowing lineage mapping from hematopoietic stem cell to terminal differentiation34.

External factors can also cause aberrant methylation, such as ultraviolet (UV) radiation3. While methylation marks typically block polymerases, UV radiation can induce pyrimidine dimers preferentially at mCpGs38. These dimers do not block polymerases, allowing for error-free bypass. UV radiation has also shown to cause global hypomethylation as well as localized hypermethylation of many genes3. Environmental disasters may also drive epimutations. Nutrition has been linked to methylation39, and it has been suggested that past famines have disrupted methylation patterns, causing low birth weight and metabolic defects in the grandchildren of famine survivors40. Heavy metals are known to induce epimutations through both hypo- and hypermethylation of numerous loci, and similar effects were seen with persistent organic pollutants (POPs) like dichlorodiphenyltrichloroethan (DDT) and polychlorinated biphenyls (PCBs)41. This is further problematic since, as stated above, epimutations are heritable during cell replication, but are also heritable transgenerationally from parent to offspring as well. Diethylstilbestrol, a formerly used nonsteroidal estrogen medication prescribed to pregnant women but now known to induce aberrant methylation42, has shown links to breast cancer in their children43 and ADHD in their grandchildren44. Aberrant epimutations have significant, and likely understated, contribution to human disease.

## Implications to disease

There are numerous diseases solely attributed to epigenetic abnormalities. These often happen via genomic imprinting, where genes are differentially expressed based on which parent the gene was inherited from45. For instance, insulin-like growth factor 2 (IFG2) is only expressed from the paternal allele46. The maternal gene is methylated upstream which prevents CTCF binding and transcription, but the parental gene remains unmethylated to allow CTCF and enhancers to bind. The end result on the maternal gene is similar to X-chromosome inactivation and will never be expressed, though the latter happens via an unrelated process2. If a gene is incorrectly imprinted, disease can result45. The maternally-expressed gene 3 (MEG3) is an imprinted lncRNA-coding gene that regulates hepatic glucose production via the gluconeogenic FoxO1, and overexpression of MEG3 has shown to increase insulin resistance, a major characteristic of type 2 diabetes mellitus47. However, disease is not limited to epigenetic defects of individual genes. Immune cells show epigenetic reprogramming during development, activation, and differentiation30,48, so epimutations may indirectly affect the outcomes of all diseases.

### Cancer

Aberrant methylation is ubiquitous in cancer, typically showing global hypomethylation and localized hypermethylation in CpG islands49. Mutations in DNMT proteins have been found in multiple cancers (e.g., DNMT3A in 25% of acute myeloid leukemia cases and DNMT1 in 12% of uterine cancer cases50), and TET mutations are also common (e.g., up to 42% of chronic myelomonocytic leukemia cases51). This is problematic with pro-tumor and tumor-suppressing genes (TSGs), as single epimutations can have significant effect; hypermethylation of single CpGs is found in TSGs at 10x the rate in breast cancer patients versus controls52. Genetic imprinting also has shown links to cancer45. MEG3, described above, is also known to aid in tumor suppression and p53 activation, and is inactivated in most types of cancer53.

### Glioma

Glioma is a devastating, rapidly progressing disease with one of the poorest outcomes of any cancer54. Each year, about 250 000 individuals are diagnosed with glioma (1.4% of new cancer diagnosis)55. The disease progresses from mild episodic effects like headaches and seizures, to severe functional decline including motor dysfunction and neurocognitive failure54. Tumor samples are difficult to obtain, and a very small amount of tissue is taken. Mechanisms and risk factors for glioma incidence are not well understood, but epigenetics is strongly implicated in its development56. Six major glioma subtypes have been classified via methylation patterns, and there is strong correlation between these patterns and mutations of one specific gene also implicated in glioma (e.g., IDH3, K27, G34)57. As well, methylation of specific genes (e.g., MGMT58) drastically reduce effectiveness of certain anti-cancer drugs that work in tandem with the immune system in combating tumors.

Glioma tumor microenvironment is distinct from other types of tumors. The brain is physically protected from the rest of the body due to the blood-brain barrier, which can selectively prevent circulating molecules and cells from passing through59. Composition of immune cells drastically differs, with lymphocytes making up < 1% of immune cells in healthy brains60, compared to 70 - 90% of immune cells in the blood61. Instead, macrophages are the dominant immune cell, making about 80% of immune cells in the brain60. They also play a key role in cancer, as up to 50% of the tumor bulk in glioblastoma is made up of tumor-associated macrophages (TAMs)62.

### Tumor-associated macrophages

Microglia (MG) are the resident macrophages unique to the CNS63. These cells differentiate only during embryogenesis from primate primitive myeloid precursors that arise in the yolk sac, but survive with long lifespans and marginal local proliferation (< 2% of total)63. One could think this would be problematic when combatting an immune threat, as their limited population may become depleted. However, brain lesions caused by glioma or other diseases cause circulating monocytes (Ly‑6ChiCCR2+) to be preferentially recruited into the brain where they differentiate into microglia-like cells64,65. These monocyte-derived microglia-like (MDMi) cells are nearly indistinguishable but do show functional differences compared to their native counterparts and can be a significant factor for disease progression. For instance, in Alzheimer's disease, only MDMi cells are able to phagocytose the amyloid plaques that contribute to disease onset, whereas native MGs are unable to do so66. The mechanism behind this is not clear, but transgenic overexpression of interleukin-1β (IL-1β) in brain tissue has shown as a key signal for plaque phagocytosis67. Differential activation effects could have relevance for cancer progression, as glioblastoma cells aberrantly express a wide variety proteins, including IL-1β68. Identifying the types of macrophages present in glioma and their respective 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 macrophages69. However, *in situ*, macrophages (including MG) 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-types70. For instance, in cytokine-deficient medium, M1 macrophages can transition to M2 type and start expressing CD11b(+)CD209(+) markers71. Thus, immunophenotypic markers are weakly reliable in classifying macrophages *in situ*. Transcriptomes can be used to identify M-type *in vivo*70, but their inherent plasticity *in vitro* 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 cells72. Lineage can also be observed, such as monocyte-to-macrophage differentiation, as there are highly localized DNAme changes73.

### Potential for treatments

As mentioned previously, methylation can influence CD45 isoforms, and may be a target for treatment. In HIV, CD45 isoforms have shown significant effect on viral replication74. Targeting antibodies for one specific isoform, CD45-RO, virtually stopped RO-expressing microglial proliferation as well as strongly inhibiting overall HIV-1 production75. With glioma, CD45 isoforms haven’t been extensively studied, but it is still a key marker protein for MG (CD45low) and MDMi cells (CD45high)76. More research into glioma is necessary to discover the contributory effects of CD45 isoforms caused by aberrant methylation, and the epigenetic cell lineage of the TAMs that present them. This could allow indirect treatment of glioma by identifying and targeting of pre-tumorigenic macrophages, or direct treatment by antibodies against specific TAMs.

