INAUGURAL – DISSERTATION

submitted
to the

Combined Faculty of Mathematics,
Engineering, and Natural Sciences
of
Ruprecht-Karls-University, Heidelberg

for the degree of
Doctor of Natural Sciences

Put forward by

Yunhee Jeong, M.Sc.
born in Seoul, Republic of Korea

Oral examination: _____

Cell-type deconvolution model for read-level DNA methylomes

Advisor: PD Dr. Karl Rohr
Prof. Dr. Christoph Plass

Abstract

The cell-type composition in bulk samples serves as key evidence for examining disease progression, phenotypic characterisation and treatment responses. Therefore, cell-type deconvolution has been spotlighted as a computational approach to estimating cell-type composition in bulk samples. DNA methylation (DNAm) has been broadly used as epigenetic marks for cell-type deconvolution because it carries cell type-specific signals at CpG sites in mammal genomes. In particular, cell-type deconvolution methods using DNAm data can be used for tumour purity estimation due to distinctive DNAm patterns dominantly found in tumour cells.

There are two major approaches for generating DNAm data: sequencing-based and array-based profiling. Sequencing-based DNAm data, which profiles every single DNAm pattern in the format of sequencing reads, provides broader genomic coverage and better captures rare cell-type signals compared to array-based data. Thus, sequencing-based DNAm data is more suitable for accurate cell-type deconvolution compared to array-based data which contains average methylation levels only at designated CpG sites. Nevertheless, so far, array-based data has been the primary target of cell-type deconvolution methods because the matrix shape makes it easier to apply linear algebraic algorithms. Hence, sequencing-based cell-type deconvolution still needs intensive explorations to accomplish accurate estimation of cell-type compositions by taking the benefits of read-level methylation patterns into account.

In this thesis, we introduce a new sequencing-based cell-type deconvolution method using DNAm data and perform a systematic evaluation of existing cell-type deconvolution methods. We divide the evaluation into two major steps of sequencing-based cell-type deconvolution: informative region selection and cell-type composition estimation. In order to cover diverse scenarios of biological samples, we test the methods using DNAm data from mouse neurons and human tumours. The evaluation shows that existing sequencing-based methods do not outperform array-based methods despite having the advantage of exploiting read-level methylomes. This underscores the need for the development of new sequencing-based cell-type deconvolution methods that accurately identify cell type-specific methylation patterns and eliminate confounding factors.

To address the limitations of existing methods, we developed the deep learning method *MethylBERT* based on Bidirectional Encoder Representations from Transformers (BERT). The proposed method is specifically designed for tumour purity estimation. MethylBERT

classifies sequencing reads into tumour and normal cell types, and infers the proportion of tumour cell type via maximum likelihood estimation. The tumour purity is inferred with a likelihood function computed using the estimated posterior probability of cell types. We also employ the Fisher information to calculate the precision of MethylBERT tumour purity estimation. Furthermore, in order to address dissimilar region-wise cell-type distributions in a bulk sample, we developed an algorithm to adjust the estimated tumour purity by minimising the skewness of inferred region-wise tumour purities.

We thoroughly evaluate MethylBERT and perform a comparison with previous methods. The evaluation demonstrates the good performance of the proposed method for read-level methylation pattern classification and estimation of tumour purity. In particular, the read-level methylation pattern classification shows that MethylBERT outperforms other methods regardless of the pattern complexity, read length, and read coverage. Our investigation also includes an analysis of the model training and the estimated cell-type posterior probabilities. We also explore which kinds of DNA sequence features are learnt during MethylBERT pre-training and emphasise the importance of pre-training. In addition, we show that MethylBERT is capable of detecting rare tumour signals by yielding accurate tumour purity estimation results for bulk samples with a very low tumour percentage ($<1\%$). This demonstrates the potential of MethylBERT for non-invasive early cancer diagnostics via blood tests, and accurate circulating tumour DNA analyses.

Zusammenfassung

Die Zusammensetzung von Zelltypen in Massenproben dient als wichtiger Nachweis für die Untersuchung des Krankheitsverlaufs, der phänotypischen Charakterisierung und des Ansprechens auf die Behandlung. Daher wurde die Zelltyp-Deconvolution als rechnerischer Ansatz zur Bestimmung der Zelltyp-Zusammensetzung in Massenproben hervorgehoben. DNA-Methylierung (DNAm) wird häufig als epigenetische Markierung für die Zelltyp-Deconvolution verwendet. Sie erzeugt zelltyp-spezifische Signale an CpG-Stellen in Säugetiergenomen. Insbesondere Zelltyp-Deconvolutionsmethoden unter Verwendung von DNAm-Daten können aufgrund der in Tumorzellen vorherrschenden charakteristischen DNAm-Muster zur Bestimmung der Tumorreinheit verwendet werden.

Es gibt zwei Hauptansätze zur Generierung von DNAm-Daten: sequenzierungsbasiertes und array-basiertes Profiling. Sequenzierungsbasierte DNAm-Daten, die jedes einzelne DNAm-Muster im Format von *Sequenzierungs-Reads* profilieren, bieten eine breitere genomische Abdeckung und erfassen seltene Zelltypsignale besser als Array-basierte Daten. Daher eignen sich sequenzierungsbasierte DNAm-Daten besser für eine genaue Zelltyp-Deconvolution als Array-basierte Daten, die durchschnittliche Methylierungsgrade nur an bestimmten CpG-Stellen enthalten. Dennoch waren Array-basierte Daten bisher das Hauptziel von Zelltyp-Deconvolutionsmethoden, da die Matrixform die Anwendung Algorithmen aus der linearen Algebra erleichtert. Daher bedarf die sequenzierungsbasierte Zelltyp-Deconvolution noch intensiver Erforschung, um eine genaue Schätzung der Zelltyp-Zusammensetzungen unter Berücksichtigung der Vorteile von Methylierungsmustern auf Read-Level zu erreichen.

In dieser Arbeit stellen wir eine neue Zelltyp-Deconvolutionsmethode, welche sequenzierungsbasiert ist und DNAm-Daten verwendet. Zudem führen wir eine systematische Bewertung bestehender Zelltyp-Deconvolutionsmethoden durch. Wir unterteilen die Auswertung in zwei Hauptschritte der sequenzierungsbasierten Zelltyp-Deconvolution: informative Regionsauswahl und Schätzung der Zelltyp-Zusammensetzung. Um verschiedene Szenarien biologischer Proben abzudecken, testen wir die Methoden anhand von DNAm-Daten von Mausneuronen und menschlichen Tumoren. Die Auswertung zeigt, dass bestehende sequenzierungsbasierte Methoden Array-basierte Methoden nicht übertreffen, obwohl sie den Vorteil haben, Methylome auf Read-Level zu nutzen. Dies unterstreicht die Notwendigkeit der Entwicklung neuer sequenzierungsbasierter Zelltyp-Deconvolutionsmethoden, die zelltypspezifische Methylierungsmuster genau identifizieren und Störfaktoren beseitigen.