Specificity is an important factor for treatment, as macrophages are the dominant immune cells in the glioma tumor microenvironment, and paradoxically, both the major defender and driver of tumorigenesis. Clodronate-filled liposomes can be selectively induce apoptosis in TAMs77, as well as increase recruitment of non-MG TAMs into peripheral tumor regions78. Increased recruitment is generally undesirable but can provide additional treatment options. The blood-brain barrier can prevent passage of anti-cancer agents, but infiltrating macrophages can be hijacked and loaded with gold nanoshells containing tumor-targeting vectors79. CXCR4 is absent in MG, so can be used for targeting vector80, and inhibition of CXCR4 can decreasing macrophage infiltration and promote M1 phenotype in TAMs81. Studies of other inflammatory diseases have suggested targeting epigenetic modifiers to induce phenotype changes82, but this remains challenging due to plasticity of TAMs, as described above, so accurate identification of TAMs and their epigenetic pattern is essential.

## 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, in contrast to expression counts. (4) Special handling is not required, as is necessary with RNA. (5) Assays for DNAme easily integrated into a clinical setting as they are similar to normal DNA sequencing. NP95 and DNMT1 are not present during polymerase chain reaction (PCR) or other types of synthetic replication, so there is no direct passing of epigenetic marks, but additional treatments allow detection. There are many methods of DNAme quantification83, but two show the majority of use: microarray- and sequencing-based.

### Methylation arrays

Microarrays are attractive for many assays due to low cost, low DNA requirement, and low sample processing time83. This allows for high throughput of large numbers of clinical samples. The major disadvantage to this technique is the limited number of probes available. Only about 850K CpG sites can be profiled via the premier 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 arrays84. For complex regulatory systems, this may not be sufficient though, as single CpG sites can drastically affect gene expression, so a technique with a wider genomic range may be necessary.

### Whole genome bisulfite sequencing

Bisulfite sequencing (BS-seq) is the gold standard sequencing-based technique for studying the epigenomic landscape since its development in 199283. In this technique, sodium bisulfite deaminates unmethylated cytosines into uracil, whereas the 5mC bases are bisulfite conversion. Subsequent 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 reproducible85. However, the treatment is very harsh; bisulfite treatment randomly induces single strand nicks, so up to 95% of the CpG sites lost during sequencing86, so a coverage of at least 30x is recommended.

Unfortunately, normal BS-seq suffers some drawbacks due to bulk processing where numerous cells are sequenced in the same run. 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 as well pathogenesis of cancer87. This cell might differ in only a few CpG sites, and thus lost as noise during WGBS. As mentioned previously, single CpG hypermethylation is implicated in breast cancer52, and is difficult to discover via bulk WGBS. The ability to analyze the methylome of single cells would allow this level of analysis.

### Single cell bisulfite sequencing

Single cell bisulfite sequencing (scBS-seq) has more recently been developed (2014)88, 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.

Traditional techniques for studying lineage of cells (e.g., inducible reporter genes tracked through cell divisions89) can accurately determine heritage but cannot be applied *in situ* and typically require a significant time commitment. Transcriptomic-based techniques have shown reliable lineage tracing90, but requires many samples to accurately capture transitional states, and is difficult to determine far ancestors. Cells retain a strong epigenetic memory, as mentioned above30,34,35, and DNAme sequencing is becoming quick and economical, so DNAme lineage analysis can address the drawbacks of both traditional and transcriptomic tracing. Bulk WGBS has shown success, and distinct cell states can be observed, but does not have the resolution necessary for the mapping the trajectory of intermediate cell types91. scBS-seq is able to show individual differences between cells and accurately map lineage trajectory but is still an emerging technique and has its own drawbacks.

While scBS-seq can allow for high resolution analysis, data sparsity is a significant challenge, as a vast majority of CpGs will be undetected in each cell (> 90%)92. Robust statistical and clustering methods have been developed to mitigate this drawback, such as the Epiclomal software package93. This can allow discovery of hidden subpopulations where bulk WGBS is unable to do so. As sequencing techniques improve, this drawback will become irrelevant, but can be overcome currently by utilizing transcriptomic data as well to generate a comprehensive analysis on current cell state and past lineage94.

As scBS-seq is a relatively new technique to the mainstream, 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 necessary to 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, the methrix R package95, 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. This will involve three steps: (1) Generate a methylation signature matrix for the major cells in the tumor microenvironment, (2) Classify glioma and glioblastoma sub-types using known reference tools, (3) Classify sub-types of monocytes, macrophages, and microglia by epigenetic lineage, and (4) Use these new signatures to analyze bulk tumor cell samples for TAM presence. 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. An 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. |

The scMethrix experiment object extends the SingleCellExperiment96 container and will interface with many other packages in the Bioconductor ecosystem. It largely uses the data.table97 and DelayedMatrixStats98 R packages for quick and efficient operations.

As this is an S4 object99, 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 Bismark100, MethylDackel101, methylCtools102, BisSNP103, and BS-Seeker2104. 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()). Data from array-based formats, like 450k arrays, can also be imported directly via minfi::GRset objects. Experiment objects may 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 HDF5Array105), 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 GenomicRanges106, 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, metilene107. If coverage data is present, bigWig108, and BSSeq109 objects can be created. There is partial compatibility Minfi via GRset. Seurat110 objects can be created, 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., scater111 or scran112, 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 verify data quality, select relevant features, reduce data, and 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 regions. Masked sites can be removed via remove\_uncovered().

**Feature selection:** Descriptive statistics can be generated to help identify usefulfeatures and samples by get\_region\_summary(), get\_rowdata\_stats(), and get\_coldata\_stats(). 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.

**Merging:** It may be necessary to import data at different times or constructed for other input types, so experiments can be merged together. Metadata inside metadata(), rowData(), and colData() will be retained during merging, and overlapping metadata will be condensed if identical, or appended if different.

**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 for each 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. Genomic loci for CpGs are stored as GenomicRanges in rowData(). The rows and columns of each assay() represent each CpG and sample, respectively. Sample information is stored as a data.frame in colData(), as is the dimensionality reduction data stored in reducedDims(). The green and blue colored bands among each component indicates CpGs and samples, respectively. For synchronicity, each CpG and sample in assays must have a corresponding match among the row and column metadata. |

**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 impute113), iterative principal component analysis (iPCA; via missMDA114), and random forest (RF; via missForest115). 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 biodist116). Using this metric, clusters can be generated via hierarchy or partitions (via hclust and kmeans in base R), as well as model-based clustering (via mclust117). Other clustering algorithms and distance metrics can be used via arbitrary function. The discovered clusters will be stored in colData().

**Dimensionality reduction:** For plotting and clustering visualization, dimensionality must be reduced. Using the dim\_red\_scMethrix() function, reduction can be done by PCA (stats), uniform manifold approximation and projection (UMAP; via umap118), or t-distributed stochastic neighbor embedding (tSNE; via tsne119). 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. An assay can be transformed using some arbitrary function, such as the included binarize() function or any user-defined vectorized function, and then stored in the experiment object as a new assay. Unneeded assays, excluding the score matrix, can easily be removed from the scMethrix object. Assays can additionally be added to the experiment using the add\_assay() function. Compared to using a direct assignment operation, this function will perform validation checks of the input assay format.

**Visualization:** Using ggplot2120, dataset characteristics can be plotted, including methylation/coverage (plot\_violin(), plot\_density(), plot\_coverage()), quality control (plot\_sparsity(), plot\_stats()), as well as previously performed dimensionality reductions (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.  This describe the basic workflow progress |

### Additional Features

**Input validation:** A robust functionality for argument checking was implemented. This supports type- or value-checking where multiple comparisons can be made in a single function. Partial argument matching for strings is enabled for all functions without affecting downstream functions. Other functions for checking experiment or assay existence, and available threads for parallelization are also implemented. Each of these functions has a verbose error message which indicates the particular variable that has violated a condition as well as displays what the correct conditions are.

**Enhanced save function:** By default, HDF5 experiments are saved into a specified folder in two files: assays.h5 and se.rds. These filenames are hardcoded into HDF5Array objects. As well, all contents of the specified save folder will automatically be deleted during saving. The save\_scMethrix() function wraps the default save function, adding support for custom file names, as well as removing the auto-deletion of contents in the save folder. This function also allows saving of in-memory scMethrix objects, and these will be stored as .RDS files.

**Convenience functions:** Numerous accessory functions were implemented to increase ease-of-use. GenomicRanges, and the equivalent data.table objects via automatic casting, have additional options for splitting, subsetting, and creating windows. A simple stopwatch-like system was added to track multiple function running times at once.

### Performance

**Computations:** For most functions, assays are converted to data.table format for increased performance.

For calculation on HDF5 matricies, DelayedMatrixStats is used for all computations. These functions wrap the standard MatrixStats, so all non-data.table stats functions instead use DelayedMatrixStats functions. Storage on HDF5 is implicitly ‘chunked’ into sub-matrices for efficiency to access. In scMethrix, these are defined as individual columns

**Batch processing:** To prevent memory overflows, functions that require significant memory can be processed in batches. Batch processing is automatically used in the HDF5 backend.

**Parallelization:** Some 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 parallel121 and doParallel R packages122, both Windows and Unix-based systems are supported. However, data.table is internally parallelized, and does not support Windows forking.

### Usage

To show simple functionality of the package, a small example experiment was created from GSE5687988. From this, fourteen samples from C57BL/6Babr strain mice were obtained, consisting of E14 embryonic stem cells (ESCs) that were cultured in either serum/leukemia inhibitory factor (LIF) or 2i/LIF. The goal is to determine any differences between the populations due to the culture medium. See Supplemental Method 1 for condensed source code.

The raw data from GEO were imported into scMethrix as an in-memory experiment due the lower number of samples. Max-min normalization is automatically performed during conversion to β-value. Next, the methylation, sparsity, and coverage data were plotted to identify any irregularities such as non-bimodal β-density or too-high coverage. Then, quality control was performed to remove artifacts and low usefulness CpGs. Sites that were unread or read by only one sample were removed, as well as any site with a homogeneous methylation state. For single cell unstranded DNAme data, only 1 read is expected per allele, so CpG sites with an average coverage of > 2 were masked. Then, statistics were generated across samples, CpGs, and chromosomes.

To identify potential functional difference between the samples, the experiment was subset to include only CpGs located in promoter regions obtained from the Eukaryotic Promoter Database (EPD)123 for the Genome Reference Consortium mouse build 38 (mm10)124. From the transcription start site (TSS), 2000 bp upstream and 200 bp downstream defined the promoter regions. Each region was binned by averaging the methylation of the overlapping CpGs, and coverage was summed, then statistics were visualized for the new bins. To prepare for dimensionality reduction and clustering, missing methylation values were imputed via kNN using k = 5. Dimensionality reduction was performed with tSNE to check for unique clusters of the two input populations.

Differentially methylated promoters were then chosen as the promoters with the highest β-value variance (n = 500). These promoters were then hierarchically clustered with stats::hclust using k = 5, and visualized on the previous dimensionality reduction using the discovered clusters. To visualize patterns of methylation, a heatmap was generated for the two clusters. Promoter functions were then identified by ontology via Gene Ontology (GO) and pathway involvement via the Kyoto Encyclopedia of Genes and Genomes (KEGG).

## Glioma tumor-associated macrophages

### Raw data

**Datasets:** Methylationdata files were obtained from the Gene Expression Omnibus (GEO). This includes immune cell types, including monocyte, neutrophil, natural killer (NK) cells, eosinophils (GSE35069125, GSE49618126, GSE49667127, GSE88824128, GSE166844129); dendritic cells (GSE8345831); brain-specific cells, glia and neuron (GSE66351130), microglia72 and microglia-like (GSE121483131), and glioma-related immune cells (GSE15150694). Neoplastic cells were also obtained, including lower grade glioma (LGG; GSE104293132, GSE152035133, GSE137845134) and glioblastoma (IDH-WT and IDH-mut; GSE151506, GSE103659135). For lineage analysis, precursors to immune cells were also obtained, including hematopoietic stem cells (HSC; from bone and blood), multipotent progenitors (MPP; from bone and blood), common lymphoid progenitors (CMP), granulocyte-monocyte progenitors (GMP), common myeloid progenitors (CMP), and common monocyte progenitors (cMOP) from GSE49618126, GSE63409136, GSE8719630, GSE121483131. Non-glioma cancer cell lines were obtained from GSE68379137.

Data for microglia and GSE151506 was obtained directly from the respective authors. For datasets where cell types are identified only by protein markers, archetypal cell type was used instead (e.g., CD45+ representing NK cells138). As well, for studies investigating methylation and disease (e.g., GSE66351 with Alzheimer’s disease), only the data from control subjects was taken. See Supplementary Table 1 for full breakdown of cell types and sample sizes.