Um die Einschränkungen bestehender Methoden zu beseitigen, haben wir die Deep Learning Methode *MethylBERT* entwickelt, die auf *Bidirectional Encoder Representations from Transformers (BERT)* basiert. Die vorgeschlagene Methode ist speziell für die Schätzung der Tumorreinheit konzipiert. MethylBERT klassifiziert Sequenzierungsablesungen in Tumor und normale Zelltypen über eine Maximum-Likelihood-Schätzung, um den Anteil der Tumorzelltypen zu bestimmen. Die Reinheit des Tumors wird mit Hilfe einer Wahrscheinlichkeitsfunktion bestimmt, die anhand der geschätzten A-posteriori-Wahrscheinlichkeit der Zelltypen berechnet wird. Wir verwenden auch die Fisher-Information, um die Präzision der MethylBERT-Tumorreinheitsschätzung zu berechnen. Um unterschiedliche regionale Zelltypverteilungen in einer Massenprobe zu berücksichtigen, haben wir außerdem einen Algorithmus entwickelt, um die geschätzte Tumorreinheit anzupassen, indem die Schiefe der abgeleiteten regionalen Tumorreinheiten minimiert wird.

Wir evaluieren MethylBERT gründlich und führen einen Vergleich mit früheren Methoden durch. Die Bewertung zeigt die gute Leistung der vorgeschlagenen Methode zur Klassifizierung von Methylierungsmustern auf Read-Level und zur Schätzung der Tumorreinheit. Insbesondere zeigen die Klassifizierung des Methylierungsmusters auf Read-Level, dass MethylBERT andere Methoden übertrifft, unabhängig von der Musterkomplexität, der Read-Länge und der Read-Abdeckung. Unsere Untersuchung umfasst auch eine Analyse des Modelltrainings und der geschätzten Zelltyp-Posteriori-Wahrscheinlichkeiten. Wir untersuchen auch, welche Arten von DNA-Sequenzmerkmalen während des MethylBERT-Vortrainings gelernt werden, und betonen die Bedeutung des Vortrainings. Darüber hinaus zeigen wir, dass MethylBERT in der Lage ist, seltene Tumorsignale zu erkennen, indem es genaue Ergebnisse zur Schätzung der Tumorreinheit für Massenproben mit einem sehr geringen Tumoranteil (<1%) liefert. Dies zeigt das Potenzial von MethylBERT für die nicht-invasive Frühdiagnose von Krebs durch Bluttests und genaue Analysen zirkulierender Tumor-DNA.

Acknowledgements

I am grateful for the enriching experience of completing my doctoral studies in the Faculty of Mathematics and Computer Science at Heidelberg University, and in the Division of Cancer Epigenomics at the German Cancer Research Center (DKFZ). I extend my sincere appreciation to my co-supervisors, Prof. Dr. Pavlo Lutsik[1,2], PD Dr. Karl Rohr[3], and Prof. Dr. Christoph Plass[1], whose support was instrumental in both the scientific and personal aspects of my journey. Special thanks to Prof. Dr. Verena Wolf[4], who generously contributed as an external member of the Thesis Advisory Committee (TAC). Their consistent support provided a solid foundation for my doctoral studies, allowing me to successfully attain my PhD.

The benchmarking study involved collaborative efforts with Reka Toth[1], Kersten Breuer[1], Marlene Ganslmeier[1], Lisa Barros de Andrade e Sousa[5], and Dominik Thalmeier[5]. Again, special thanks to Marie Piraud who is the leader of the Helmholtz AI consultant team. I express my gratitude for their invaluable contributions to the project, which significantly contributed to its successful completion and subsequent publication.

I also extend my thanks to Jonathan Ronen[6], Wolfgang Kopp[6], and Altuna Akalin[6] for their collaboration on the project involving the development of the single-cell multi-omics integration method, scMaui. I appreciate their warm welcome into their group as an external collaborator, broadening my perspectives in the fields of bioinformatics and machine learning. I would like to acknowledge the support received from the Helmholtz Information & Data Science Academy (HIDA) Trainee Network program.

In addition to the individuals mentioned above, I extend my gratitude to all the current and former members of the Division of Cancer Epigenomics at DKFZ and the group of Computational Cancer Biology and Epigenomics at KU Leuven. Their collective efforts have contributed to creating an encouraging and friendly atmosphere, shaping the backdrop of my PhD. I especially acknowledge Elisabeth Pezzuto for her time and efforts in

[1]Division of Cancer Epigenomics, German Cancer Research Center (DKFZ), Heidelberg, Germany
[2]Department of Oncology, KU Leuven, Leuven, Belgium
[3]Biomedical Computer Vision Group, Heidelberg University, BioQuant, IPMB, Heidelberg, Germany
[4]Modeling and Simulation Group, Saarland University, Saarland, Germany
[5]Helmholtz AI, Helmholtz Munich, Neuherberg, Germany
[6]Bioinformatics and Omics Data Science Platform, Berlin Institute for Medical Systems Biology (MDC-BIMSB), Berlin, Germany

proofreading this thesis. My friend Hojin deserves acknowledgement for introducing me to the BERT model, sparking the primary project of my PhD. I would like to express my appreciation to Prof. Efrat Shema and Prof. Guy Ron from the Weizmann Institute and the Hebrew University of Jerusalem for their guidance in the Cancer-TRAX project.

Beyond the academic realm, my steadfast friend group, 'Good Boys,' provided constant companionship, ensuring I was never alone during my study abroad. Having an old friend, Jeongmin, in Germany warms my heart every holiday season. My heartfelt thanks go to my friends in South Korea—Jihyun, Junku, Hyerim, Jiwon, and Inbee—who have consistently supported me from afar, spanning a distance of 8,602 km. I reserve one of the major honours for Valentin Maurer, who not only shared insightful thoughts about science but also provided encouragement in every aspect of my life throughout my PhD journey.

Foremost, my deepest gratitude goes to my family members, especially my parents, whose unwavering support and presence have played an important role in my personal growth. I take immense pride in my brother's career achievements and appreciate him for standing by our parents in South Korea during my time away. To my grandparents, aunties, uncles, and cousins, I owe cherished memories that have served as powerful sources of encouragement throughout my study abroad experience.

In conclusion, I extend my sincere thanks to everyone who has played a role, either directly or indirectly, in my educational journey—from my early years in primary school to the completion of my doctorate. The dedication of teachers, lecturers, professors, mentors, and supervisors has paved the way for the completion of my studies. I am committed to carrying forward the valuable lessons they imparted, both in science and life, and strive to contribute to creating a better world through my endeavours.

Contents

List of Figures

List of Tables

Acronyms

ASM	Allele-specific Methylation.
AUC	Area under the Curve.
BERT	Bidirectional Encoder Representations from Transformers.
CDF	Cumulative Distribution Function.
cfDNA	Cell-free DNA.
CGI	CpG Island.
CRC	Colorectal Cancer.
ctDNA	Circulating Tumour DNA.
DLBCL	Diffuse Large B-cell Lymphoma.
DML	Differentially Methylated Locus.
DMR	Differentially Methylated Region.
DNAm	DNA Methylation.
DNMT	DNA-methyltransferase.
EM	Expectation-Maximisation.
GEO	Gene Expression Omnibus.
GRU	Gated Recurrent Unit.
HMM	Hidden Markov Model.
LSTM	Long Short-Term Memory.
MAE	Median Absolute Error.
MAPE	Median Absolute Percentage Error.
MLE	Maximum Likelihood Estimation.
MLM	Masked Language Model.
NLP	Natural Language Processing.
NMF	Non-negative Matrix Factorisation.
NSP	Next Sentence Prediction.