**Annotation:** Identification data for Illumina 450k/EPIC methylation array probes was obtained via Bioconductor139,140, then converted to genomic positions via Minfi141. Probes that have previously shown to be cross-reactive, low-quality genomic mapping, or target polymorphic CpGs were removed142. Array data and bisulfite sequencing are suitable for comparison (Pearson's r = 0.95 – 0.9786). Data sourced from mouse models (GSE121483) was limited 19,420 probes common with humans that are highly conserved in brain tissue143,144. All genome assemblies for sample and reference data was translated to Genome Reference Consortium Human Build 38 (hg38145) by liftOver146 using chains from AnnotationHub147. Human promoter loci were obtained from the EPD and limited to include only the most representative promoter for each gene (N = 16,455).

### Data processing

**Data importation:** For array-based data, datasets were generally provided as raw 450K/EPIC Illumina Methylation array (.idat) files or as β-values scraped from Series Matrix Files or SOFT family files associated with the GEO accession. This data was processed, quality control (QC) checked, and normalized by the single sample noob function in Minfi148, then converted into 450k arrays. All non-idat datasets were previously processed and used directly (see Supplementary Table S1 for processing details).

**Data filtering:** All array-based samples were subset to include only 450k probes using 6 bp windows around each probe to account for differences due to 0- and 1-based genomic coordinates, strand sense offset, and conversion between hg19 and hg38 reference genomes. Probes common to less than 5% cells or showed homogenous methylation (β-value SD < 0.05) were removed. Samples with less than 50,000 identified CpGs were filtered out (excluding WGBS samples). To bridge bisulfite and array-based data, a window of 1000 bp was defined for each probe, and overlapping windows were combined and tiled into a minimum number of 1000 bp windows, as per Chaligne et al. (2021)94. This resulted in 207,623 disjoint windows, and sample methylation values were calculated from the mean of all overlapping sites in each window.

### Cell deconvolution

**Signature Matrix:** A reference tumor microenvironment signature matrix for the brain was generated using sixteen different cell types (n = 232; Supplementary Table 2): B‑cells, CD4+ T‑cells, CD8+ T‑cells, dendritic cells, endothelial cells, eosinophils, glial cells, glioma tumor cells, granulocytes, microglia, monocytes, neurons, neutrophils, NK cells, Treg cells, and whole blood. The missing values in the 1000 bp windows, described above, were imputed by kNN with k = 10. Then, to find cell-specific signatures, methylCIBERSORT was used for deconvolution149. To identify features, a pair-wise comparison was done between each cell type using a modified version of FeatureSelect.V4, as described in Grabovska et al. (2020), to select the 100 top pairwise features using a median β-value difference of 0.25 and false discovery rate of 0.01. The discovered feature set was visualized to determine if each cell type is represented by a unique DNAme pattern by tSNE. Discovered features were further visualized by a heatmap with cluster::agnes150 hierarchical clustering.

**Microglia signatures:** Broad methylation patterns of microglia and glioma-related TAMs are weakly studied. To generate additional signatures, cells were identified from unannotated glioma-related CD45+ immune cells found in Chaligne et. al. (2021)94. As well, differentially expressed genes were previously found between microglia (e.g., P2RY13, GPR34) and peripheral macrophages (e.g., F10, EMILIN2; see Supplementary Table S3 for full list)151. In Chaligne et al., the unannotated immune cells were individually profiled by both scRNA-seq and scBS-seq, so the methylome data for microglia and TAMs can be found by classifying the cells based on their gene expression. To broaden potential future studies, the discovered signatures were also compared to mouse microglia as they are substantially easier to obtain for study (GSE121483).

**Signature Validation:** To measure deconvolution performance, the signature matrix was first benchmarked against homogenous cell samples. From GSE35069, 6 samples of CD4+ T cells, CD8+ T cells, NK cells, B cells, and monocytes, as well as 18 samples of granulocytes were obtained. The data was input into scMethrix, then imputed by kNN with k = 7. A collection of 100 synthetic cell mixtures was generated from a random uniform distribution of input cells, as described in Grabovska et al. (2020)151. Each synthetic mixture was classified by the signature matrix for the expected cell type proportion, and this was compared to the known cell type proportion of each synthetic mixture. CIBERSORT152 was used to generate the comparison, using 1000 iterations and quantile normalization. Correlation between the two proportions was visualized then calculated by Spearman’s rank method.

A second set of samples was also tested against the signature matrix. Samples consisting of known proportions of mixed cell types were obtained from two libraries, and cell proportion was verified by flow-cytometry (GSE110554) and FACS (GSE112618). These contained the same cell types as GSE35069 and the same validation method was used, but the β-values for the cell mixes were used directly instead of generating a synthetic mixture.

**Feature Analysis:** For each discovered feature, the genes overlapping with the respective feature window were collected. The molecular function for each gene was determined by cross-referencing the Gene Ontology (GO) project153.

This was generated using the 100 top pairwise features

### Glioma TAMs

**Glioma Classification:** Glioma cells were first validated by glioma-specific CpG probes taken from by a previous TCGA bulk DNA methylation study154 (n = 1300). These were compared with unrelated cancer cell lines to ensure unique grouping (n = 25; see Supplementary Table 2). To better further classify subtype, the TCGAbiolinks::gliomaClassifier155 was used.

**Glioma TAMs:** Each glioma sample was tested against the previously generated tumor microenvironment signature matrix to determine cell presence.

### Lineage reconstruction

There are no methylation-specific pipelines for lineage analysis as of yet, but some tools developed towards analyzing RNA-seq data will accept low dimensional data. Datasets containing immune cell precursors and terminally differentiated immune cells were combined (n = 339; see Supplementary Table 2), and subset to include only promoter regions.

Imputation by kNN (k = 9) was performed, then feature selection by methylCIBERSORT with the same settings as above. However, due to high sparsity, feature selection was done only using progenitor cells. Next, Dimensionality reduction was performed, and visualized to confirm unique cell groups, but five dimensions were retained after reduction for use in cell lineage reconstruction via slingshot156. Through minimum spanning trees and simultaneous principal curves, a connected and edge-weighted graph can be generated where each cell type is linked to the closest related cell types but will not allow any cycles. Applied to cell types, this will generate a pseudo lineage tree stemming from the common progenitor.

### 

# Results

## scMethrix

For the sample experiment, the data showed good initial quality. The β-values showed an expected bi-modal distribution, indicating homogenous methylation status for most CpG sites (Figure 4A). Samples with high unmethylated density showed low methylated density, and vice versa (Figure 4B). Coverage showed most sites with mean reads per CpG <= 2, and this was consistent between samples. Mean sparsity was 90.8 ± 1.84% (Figure 4C), so many sites were not read during sequencing. Of the 21.8M CpGs in the mm10 genome build, 9.3M (43.8%) CpGs were either unread or read by only one sample, so these were removed. Homogenously methylated CpGs were also removed (4.7M; 21.4%). Sites with coverage > 2 were removed due to likely technical contamination (695k; 3.2%). After removal, 7.1M CpGs remained, and sparsity dropped to 84.4 ± 3.9%, but with no significant change in overall sample sparsity pattern. Read numbers between chromosomes was generally consistent, though X, Y, and mitochondrial chromosomes showed significantly less reads (Figure 4D).

|  |
| --- |
|  |
| Figure 4. Initial quality checks of β-value, coverage, and sparsity of scWGBS data for mouse E14 embryonic stem cells.  **(A)** Density plot showing distribution of β-values. Distribution is a bi-modal, indicating either unmethylated (0) or methylated (0) status for most CpGs. A minor bump at β-value = 0.5 indicates some presence of few partially methylated CpG sites. **(B)**Scatterplot of sparsity for each sample. Read sparsity is consistent between samples. **(C)** Boxplot showing the number of reads for each chromosome. Mean = 92 ± 48k reads. **(D)** Density plot showing coverage of non-zero read CpG sites. Most sites having 1 or 2 reads with a low proportion having > 4 reads. Mean = 90.7 ± 2.4%. |

The cleaned data was binned into 19,535 promoter regions; 678 genes from the reference EPD promoters showed no reads after the quality control steps. The mean β-value density showed a significant skew in one sample of the 2i medium group (Figure 5A). This sample had a β-value = 0.10 ± 0.22, but most reads were not in highly variable regions. Feature selection chose 1000 promoters with the highest β-value variance, and the experiment was subset to include only these features. After feature selection of highly variable reads (n = 1000), the observed skew for the sample in the 2i medium group was significantly reduced, and it returned to a bi-modal density similar in magnitude and shape to the original input (Figure 5B). Dimensionality reduction by UMAP of the highly variable promoters showed a partition between the two medium groups, although two outliers from the serum medium group are present in the 2i group (Figure 5C).

|  |
| --- |
|  |
| Figure 5. β-value density post-binning and post-feature selection, and dimensionality reduction of selected features  **(A)** Density plot of samples after binning. A significant skew for one sample is obvious. **(B)** Density plot after selecting highly variable promoters (n = 1000). Previously observed skew is gone, and there is a near even split of methylated and unmethylated. **(C)** UMAP showing clusters of samples for 2i and serum medium. Two cluster are visible, with two serum outliers. |

To validate the groupings observed in the UMAP, hierarchical clustering was done. The two outlier serum samples were instead grouped into the 2i medium cluster (not shown). One serum outlier showed a significantly lower β-value than the group mean (β = 0.19 and = 0.68, respectively). The raw methylation values for this sample showed a similar spread, with β = 0.24 and = 0.66, respectively. The other outlier was approximately centered between the mean methylation of both mediums.

|  |
| --- |
|  |
| Figure 6. Heatmap of β-values among serum- and 2i‑cultured samples. Features were chosen from promoters with the highest variance (n = 1000). |

## Glioma tumor-associated macrophages

### Array-based

Methylation arrays showed

Dendritic cells pretty much only in brain during neuroinflammation

# Discussion

## scMethrix

New techniques in molecular genetics are continuously being developed, and data is being generated at an unprecedented rate. The creates an ever-growing need for reliable and efficient software analysis tools. Unfortunately, this often leads to bespoke software that fills a specific need at a certain time, but is unable to adapt to changing data needs, or is simply abandoned. To combat this, projects like Bioconductor aim to standardize this process, implementing guidelines to ensure that software is extensible, scalable, interoperable, and have high quality documentation. Software designed around these goals can help scientists produce reproduceable research as well as eliminates needless redundancy and wasted effort towards producing tools that may already exist. The package we have developed here, scMethrix, is intended to join that ecosystem. It is based around existing data structures in Bioconductor, the, as well as adapting to the common function conventions of BiocGenerics157. As it extends the SummarizedExperiment and SingleCellExperiment software classes, existing projects using those experiments objects will be able to utilize most of the functionality within this project without having to convert to an scMethrix object.

A major goal of this tool was flexibility. The user should not be limited to pre-programmed functions, as they may not work for specific use cases, or newer algorithms and techniques are developed, but the tool is never updates to use them. Most of the major operations in the package, such as binning, clustering, and transforming, all allow arbitrary function input with minimal conditions for usage. For instance, any imputation algorithm can be used so long as it accepts an input matrix and will return an identical-in-dimension output matrix with the same column/row ordering. The end user will not have to wait for the developers to implement new functions, and instead can do it themselves. Continuously adding functionality is also a concern, as software dependencies can spiral out of control, and creating issues once relied upon packages are no longer being updated. Developing with a mindset for flexibility can help mitigate these issues and increase functionality for the user.

As scBS-seq is relatively new, few methylation analysis tools support single cell data.

Most algorithms are designed for contiguous data, or data with very few missing values.

### Validation

In the validation, we expected to see a clear difference between the two culture methods (serum/LIF and 2i/LIF). LIF activates transcriptional factor Stat3 which prevents cell differentiation and promotes self-renewal158, whereas 2i contains kinase inhibitors against FGF signaling and this results in genome-wide hypomethylation159. We observed this with our test samples. The 2i-treated cells showed lower methylation across the genome, while the serum-cultured samples showed expected hypermethylation patterns.

The data downloaded from GEO was high quality. β-value showed the expected bi-modal distribution and coverage was <= 2 for most CpG sites as expected for scBS-seq. Sparsity was consistent with previous scWGBS studies92. Single nucleotide polymorphisms were not considered, as homozygosity is very high between mice of the same strain. Genotyping of multiple C57BL/6-derived strains, though not including C57BL/6Babr as used here, show 12 SNPs that vary160. Of those, 5 were intronic and the other 7 did not lie in promoter regions tested here. As well, intra-strain SNP variation is expected to be even less. Looking at the count per chromosome, there was nearly zero CpG sites read on the mitochondrial chromosome, but there is only 287 CpGs as per the mm19 genome build124, so this was not unexpected. However, the X chromosome surprisingly showed very low reads.

One significant outlier was present in the serum samples (β < 30% of the group mean). This could also possibly be due to LIF-resistance, as this could indirectly cause methylation to fall. The cells used in this experiment were E14 mouse ESCs, and β-values at E14 are about 0.6 around promoter regions161, which is slightly less than we observed ( = 0.68), so the cells used here may have an epiphenotype more similar to an E13 ESC. If a cell was LIF-resistant, it could start differentiating and continue demethylation until approximately E14.5, though this drop would be considerably less than what this outlier shows (= 0.19). Alternatively, this could be an example of detecting a rare cell where gene expression or methylation pattern shows significant difference from the characteristic cell. This may represent a mutator phenotype, characterized by widespread DNA or chromosomal changes162. This typical refers to somatic mutations, but widespread epigenomic changes would have comparable effect. Epimutations of genes regulating the methylation machinery can drastically change DNAme patterns12,27. If these cells replicate unchecked, a teratoma could result, a tumor made up of embryonic stem cells163.