PDAC Pancreatic Ductal Adenocarcinoma.
PMD Partially Methylated Domain.
PMF Probability Mass Function.

RNN Recurrent Neural Network.
RRBS Reduced Representation Bisulfite Sequencing.

WGBS Whole Genome Bisulfite Sequencing.

Nomenclature

$\|a\|$	Absolute value of number a
$\{\cdot\}$	Set of discrete elements
∞	Infinity
$\mathbb{R}$	Real numbers set
$\mathbb{R}^+$	Positive real numbers set
$\mathbb{Z}$	Integer numbers set
$\mathbb{N}$	Natural numbers set
$[a, b]$	Closed interval, $a \leq x \leq b$
$\lceil \rceil$	Ceiling function, $\lceil x \rceil = min\{n \geq x\}$ where $n \in \mathbb{Z}$
$\mathbb{1}_A$	Indicator function, 1 if the condition A is satisfied else zero
$\mathbf{1}_N$	All-ones vector with N dimension, (e.g., $1_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$)
$\mathbf{a} \circ \mathbf{b}$	Hadamard product, Element-wise multiplication of two vectors $\mathbf{a}$ and $\mathbf{b}$
$\mathcal{U}_{[a,b]}$	Continuous uniform distribution with boundaries a and b
$P(A)$	Probability that event A is to occur
$\hat{x}$	Estimated value of variable x
$A \cup B$	Union of two sets A and B
$A \cap B$	Intersection of two sets A and B

$\bigcup_{i=1}^{n} A_i$	A finite union of sets $A_1, ..., A_n$ (e.g., $\bigcup_{i=1}^{3} A_i = A_1 \cup A_2 \cup A_3$)
$\binom{n}{k}$	Combination that k elements are selected from a set containing n members, $\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}$

Chapter 1

Introduction

1.1 Motivation

Analysing cell-type compositions of complex tissue or cell-mixture samples, also known as *bulk* samples, provides key evidence to characterise genomic features and phenotypes in large cohorts [Prince et al., 2007, Wen et al., 2017]. For example, immune cell composition in the tumour microenvironment is related to tumour prognosis and immunotherapy responses [Stankovic et al., 2019]. Nonmalignent cell populations have been used as a biomarker for clinical analyses and targeted therapy for tumour [Chu et al., 2022].

In cancer biology, cell-type composition analysis enables the examination of *tumour purity*. Tumour purity has long been used as a prognostic indicator for many cancer types including liver cancer, T-cell lymphoma, and nasopharyngeal cancer [Chuanben et al., 2018, Ho et al., 2021, Schmid et al., 1999]. In addition to estimating cancer prognosis, tumour purity is used for predicting immunotherapy responses [Kim et al., 2021, Deng et al., 2021].

Purification of cell types from bulk samples can be achieved via *in vitro* techniques such as cell sorting. However, *in vitro* cell-type purification often suffers from additional sources of variation introduced during the experiments (e.g., expression level changes). Even though recent advancements in sequencing technology have enabled the acquisition of single-cell profiles, it is still costly and generates highly sparse features. Alternatively, many studies have used a computational approach called *cell-type deconvolution*, which estimates cell-type compositions from bulk samples *in silico*.

DNA methylation (DNAm), which refers to the addition of methyl groups to nucleotide bases in DNA, is one of the most extensively studied epigenetic modifications. In mammals, DNAm particularly at CpG sites involves *cell type-specific* signals [Wen et al., 2017, Hui et al., 2018]. Cell type-specific patterns present in DNAm data can be used for cell-type

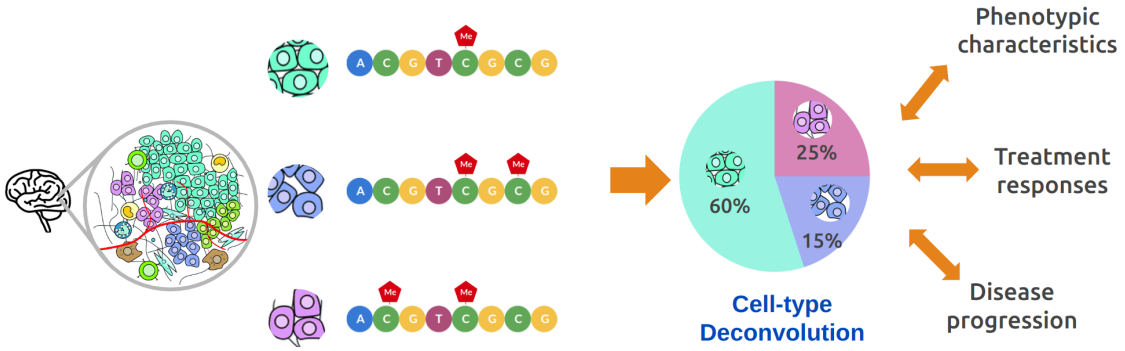


Figure 1.1: Cell-type deconvolution using DNAm data. Based on heterogeneous methylation patterns in different cell types, cell-type deconvolution methods estimate the cell-type composition in given bulk samples. The result can be used for various biological studies and clinical applications (Figure modified from [Leung et al., 2020]).

deconvolution (Figure 1.1). Various machine learning approaches have been suggested to use such cell type-specific patterns in DNAm data for disentangling bulk samples [Decamps et al., 2020, Song and Kuan, 2022].

Sequencing-based profiling of DNAm such as reduced representation bisulfite sequencing (RRBS) or whole genome bisulfite sequencing (WGBS) particularly offers broad genomic coverage and a single-CpG resolution of methylation patterns in a format of *sequencing reads*. The decreasing sequencing costs have more popularised assorted types of bisulfite sequencing data. On the contrary, *array-based* profiling, which is another commonly applied technology for generating DNAm data, covers only a limited number of CpGs.

With these benefits, *read-level methylomes* should provide more information to perform accurate cell-type composition estimation compared to array-based DNAm data. Nevertheless, the majority of cell-type deconvolution methods use array-based DNAm data [Scherer et al., 2020, Houseman et al., 2012]. The existing sequencing-based methods often do not capitalise on the benefits of read-level methylomes. For example, coMethy reshapes the acquired DNAm profiles from sequencing data into a matrix losing the single-CpG resolution [Yin et al., 2019]. On the other hand, DXM estimates cell-type proportions by finding the best fit of distribution to given DNAm data out of a thousand randomly generated distributions [Fong et al., 2021]. This approach limits model optimisation compared to regression-based models by highly relying on the random sampling result.

Consequently, the increasing demand for sequencing data and the limitations of existing methods encourage the development of new methods for accurate cell-type deconvolution using read-level methylomes.

1.2 Contributions

This thesis consists of two parts: (i) A comprehensive benchmarking of existing cell-type deconvolution methods for sequencing-based DNAm data, and (ii) development of a novel sequencing-based cell-type deconvolution method using Transformers for tumour methylomes, named *MethylBERT*.