# Future Outlooks

There is significant room for scMethrix to develop and expand. It does encompass most of the basic operations necessary after methylation calling For tasks like quality control masking, it is highly reliant on the users experience and knowledge to adequately mask the noisy or irrelevant sites, and abstain from masking off the informative sites. Implementing additional algorithms, such as optimal low variance algorithms164, would remove burden from the user as well as provide statistical evidence behind the masking steps. Expanding the vignette would also assist with this, but a grand solution would be a wiki for Bioconductor. Instead of having to search out packages and evaluate their functionality for a task, a centralized hub that package developers could collate their respective documentation and develop full pipelines for particular tasks. The user could instead search for their task and find all relevant information for performing it.

There is great potential for improving the backend data storage structures. Sparse data structures already exist within R, but these have the presupposition that missing values are equal to zero. These structures are stored as run length encoding vectors, where instead of a long string of zeros, it simply records the number of adjacent zeros between the non-zero numbers. This is acceptable for RNA-seq data where no reads mean no expression, but with methylation data, zero has a functional purpose. If one were to store methylation data in this data structures, it would suddenly de-methylate most of the genome. A new data structure could be built where zero can be an acceptable input and be fully functional with existing math operations. This would allow for significant memory savings, potentially allowing in-memory objects with hundreds of samples to exist on a simple workstation. Even more flexible would be a user-defined value for missing information. In a highly skewed or biased dataset, the user could select the dominant value to collapse, again freeing up computational resources. This is implicitly done already in HDF5 files, but there is unfortunately no interface in R to make use of it.

# 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 Table 2. Data source and list of included cells for generating methylation signatures matricies and feature sets |
|  |

|  |
| --- |
| Supplementary Table 3. Gene IDs for glioma signatures genes as determine by |
|  |

|  |
| --- |
| **Supplementary Method 1. Condensed workflow for analysis of** GSE56879 using **scMethrix.**  Example code shown here is not complete and only reflects the significant functional commands. See Data Availability for full source code. |
| # Download the GSM files from GEO ---------------------------------------  gsm\_list **<-** c**(**"GSM1370535", "GSM1370536", "GSM1370537", "GSM1370538", "GSM1370539", "GSM1370540", "GSM1370541", "GSM1370555", "GSM1370556", "GSM1370557", "GSM1370558", "GSM1370559", "GSM1370560", "GSM1370561"**)**  bed\_files **<-** sapply**(**gsm\_list, **function(**gsm**){**  GEOquery**::**getGEOSuppFiles**(**GEO **=** gsm, baseDir **=** data\_dir, makeDirectory **=** **FALSE**, filter\_regex **=** ".\*.cov.txt.gz"**)**  **})**  # Import the BedGraph files into scMethrix ------------------------------  scm **<-** scMethrix**::**read\_beds**(**  files **=** bed\_files, ref\_cpgs **=** mm10\_cpgs, chr\_idx **=** 1, start\_idx **=** 2, end\_idx **=** 3, beta\_idx **=** 4, M\_idx **=** 5, U\_idx **=** 6, stranded **=** **FALSE**, zero\_based **=** **FALSE**, colData **=** GSE\_colData  **)**  # Plots for initial quality checks (Figure 4.) --------------------------  density **<-** plot\_density**(**scm**)**  coverage **<-** plot\_coverage**(**scm,type**=**"density",max\_cov**=**10**)**  sparsity **<-** plot\_sparsity**(**scm,type**=**"scatter",pheno**=**"Sample"**)** |

# 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. Heusipp, G., Fälker, S. & Alexander Schmidt, M. DNA adenine methylation and bacterial pathogenesis. *International Journal of Medical Microbiology* **297**, 1–7 (2007).

7. Couturier, M. & Lindås, A.-C. The DNA Methylome of the Hyperthermoacidophilic Crenarchaeon Sulfolobus acidocaldarius. *Front Microbiol* **9**, 137 (2018).

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

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

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

11. 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).

12. Gagliardi, M., Strazzullo, M. & Matarazzo, M. R. DNMT3B Functions: Novel Insights From Human Disease. *Frontiers in Cell and Developmental Biology* **6**, 140 (2018).

13. 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).

14. 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).

15. Chen, T., Ueda, Y., Dodge, J. E., Wang, Z. & Li, E. Establishment and Maintenance of Genomic Methylation Patterns in Mouse Embryonic Stem Cells by Dnmt3a and Dnmt3b. *Mol Cell Biol* **23**, 5594–5605 (2003).

16. Nabel, C. S., Manning, S. A. & Kohli, R. M. The Curious Chemical Biology of Cytosine: Deamination, Methylation and Oxidation as Modulators of Genomic Potential. *ACS Chem Biol* **7**, 20–30 (2012).

17. Campanero, M. R., Armstrong, M. I. & Flemington, E. K. CpG methylation as a mechanism for the regulation of E2F activity. *PNAS* **97**, 6481–6486 (2000).

18. Lu, T. & Stark, G. R. NF-κB: regulation by methylation. *Cancer Res* **75**, 3692–3695 (2015).

19. Prendergast, G. C. & Ziff, E. B. Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. *Science* **251**, 186–189 (1991).

20. Spruijt, C. G. *et al.* Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized Derivatives. *Cell* **152**, 1146–1159 (2013).

21. Hu, S. *et al.* DNA methylation presents distinct binding sites for human transcription factors. *eLife* **2**, e00726 (2013).

22. Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74–79 (2011).

23. Tchilian, E. Z. & Beverley, P. C. L. Altered CD45 expression and disease. *Trends in Immunology* **27**, 146–153 (2006).

24. Yearim, A. *et al.* HP1 Is Involved in Regulating the Global Impact of DNA Methylation on Alternative Splicing. *Cell Reports* **10**, 1122–1134 (2015).

25. Jackson, M. *et al.* Severe Global DNA Hypomethylation Blocks Differentiation and Induces Histone Hyperacetylation in Embryonic Stem Cells. *Molecular and Cellular Biology* **24**, 8862–8871 (2004).

26. Zhang, R.-R. *et al.* Tet1 Regulates Adult Hippocampal Neurogenesis and Cognition. *Cell Stem Cell* **13**, 237–245 (2013).

27. Moran-Crusio, K. *et al.* Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell* **20**, 11–24 (2011).

28. An, J. *et al.* Acute loss of TET function results in aggressive myeloid cancer in mice. *Nat Commun* **6**, 10071 (2015).

29. Suelves, M., Carrió, E., Núñez-Álvarez, Y. & Peinado, M. A. DNA methylation dynamics in cellular commitment and differentiation. *Briefings in Functional Genomics* **15**, 443–453 (2016).

30. Farlik, M. *et al.* DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. *Cell Stem Cell* **19**, 808–822 (2016).

31. Vento-Tormo, R., Álvarez-Errico, D., Rodríguez-Ubreva, J. & Ballestar, E. Gains of DNA methylation in myeloid terminal differentiation are dispensable for gene silencing but influence the differentiated phenotype. *The FEBS Journal* **282**, 1815–1825 (2015).

32. Yatabe, Y., Tavaré, S. & Shibata, D. Investigating stem cells in human colon by using methylation patterns. *PNAS* **98**, 10839–10844 (2001).

33. Probst, A. V., Dunleavy, E. & Almouzni, G. Epigenetic inheritance during the cell cycle. *Nat Rev Mol Cell Biol* **10**, 192–206 (2009).

34. Salas, L. A. *et al.* Tracing human stem cell lineage during development using DNA methylation. *Genome Res* **28**, 1285–1295 (2018).

35. Shibata, D. Inferring human stem cell behaviour from epigenetic drift. *The Journal of Pathology* **217**, 199–205 (2009).

36. Horsthemke, B. Epimutations in Human Disease. in *DNA Methylation: Development, Genetic Disease and Cancer* (eds. Doerfler, W. & Böhm, P.) 45–59 (Springer, 2006). doi:10.1007/3-540-31181-5\_4.

37. Chu, M. *et al.* Inferring relative numbers of human leucocyte genome replications. *British Journal of Haematology* **141**, 862–871 (2008).

38. Ikehata, H. & Ono, T. The mechanisms of UV mutagenesis. *J Radiat Res* **52**, 115–125 (2011).

39. Kadayifci, F. Z., Zheng, S. & Pan, Y.-X. Molecular Mechanisms Underlying the Link between Diet and DNA Methylation. *Int J Mol Sci* **19**, 4055 (2018).

40. In utero undernourishment perturbs the adult sperm methylome and intergenerational metabolism. https://www.science.org/doi/10.1126/science.1255903.

41. Baccarelli, A. & Bollati, V. Epigenetics and environmental chemicals. *Curr Opin Pediatr* **21**, 243–251 (2009).

42. Block, K., Kardana, A., Igarashi, P. & Taylor, H. S. In utero diethylstilbestrol (DES) exposure alters Hox gene expression in the developing mullerian system. *The FASEB Journal* **14**, 1101–1108 (2000).

43. Hilakivi-Clarke, L. Maternal exposure to diethylstilbestrol during pregnancy and increased breast cancer risk in daughters. *Breast Cancer Research* **16**, 208 (2014).

44. Kioumourtzoglou, M.-A., Coull, B. A., O’Reilly, É. J., Ascherio, A. & Weisskopf, M. G. Association of Exposure to Diethylstilbestrol During Pregnancy With Multigenerational Neurodevelopmental Deficits. *JAMA Pediatrics* **172**, 670–677 (2018).

45. Plass, C. & Soloway, P. D. DNA methylation, imprinting and cancer. *Eur J Hum Genet* **10**, 6–16 (2002).

46. Bergman, D., Halje, M., Nordin, M. & Engström, W. Insulin-Like Growth Factor 2 in Development and Disease: A Mini-Review. *GER* **59**, 240–249 (2013).

47. Zhu, X., Wu, Y.-B., Zhou, J. & Kang, D.-M. Upregulation of lncRNA MEG3 promotes hepatic insulin resistance via increasing FoxO1 expression. *Biochem Biophys Res Commun* **469**, 319–325 (2016).

48. Placek, K., Schultze, J. L. & Aschenbrenner, A. C. Epigenetic reprogramming of immune cells in injury, repair, and resolution. *J Clin Invest* **129**, 2994–3005 (2019).

49. Ehrlich, M. DNA methylation in cancer: too much, but also too little. *Oncogene* **21**, 5400–5413 (2002).

50. Zhang, J., Yang, C., Wu, C., Cui, W. & Wang, L. DNA Methyltransferases in Cancer: Biology, Paradox, Aberrations, and Targeted Therapy. *Cancers (Basel)* **12**, 2123 (2020).

51. Abdel-Wahab, O. *et al.* Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies. *Blood* **114**, 144–147 (2009).

52. 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).

53. Al-Rugeebah, A., Alanazi, M. & Parine, N. R. MEG3: an Oncogenic Long Non-coding RNA in Different Cancers. *Pathol Oncol Res* **25**, 859–874 (2019).

54. Weller, M. *et al.* Glioma. *Nat Rev Dis Primers* **1**, 1–18 (2015).

55. Leece, R. *et al.* Global incidence of malignant brain and other central nervous system tumors by histology, 2003-2007. *Neuro Oncol* **19**, 1553–1564 (2017).

56. Gusyatiner, O. & Hegi, M. E. Glioma epigenetics: From subclassification to novel treatment options. *Seminars in Cancer Biology* **51**, 50–58 (2018).

57. Sturm, D. *et al.* Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. *Cancer Cell* **22**, 425–437 (2012).

58. Butler, M. *et al.* MGMT Status as a Clinical Biomarker in Glioblastoma. *Trends in Cancer* **6**, 380–391 (2020).

59. Daneman, R. & Prat, A. The Blood–Brain Barrier. *Cold Spring Harb Perspect Biol* **7**, a020412 (2015).

60. Mrdjen, D. *et al.* High-Dimensional Single-Cell Mapping of Central Nervous System Immune Cells Reveals Distinct Myeloid Subsets in Health, Aging, and Disease. *Immunity* **48**, 380-395.e6 (2018).

61. Kleiveland, C. R. Peripheral Blood Mononuclear Cells. in *The Impact of Food Bioactives on Health: in vitro and ex vivo models* (eds. Verhoeckx, K. et al.) 161–167 (Springer International Publishing, 2015). doi:10.1007/978-3-319-16104-4\_15.

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

63. Quail, D. F. & Joyce, J. A. The microenvironmental landscape of brain tumors. *Cancer Cell* **31**, 326–341 (2017).

64. 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).

65. 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).

66. 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).

67. 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).

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

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

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

71. 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).

72. 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).

73. Dekkers, K. F., Neele, A. E., Jukema, J. W., Heijmans, B. T. & de Winther, M. P. J. Human monocyte-to-macrophage differentiation involves highly localized gain and loss of DNA methylation at transcription factor binding sites. *Epigenetics & Chromatin* **12**, 34 (2019).

74. Cosenza‐Nashat, M. A., Kim, M., Zhao, M., Suh, H. & Lee, S. C. CD45 Isoform Expression in Microglia and Inflammatory Cells in HIV‐1 Encephalitis. *Brain Pathol* **16**, 256–265 (2006).