The benchmarking study systematically evaluates existing sequencing-based cell-type deconvolution methods together with two array-based deconvolution methods. So far, benchmarking studies have been mainly published for deconvolution methods targeting array-based DNAm data. However, considering the advantages and popularity, it is necessary to analyse the current state of deconvolution methods for sequencing-based DNAm data taking its specific features into account. Our benchmarking study proposes a new perspective to evaluating cell-type deconvolution methods by distinguishing between *informative region selection* and *cell-type composition estimation*, which has not been suggested before. The separate analysis proved the importance of informative region selection results and the algorithmic design, for accurate deconvolution. Furthermore, from the comparison of the methods, one can conclude that many of the existing sequencing-based deconvolution methods do not significantly outperform array-based methods, despite the abundance of detailed information about cell type-specific methylation in sequencing-based data. This underscores the necessity of developing new sequencing-based deconvolution methods. This work has been published in [Jeong et al., 2022].

The outcome of the benchmarking study motivated the development of a new model for sequencing-based DNAm deconvolution. We propose MethylBERT as a new Transformer-based cell-type deconvolution model for tumour read-level methylomes. MethylBERT uses bidirectional encoder representations from Transformers (BERT) for the classification of read-level methylation patterns into cell types. Although Transformers have previously been applied to DNAm-related tasks such as DNAm pattern imputation [De Waele et al., 2022], they have not been used for cell-type deconvolution using DNAm data. MethylBERT also determines the Fisher information and includes a new algorithm for adjustment of the estimated tumour purity. The Fisher information yields the precision of tumour purity estimation. This is essential information for the analysis of tumour data which can involve highly noisy methylation patterns. The adjustment of tumour purity estimation is performed by minimising the skewness of estimated region-wise tumour purity. We prove that the adjustment algorithm improves the accuracy of tumour purity estimation by considering region-wise cell-type distributions which can differ from the global cell-type distribution. This work has been published in [Jeong et al., 2023a].

Our experiments show that MethylBERT significantly improves the performance of tumour deconvolution compared to previous methods and provides a highly sensitive tumour signal detection which can be potentially used for early tumour diagnosis via blood test. These results highlight the broad applicability of MethylBERT to diverse types of

biological samples. The existing methods were developed for either general tumour bulk samples [Houseman et al., 2012] or blood biopsy samples [Li et al., 2018, Li et al., 2021], thus their performance was found to be biased toward the targeted sample type. However, MethylBERT can accurately estimate tumour purity for both high and very low percentages of tumour cells. In addition to the performed evaluation, our in-depth analyses of model training and estimated cell-type posterior probabilities provide valuable inputs for the field of bioinformatics. Even though a variety of studies have shown the successful application of Transformer-based models for genomic data analysis [Ji et al., 2021, Gwak and Rho, 2022], to the best of our knowledge, there has not been an explanation of what Transformer-based models learn from DNA sequences. In this thesis, the efficacy of BERT pre-training using 3-mer DNA sequences and the variation of gained knowledge during the MethylBERT training are thoroughly analysed. The examination results confirm that the BERT pre-training on DNA sequences lets the model recognise important DNA sequence features such as DNA nucleotide pairs and CpG-context, without supervision. The pre-training also leads the MethylBERT model to unbiased fine-tuning towards dominant methylation patterns, and to identify correct CpG-specific methylation patterns. The evaluation of MethylBERT has been published in [Jeong et al., 2023a].

A part of the doctoral studies was dedicated to developing a new single-cell multi-omics integration model using variational autoencoder (VAE), named *scMaui* [Jeong et al., 2023b]. This work is not included in this thesis due to its divergence from the scope. However, scMaui has achieved superior performance in cell-type clustering and molecular profile imputation compared to previous methods, as well as identified hidden cell subpopulations.

1.2.1 Publication

Peer-reviewed publication

- **Yunhee Jeong**, Lisa Barros de Andrade e Sousa, Dominik Thalmeier, Reka Toth, Marlene Ganslmeier, Kersten Breuer, Christoph Plass, Pavlo Lutsik, *Systematic evaluation of cell-type deconvolution pipelines for sequencing-based bulk DNA methylomes*, Briefings in Bioinformatics, Volume 23, Issue 4, July 2022, bbac248, `https://doi.org/10.1093/bib/bbac248`,[Jeong et al., 2022]

Preprint

- **Yunhee Jeong**, Karl Rohr, Pavlo Lutsik, *MethylBERT: A Transformer-based model for read-level DNA methylation pattern identification and tumour deconvolution*, bioRxiv, 2023.10.29.564590; doi: `https://doi.org/10.1101/2023.10.29.564590`, [Jeong et al., 2023a].

- **Yunhee Jeong**, Jonathan Ronen, Wolfgang Kopp, Pavlo Lutsik, Altuna Akalin, *Decoding single-cell multiomics: scMaui - A deep learning framework for uncovering cellular heterogeneity in presence of batch effects and missing data*, bioRxiv 2023.01.18.524506; doi: `https://doi.org/10.1101/2023.01.18.524506`, [Jeong et al., 2023b]

1.2.2 Conference presentations

- **Yunhee Jeong**, Jonathan Ronen, Wolfgang Kopp, Pavlo Lutsik, Altuna Akalin, *scMaui: variational autoencoders combined with adversarial learning reveal cellular heterogeneity from single-cell multiomics data and handle multiple batch effects independently*, Genes 2023: Single-cell multiomics moving forward (Oral Presentation), Barcelona, Spain.

1.3 Thesis outline

This thesis is organised as follows. Chapter 2 introduces the background knowledge about epigenomics and machine learning necessary for understanding this thesis. The epigenomic section focuses on DNAm which is the targeted epigenetic modification in the thesis, while the machine learning section describes machine learning models for sequential data along with examples of applications to sequencing-based methylation data. The last section describes the general concept of cell-type deconvolution and existing methods categorised into two groups according to the requirement of the reference data.

Chapter 3 is devoted to the evaluation of six existing sequencing-based cell-type deconvolution methods. Two array-based methods are included as a comparison group, thus we clarify the separate test pipelines designed for sequencing-based and array-based deconvolution methods, respectively. The two major steps of cell-type deconvolution, *informative region selection* and *cell-type composition estimation* steps, are assessed separately, thereafter we examine the relation between the two steps.

In Chapter 4, *MethylBERT* is introduced as a new Transformer-based cell-type deconvolution model for tumour read-level methylomes. The first section explains the network architecture and the training process of MethylBERT. In the second section, the tumour purity estimation algorithm using the estimated cell-type posterior probability is described. The last section specifies the overall training schemes of the MethylBERT.

In Chapter 5, we present the experimental results of MethylBERT and compare its performance with other existing methods. The experimental results are divided into four sections. The first section evaluates MethylBERT in terms of read-level methylation pattern classification. In the second section, we explore the cell-type deconvolution performance of MethylBERT for tumour bulks. The third section demonstrates the efficacy of pretraining in MethylBERT, and the last section showcases the application of MethylBERT to circulating tumour DNA analysis.

Finally, we summarise the thesis and discuss the outcomes in Chapter 6. The future work section includes further tasks that could be achieved by an improved version of MethylBERT in the future.

Chapter 2

Background

In this chapter, background knowledge about the topics covered by this thesis is introduced. Section 2.1 explains the phenomenon of epigenetic modifications, particularly focusing on *DNA methylation (DNAm)*, and how methylation can be a crucial biomarker for tumour analysis. Also, different technologies for DNAm profiles are demonstrated. In Section 2.2, machine learning approaches for modelling read-level methylation patterns are described. Finally, *cell-type deconvolution* models are explained in detail in Section 2.3. This section describes the concept and the importance of cell-type deconvolution as well as the difference between reference-based and reference-free cell-type deconvolution methods.

2.1 Epigenomics

2.1.1 Epigenetic modifications and DNA methylation

Epigenetic modifications are referred to as DNA or histone[1] modifications affecting gene expression without changing the genetic code [Plass et al., 2013]. Different types of epigenetic modifications such as DNA methylation and histone modifications strongly support our understanding of fundamental biological processes [Han and He, 2016]. For instance, epigenetic modifications are capable of regulating cell development, thus epigenetic deregulation may cause abnormal cell development linked to diseases such as cancer [Pujadas and Feinberg, 2012, Atlasi and Stunnenberg, 2017]. Therefore, many biomedical studies have spotlighted epigenetic modifications as a primary biomarker [Prince et al., 2007, Portela and Esteller, 2010, Waldmann and Schneider, 2013].

DNAm is defined as a biological mechanism in which a methyl group is attached to DNA molecules. In mammalian DNA, methylation dominantly occurs at cytosines (C). Cyto-

[1] Histone is a protein which can support the structure of chromosomes.

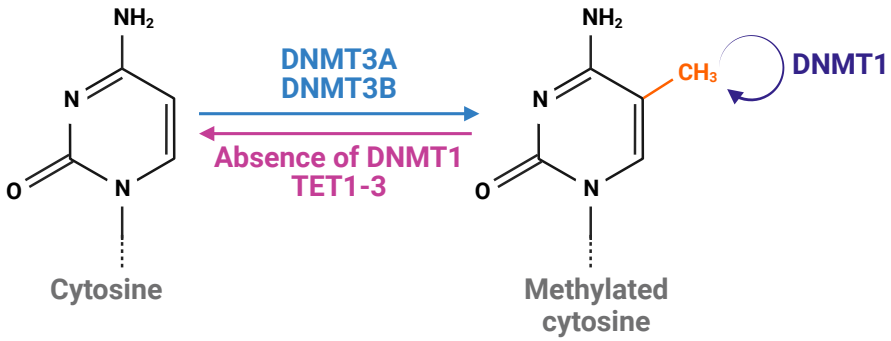


Figure 2.1: Cytosine methylation and demethylation mechanisms. The methylated cytosine has an additional methyl group, CH_3.

sine methylation at *CpG sites* is especially one of the most broadly explored epigenetic modifications due to its relevance to gene expression. CpG sites indicate the regions in DNA sequences where a cytosine is followed by a guanine. Hypomethylation at CpG-dense promoters[2] can activate gene transcription which results in high gene expression, whereas gene silencing can be associated with hypermethylation of promoters. Therefore, cytosine methylation at CpG sites carries cell type-specific patterns that influence the expression of neighbouring genes forming heterogeneous cell types [Hui et al., 2018, Wen et al., 2017].

Cytosine DNA methylation is primarily controlled by six different enzymes: DNA methyltransferase (DNMT) 3A, DNMT3B, DNMT1, and ten-eleven translocation (TET) 1-3 [Ambrosi et al., 2017] (Figure 2.1). TET1-3 refers to the three TET family members, TET1, TET2, and TET3. While DNMT3A and DNMT3B catalyse *de novo* DNA methylation at cytosines, DNMT1 maintains the methylation on the cytosine molecule after the DNA replication during cell division. Demethylation of cytosine can occur via two different mechanisms: active and passive demethylation mechanisms. If DNA replicates without the DNMT1 activity, the methylation of the mother DNA strand is not copied to the new strand of DNA (passive demethylation). Otherwise, TET enzymes catalyse the hydroxylation of the methylated cytosine which leads to DNA demethylation (active demethylation) [Melamed et al., 2018].

[Waddington, 2014] introduced the concept of epigenetic landscape to explain cellular differentiation[3] during cell development, without genetic alterations. The change of epigenetic landscapes is also shown in DNAm patterns. Cell type-specific DNAm profiles are acquired via the methylation change during embryonic development, which is also known as DNAm reprogramming. Starting from a zygote[4], cell differentiation results in different cell types by changing epigenetic landscapes including DNAm patterns. The

[2]Promoter is a genomic region where proteins bind to start transcription of a gene.
[3]Cellular differentiation means a cell-type transition occurring in a cell.
[4]Zygote is a fertilized cell created by sperm and ovum.

unique state of DNAm contributes to programming cell type-specific gene expressions which determines the cell identity eventually [Goldberg et al., 2007, Basu and Tiwari, 2021]. Therefore, DNAm has been used for investigating cell type-specificity explaining the biological development and malignant disease progression [Greenberg and Bourc'his, 2019, Zhou et al., 2017].

Individual cells can show highly heterogeneous properties or behaviours based on many factors, such as cell type or tissue type. This phenomenon is known as *cellular heterogeneity*. The measurement of cellular heterogeneity can be done in various ways including epigenomic studies using DNA methylomes [Goldman et al., 2019]. Explaining cellular heterogeneity based on DNAm data has provided valuable evidence extending our perspectives in biological and clinical studies. For instance, [Liu et al., 2013] claimed that DNAm can pave the way for reducing genetic risk in rheumatoid arthritis (RA) based on cellular heterogeneity analysis of RA patient data.

DNAm has two main strengths as a promising biomarker: stability and inheritability. Numerous previous studies have certified that DNAm keeps its status in both different stages of the cell cycle and different conditions of sample storage [Gosselt et al., 2021, Vandiver et al., 2015] which makes analyses more rigorous. DNAm is also steadily inherited over multiple cell divisions, and some of the inherited DNA methylation has shown its association with cancer susceptibility [Reid and Fridley, 2020, Joo et al., 2018].

In conclusion, the analysis of heterogeneous DNA methylation patterns contributes to understanding both phenotypic and genotypic variability [Sheffield et al., 2017, Li et al., 2016]. Previous studies have explored epidemiology, disease, as well as early mammalian development based on cell type-specificity shown in DNAm [Felix and Cecil, 2019, Greenberg and Bourc'his, 2019]. Furthermore, revealing cell type-specific effects from bulk samples is fundamental evidence to disclose cellular mechanisms associated with diseases and to determine therapeutic targets [Rahmani et al., 2019]. Specifically, a broad range of diseases are known to be detectable with DNAm alteration including cardiovascular diseases, diabetes, neurological disorders and cancer [Kulis and Esteller, 2010, Bansal and Pinney, 2017, Cuadrat et al., 2021, Rasmi et al., 2023]. Hence, varying strategies for disease therapies and prevention have been suggested based on DNAm [Yang et al., 2010, Cheng et al., 2019].

2.1.2 Aberrant DNA methylation and cancer

In cancer, which brings abnormal cell growth, aberrant methylation patterns associated with dysregulated gene expression are often found. Hypermethylation at CpG sites in promoter regions can silence tumour suppressor genes, whereas genome-wide hypomethylation is known to increase with cancer progression [Ehrlich, 2002, Kulis and Esteller, 2010, Ehrlich, 2009]. For example, hypermethylation of *P16* gene promoter region has been repetitively reported in breast, oral, gastric and colorectal cancer patients [Abbaszadegan

et al., 2008, Hall et al., 2008, Veganzones-de Castro et al., 2012, Wang et al., 2012]. This hypermethylation can silence the *P16* gene involved in cell cycle regulation, causing cancerous cell development.