75. Kim, M.-O., Suh, H.-S., Si, Q., Terman, B. I. & Lee, S. C. Anti-CD45RO Suppresses Human Immunodeficiency Virus Type 1 Replication in Microglia: Role of Hck Tyrosine Kinase and Implications for AIDS Dementia. *J Virol* **80**, 62–72 (2006).

76. Badie, B. & Schartner, J. M. Flow cytometric characterization of tumor-associated macrophages in experimental gliomas. *Neurosurgery* **46**, 957–961; discussion 961-962 (2000).

77. Markovic, D. S., Glass, R., Synowitz, M., Rooijen, N. van & Kettenmann, H. Microglia Stimulate the Invasiveness of Glioma Cells by Increasing the Activity of Metalloprotease-2. *Journal of Neuropathology & Experimental Neurology* **64**, 754–762 (2005).

78. Poli, A. *et al.* Targeting glioblastoma with NK cells and mAb against NG2/CSPG4 prolongs animal survival. *Oncotarget* **4**, 1527–1546 (2013).

79. Baek, S.-K. *et al.* Photothermal treatment of glioma; an in vitro study of macrophage-mediated delivery of gold nanoshells. *J Neurooncol* **104**, 439–448 (2011).

80. Werner, Y. *et al.* Cxcr4 distinguishes HSC-derived monocytes from microglia and reveals monocyte immune responses to experimental stroke. *Nat Neurosci* **23**, 351–362 (2020).

81. Mercurio, L. *et al.* Targeting CXCR4 by a selective peptide antagonist modulates tumor microenvironment and microglia reactivity in a human glioblastoma model. *J Exp Clin Cancer Res* **35**, 55 (2016).

82. Jin, F. *et al.* Targeting epigenetic modifiers to reprogramme macrophages in non-resolving inflammation-driven atherosclerosis. *European Heart Journal Open* **1**, (2021).

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

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

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

86. 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).

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

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

89. Fox, D. T., Morris, L. X., Nystul, T. & Spradling, A. C. Lineage analysis of stem cells. in *StemBook* (Harvard Stem Cell Institute, 2008).

90. Kester, L. & van Oudenaarden, A. Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell* **23**, 166–179 (2018).

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

92. Mulqueen, R. M. *et al.* Highly scalable generation of DNA methylation profiles in single cells. *Nat Biotechnol* **36**, 428–431 (2018).

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

94. Chaligne, R. *et al.* Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. *Nat Genet* 1–11 (2021) doi:10.1038/s41588-021-00927-7.

95. 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.

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

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

98. 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.

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

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

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

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

103. 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.

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

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

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

107. Jühling, F. *et al.* metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res* **26**, 256–262 (2016).

108. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).

109. Hansen, K. D., Langmead, B. & Irizarry, R. A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology* **13**, R83 (2012).

110. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

111. 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).

112. 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.

113. 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.

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

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

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

117. 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).

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

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

120. Wickham, H. *et al.* *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. (2021).

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

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

123. Dreos, R., Ambrosini, G., Groux, R., Cavin Périer, R. & Bucher, P. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Research* **45**, D51–D55 (2017).

124. GRCm38 - mm10 - Genome - Assembly - NCBI. https://www.ncbi.nlm.nih.gov/assembly/GCF\_000001635.20/.

125. 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).

126. Cancer Genome Atlas Research Network *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**, 2059–2074 (2013).

127. Zhang, Y. *et al.* Genome-wide DNA methylation analysis identifies hypomethylated genes regulated by FOXP3 in human regulatory T cells. *Blood* **122**, 2823–2836 (2013).

128. 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).

129. 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).

130. 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).

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

132. Bady, P. *et al.* The DNA methylome of DDR genes and benefit from RT or TMZ in IDH mutant low-grade glioma treated in EORTC 22033. *Acta Neuropathol* **135**, 601–615 (2018).

133. 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).

134. 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).

135. Kickingereder, P. *et al.* Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma. *Neuro Oncol* **20**, 848–857 (2018).

136. Jung, N., Dai, B., Gentles, A. J., Majeti, R. & Feinberg, A. P. An LSC epigenetic signature is largely mutation independent and implicates the HOXA cluster in AML pathogenesis. *Nat Commun* **6**, 8489 (2015).

137. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016).

138. Thermo Fisher Scientific. Immune Cell Guide - Human and mouse antigens.

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

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

141. 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).

142. markgene/maxprobes: Methylation Array Cross-Reactive Probes version 0.0.2 from GitHub. https://rdrr.io/github/markgene/maxprobes/.

143. 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).

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

145. GRCh38 - hg38 - Genome - Assembly - NCBI. https://www.ncbi.nlm.nih.gov/assembly/GCF\_000001405.26/.

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

147. Morgan, M. *et al.* *AnnotationHub: Client to access AnnotationHub resources*. (Bioconductor version: Release (3.14), 2021). doi:10.18129/B9.bioc.AnnotationHub.

148. Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res* **41**, e90 (2013).

149. Chakravarthy, A. *et al.* Pan-cancer deconvolution of tumour composition using DNA methylation. *Nat Commun* **9**, 3220 (2018).

150. Maechler, M. *et al.* *cluster: ‘Finding Groups in Data’: Cluster Analysis Extended Rousseeuw et al.* (2021).

151. Grabovska, Y. *et al.* Pediatric pan-central nervous system tumor analysis of immune-cell infiltration identifies correlates of antitumor immunity. *Nat Commun* **11**, 4324 (2020).

152. Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol Biol* **1711**, 243–259 (2018).

153. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**, D325–D334 (2021).

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

155. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* **44**, e71 (2016).

156. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

157. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**, 115–121 (2015).

158. Niwa, H., Burdon, T., Chambers, I. & Smith, A. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev* **12**, 2048–2060 (1998).

159. Ficz, G. *et al.* FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency. *Cell Stem Cell* **13**, 351–359 (2013).

160. Mekada, K. *et al.* Genetic Differences among C57BL/6 Substrains. *Experimental Animals* **58**, 141–149 (2009).

161. Zhao, L. *et al.* The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. *Genome Res* **24**, 1296–1307 (2014).

162. Loeb, L. A. A mutator phenotype in cancer. *Cancer Res* **61**, 3230–3239 (2001).

163. Blum, B. & Benvenisty, N. The Tumorigenicity of Human Embryonic Stem Cells. in *Advances in Cancer Research* vol. 100 133–158 (Academic Press, 2008).

164. Meng, H. *et al.* A statistical method for excluding non-variable CpG sites in high-throughput DNA methylation profiling. *BMC Bioinformatics* **11**, 227 (2010).