For successful cancer diagnosis and therapy, understanding tumour heterogeneity is key due to the vast variety of genotypic and phenotypic characteristics within a tumour and across patients, which are referred to as intra- and inter-tumour heterogeneity, respectively [Guo et al., 2019, Sollier et al., 2023]. Although many failures in cancer therapies are attributed to tumour heterogeneity giving rise to different cell populations in tumours, it is often not properly addressed in many therapeutic strategies. Therefore, in-depth knowledge of cellular heterogeneity is crucial to explain different tumour samples and establish precise clinical strategies [Waas and Kislinger, 2020].

In the field of cancer research, DNAm is particularly considered a valuable resource for examining tumour heterogeneity due to its stability as explained in Section 2.1.1. DNAm data holds major promise for diagnostics and clinical applications based on tumour heterogeneity analysis.

Previous studies have found heterogeneous methylation patterns across tumours and established tumour classification based on DNAm data [Sill et al., 2020]. For example, [Capper et al., 2018] developed a molecular classification method for the central nervous system (CNS) tumours based on DNAm profiles. This classification method has contributed to the standardised diagnostics in CNS tumours. Classification of juvenile myelomonocytic leukaemia (JMML) patients based on DNAm profiles was suggested for designing stratification of clinical trials [Schönung et al., 2021].

DNAm-based biomarkers enhance the diagnosis and prognosis of cancers. For instance, hypermethylation at the promoter region of *SHOX2* gene is used for lung cancer diagnosis [Schmidt et al., 2010], whereas hypomethylation of *TBRG4* has been identified as a biomarker for colorectal cancer metastasis [Jang et al., 2020].

Clinical applications targeting DNAm have been broadly proposed over multiple cancer types [Yang et al., 2020a, Lietz et al., 2022]. Decitabine, one of the chemotherapy drugs, that causes global hypomethylation shows efficacy for acute myeloid leukaemia (AML) patients [Klco et al., 2013]. Moreover, clinical trials which control specific oncogenes[5] via DNAm have been developed and applied to both haematological malignancies and solid tumours [Cheng et al., 2019, Blagitko-Dorfs et al., 2019, Azad et al., 2013].

Recently, early non-invasive diagnosis of tumours has been demonstrated using circulating tumour DNA (ctDNA) analysis. Cell-free DNA (cfDNA) refers to DNA fragments in body fluids and ctDNA is a type of cfDNA that originates from tumour cells in the stage of

[5]An oncogene is a gene whose mutation can cause cancer.

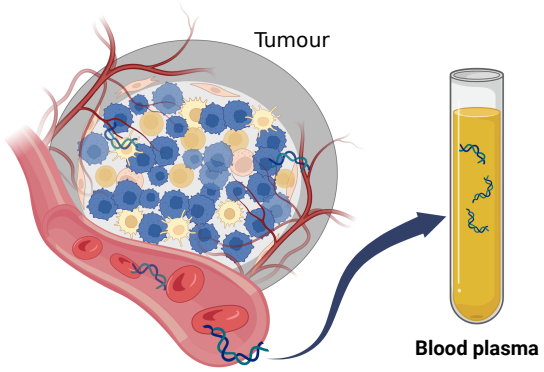


Figure 2.2: Circulating tumour DNA in blood plasma.

apoptosis[6] or circulating tumour cells (Figure 2.2). Blood plasma is the most commonly collected sample type utilised for the detection of ctDNA and the results can be used for monitoring, early diagnosis and treatment responses of cancer [Yan et al., 2021, Duffy and Crown, 2022, Yang et al., 2022]. DNAm profiling is broadly conducted to detect ctDNA signals in various cancer types [De Mattos-Arruda et al., 2013, Pantel and Alix-Panabières, 2017, Salvianti et al., 2021] because it occurs at the very early stage of cancer and provides a clearer distinction between tumour and normal tissues compared to other biomarkers, such as DNA mutation [Fiala and Diamandis, 2018].

2.1.3 Profiling and analysis of DNA methylation

DNA methylation can be profiled using a range of different technologies. Bisulfite treatment changes only unmethylated cytosines (C) into uracil (U), while does not affect methylated cytosines (Figure 2.3A). When PCR amplification is applied to the bisulfite-converted DNA, uraciles are converted to thymine (T), whereas cytosines stay intact. Based on the discriminative conversions of cytosines, methylation patterns can be detected at each CpG using multiple read-outs, including microarrays such as Illumina 450K/EPIC arrays or sequencing protocols like whole genome bisulfite sequencing (WGBS) (Figure 2.3B). In array-based profiling, the methylation intensity is normalised into a range between zero and one. The resulting measure of DNAm is known as methylation *beta-value*. The two different types of DNAm data have pros and cons (Table 2.2).

[6] Apoptosis is the programmed destruction of a cell [Hengartner, 2000].

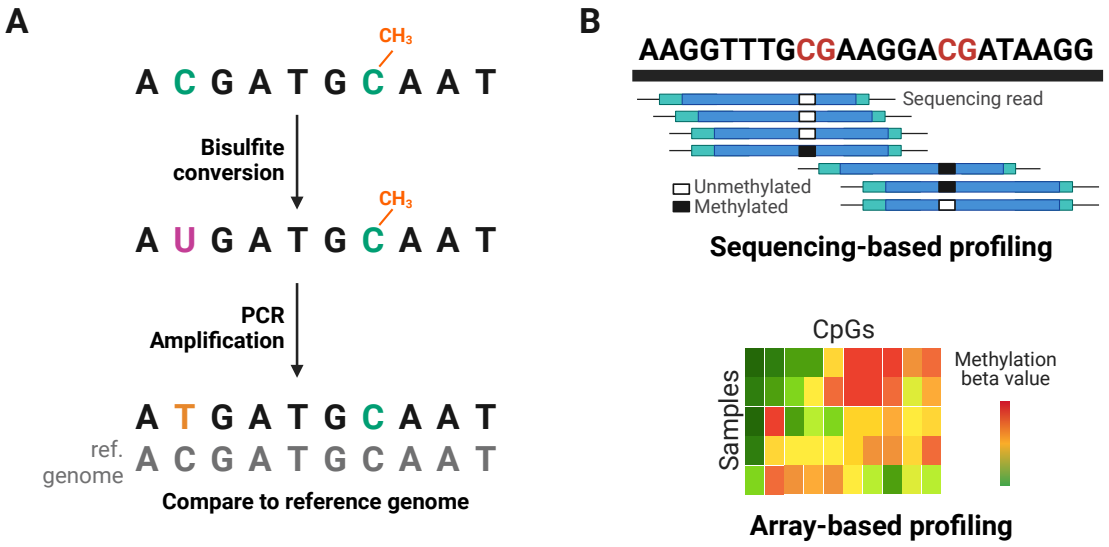


Figure 2.3: DNA methylation profiling. **(A)** The process of bisulfite sequencing data generation. **(B)** Two types of DNA methylation profiling.

Table 2.2: Comparison of bisulfite-treated DNAm profiling methods

	Array-based profiling	Sequencing-based profiling
Description	Beta-values in a CpG-by-sample array normalised from the measured methylation intensities	Sequencing reads including binary methylation patterns
Technologies	• Human Methylation 450K • EPIC array	• Illumina short read sequencer (WGBS, RRBS, scBS-Seq etc.) • Long-read sequencing
Pros	• Small file size • Well-summarised information	• Higher coverage of CpGs • Detection of rare cell type • Genomic context is provided
Cons	• Limited available CpGs • Less detailed information	• Large file size • High cost

In *array-based profiling*, the data is usually stored as a CpG-by-sample matrix comprised of beta-values. 450K arrays contain beta-values measured at ca. 450,000 sites mostly representing genic regions, whereas EPIC arrays capture ca. 850,000 CpGs that mostly lie in regions of biological importance such as enhancers[7] [Shu et al., 2020]. Both methods are broadly used for methylation profiling in large population-level cohorts due to their low cost and straightforward analysis. Array-based profiling data is considered more suitable for modelling or analysis using linear algebraic algorithms such as non-negative least squares or matrix factorization. Nevertheless, the number of CpGs captured by array-based profiling is below 5% of the entire CpGs in the human genome (ca. 28 million), so the obtainable information is largely limited. Detection of rare cell types is another difficulty faced by array-based profiling since the average methylation level can mask the methylation patterns from the minority cell type. Even if the proportion is very low, some cell types can be a crucial indicator for disease prognosis [Orkin and Zon, 2008]. ctDNA analysis, as explained in the previous section, also requires the identification of a very low ratio of tumour-derived cells.

Sequencing-based profiling is an attractive alternative allowing it to cover a much higher number of CpGs. Reduced representation bisulfite sequencing (RRBS) can cover up to 10% of CpGs, whereas WGBS ideally captures all CpGs in the genome. The high CpG coverage of sequencing-based profiling compensates for the limitation of array-based profiling and enables in-depth analysis of methylomes. Such data contains sequencing reads, including both a binary sequence of methylation patterns at cytosines and inferred DNA base pairs. Sequencing-based data provides *read-level methylation* patterns as well as genomic sequence which enables the joint analysis of genomics and epigenomic modifications. Moreover, sequencing-based profiling better preserves the signals from rare cell types through the single molecule-level methylation patterns on sequencing reads. Even though sequencing data remains more costly than array data, the cost of bisulfite sequencing data generation has dropped without losing genomic coverage, making the data more popular and accessible.

With the advance of sequencing technologies, more variations of methylation profiling are available nowadays. Single-cell bisulfite-sequencing (scBS-Seq) is able to profile methylomes of individual cells up to ca. 48% of CpG sites, and it has broadened our perspectives in cellular heterogeneity [Smallwood et al., 2014]. Long-read sequencing is also available for methylation profiling via nanopore or single-molecule real-time sequencing and provides a longer context of methylation changes. Compared to short-read sequencing technologies whose read length is usually 100 to 150 bps, long-read sequencing can generate reads up to 2 Mbps.

[7] Enhancer is a genomic region where proteins bind to enhance transcription of a gene.

Bisulfite sequencing data processing

After acquiring sequencing reads from biological samples, data processing steps must be followed to make a standard quality and analysable sequencing-based DNAm profiling. This processing encompasses alignment to a reference genome, and different sorts of quality control. All BS-seq data present in this thesis were processed via the processing steps below:

1. **Trimming**

   The raw sequencing reads generated by a bisulfite sequencer may involve technical errors, thus every nucleotide base within a sequencing read comes with a quality value, called base quality. The standard base quality is defined as a Phred quality score (Q) of base calling error probabilities (P) [Ewing and Green, 1998]:

   $$Q = -log_{10}P. \tag{2.1}$$

   During trimming, low-quality ends of sequencing reads are trimmed off with a given threshold. Adapter sequences within the sequencing reads are also removed during trimming. The adapter sequence is a short sequence ligated to DNA fragments and enables the binding of DNA fragments to a flow cell. Once the fragments are attached to the flow cell, they can be sequenced base-by-base. We used *Cutadapt* [Martin, 2011] and *Trim Garlore* (for the other samples, `https://github.com/FelixKrueger/TrimGalore`) for the trimming.

2. **Alignment**

   Alignment is a step to find the genomic location of sequencing reads on a reference genome. It requires a reference genome acquired from the same organism that the samples originated from. Because of the bisulfite conversion, BS-seq data needs to be aligned with a bisulfite-aware aligner. We used *bismark* [Krueger and Andrews, 2011] for aligning all BS-seq data sets in this thesis. As reference genome, *mm10* and *hg19* were used for mouse and human samples, respectively.

3. **Sorting**

   Sorting the aligned reads according to the genomic position not only supports fast analyses but also reduces the file size. For the data used in this thesis, we ran *samtools sort* to sort the reads [Danecek et al., 2021].

4. **Duplicate removal**

   PCR amplification conducted during BS-seq data generation results in many copies of the same DNA fragments. Therefore, the duplicates need to be removed to avoid the potential errors caused by the amplification. Here, we used *Picard* for the du-

plicate removal [Pic, 2019].

5. **Filtering** (optional)

   The mapping quality indicates the probability that each read is mapped to an incorrect location. It is also calculated using the Phred score (Equation 2.1). Only for non-cancer samples, we filtered reads whose mapping quality is lower than 30 by using *samtools view* [Danecek et al., 2021]. However, for the tumour samples, filtering based on the mapping quality score could discard the mutant signals that occurred in tumour cells.

2.1.4 Differentially methylated region

Once DNAm data is generated from different samples, one of the gold-standard analysis methods is to look at differentially methylated loci (DMLs) or regions (DMRs). DMLs and DMRs are referred to as genomic loci and regions, respectively, that exhibit methylation level differences between the groups of interest (Figure 2.4). A specific region where many DMLs are closely located to each other is generally considered a DMR. Not only are such regions considered to be associated with gene regulations, but their methylation patterns also predominantly reveal cellular or tumour heterogeneity [Ferreira et al., 2019]. Especially in bulk samples, which are comprised of multiple cell types in different states, there could be many confounding factors that conceal cell type-specific methylation patterns. DMRs identified using scBS-seq data or cell line-generated BS-seq data give an informative subset of genomic regions for downstream analysis minimising the impact of confounding factors.

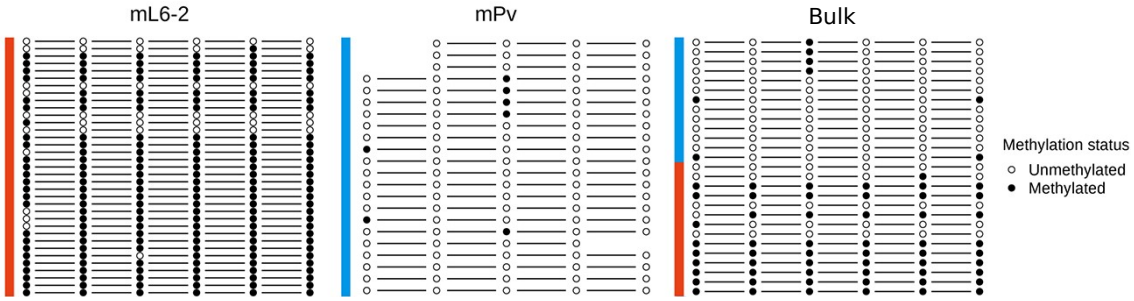


Figure 2.4: An example of DMR. This region (chr1:75244319-75244379) shows differentially methylated patterns between two mouse neuronal cell types, mL6-2 and mPv. Each row represents sequencing reads covering methylated (black) and unmethylated (white) CpGs. Missing methylation patterns at CpGs not covered by the reads remain empty in respective rows. mL6-2 dominantly has fully methylated reads, whereas reads in mPv mostly have fully unmethylated patterns. Because these two cell types (indicated by red and blue) have different methylation patterns, we observe complex methylation patterns in the bulk sample.

Throughout this thesis, the *DSS* package was used for DMR calling [Park and Wu, 2016]. When the DMR calling is conducted for D samples containing N CpG methylation patterns, DSS assumes that the number of methylated reads Y_{id} at CpG i in sample d follows a beta-binomial distribution:

$$Y_{id} \sim BetaBinom(m_{id}, \pi_{id}, \gamma_i) \tag{2.2}$$

where π_{id} and γ_i are the mean and dispersion of the beta distribution. These two parameters are explained in more details in Section 2.2.1. m_{id} is the total number of reads at CpG i in sample d. Let $\mathbf{X}$ be a $D \times P$ matrix containing P experimental designs (or phenotypes) for the samples. $\mathbf{X}$ can have both continuous and discrete values. Here, DSS assumes an *arcsin* link function of π_{id} as follow:

$$arcsin(2\pi_{id} - 1) = \mathbf{x}_d \boldsymbol{\beta}_i \tag{2.3}$$

where $\mathbf{x}_d$ is the dth row of $\mathbf{X}$. $\boldsymbol{\beta}_i$ is given as a vector of coefficients for P designs. The model coefficient vector $\boldsymbol{\beta}_i$ is estimated via a beta-binomial generalised linear model, and then used for hypothesis test to determine DMLs. CpGs whose test statistic for the standard Wald test with the null hypothesis $H_0 : \mathbf{C}^T \boldsymbol{\beta}_i = 0$ is lower than a given threshold are considered DMLs. $\mathbf{C}$ is a P-dimensional binary vector where 1 indicates the experimental design to test differentially methylated patterns.

A DMR is identified as a region of adjacent DMLs. The *areaStat* score is the sum of the test statistics of DMLs in a DMR, while the *diff.Methy* score is calculated by subtracting the mean methylation value of the tested experimental design from the value of the others.

2.2 Machine learning models for read-level methylomes

2.2.1 Beta-binomial/Bernoulli distribution model

Using a beta-binomial or Bernoulli distribution is considered to be a suitable approach for modelling read-level methylomes. This is because DNA methylation has two discrete events, methylated and unmethylated CpGs, which can be interpreted as 'success' and 'failure' in these distributions.

The beta-binomial distribution is a discrete probability distribution that models Bernoulli trials whose success probability is drawn from a beta distribution. For the Bernoulli trials, which are referred to as random trials resulting in two possible outcomes, a fixed value for the number of trials n must be given. Therefore, three parameters need to be determined for the beta-binomial model: $\alpha \in \mathbb{R}^+$, $\beta \in \mathbb{R}^+$ and $n \in \mathbb{N}$. α and β are used to characterise a beta function $B(\alpha, \beta)$ in the beta-binomial distribution $BetaBinom(n, \alpha, \beta)$. The mathematical expression of the beta-binomial model is:

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt, \tag{2.4}$$

$$BetaBinom(x; n, \alpha, \beta) = \binom{n}{x} \frac{B(x+\alpha, n-x+\beta)}{B(\alpha, \beta)}. \tag{2.5}$$

[Dolzhenko and Smith, 2014] modelled read-level methylomes from WGBS data using a beta-binomial model and utilised the model for DMR identification. Given the methylation level, the number of methylated reads m divided by the number of total reads n at a CpG site, they assume the probability mass function of m follows the beta-binomial distribution:

$$P(m|n, \alpha, \beta) = BetaBinom(m; n, \alpha, \beta). \tag{2.6}$$

The beta function in the beta-binomial can also be written with other parameters π and γ reparametrised by α and β:

$$\pi \coloneqq \frac{\alpha}{\alpha+\beta} \text{ and } \gamma \coloneqq \frac{1}{\alpha+\beta+1} \tag{2.7}$$

where π and γ are interpreted as the mean and dispersion of the beta distribution. The expectation of m is then calculated as $\mathbb{E}(m) = n\pi$, thus π can be interpreted as the average methylation level over the given samples $\frac{\mathbb{E}(m)}{n}$. The parameters are estimated via beta-binomial regression [Crowder, 1978] using WGBS data from a set of samples.

The Bernoulli distribution models a binary random variable with fixed probabilities p and $1-p$. Therefore the probability mass function $Bernoulli(x; p)$ is given as:

$$Bernoulli(x; p) = p^x(1-p)^{1-x} \text{ for } x \in \{0, 1\}. \tag{2.8}$$

This function is also a type of the binomial distribution whose trial number is 1.

[Kapourani and Sanguinetti, 2019] used Bayesian inference to cluster single-cell methylation sequencing data and found methylation variability between cells. In their model, the CpG-wise methylation has one binary value observed in an individual cell. Therefore, they modelled the methylation at CpG site i in a genomic region g with the Bernoulli distribution. The CpG methylation at the site i in the region g is denoted as $y_{g,i}$:

$$y_{g,i} \sim Bernoulli(p_{g,i}) \tag{2.9}$$

where $p_{g,i}$ is an unobserved true methylation level. They assume the cumulative distribution function (CDF) of the standard Gaussian distribution $\Phi(\cdot)$ which ensures the output value to be in the range between zero and one, for a probit regression model estimating true methylation $p_{g,i}$ from neighbouring CpG methylation patterns x_g within the same

region g. The model is written as:

$$\Phi(w_g^\intercal x_g) = p_{g,i} \tag{2.10}$$

where w_g is the unknown parameter vector for the regression.

Both beta-binomial and Bernoulli distributions have limitations in handling methylation patterns on sequencing reads because these do not consider the relationship between neighbouring CpG methylation patterns. [Dolzhenko and Smith, 2014] modelled the methylation levels of individual CpGs independently but could not capture the relations between neighbouring CpGs using these distributions. [Kapourani and Sanguinetti, 2019] introduced another probit regression to address the neighbouring CpGs. Hence, modelling read-level methylomes based on the beta-binomial or Bernoulli distribution either requires another method to explain the neighbouring CpG methylation patterns or disregards the adjacency.

2.2.2 Hidden Markov Model (HMM)

A hidden Markov model (HMM) explains unobserved hidden states following a Markov chain and observations assumed to be dependent on the hidden states [Baum and Petrie, 1966, Baum and Eagon, 1967, Baum, 1968, Baum et al., 1970, Baum et al., 1972]. A Markov chain is a sequential model in which the probability of every step only depends on the previous steps, not on the future steps (Figure 2.5A). When the hidden states $\mathbf{h} = \{h_1, ..., h_T\}$ over a discrete time series $t \in \{1, ..., T\}$ follows the first order Markov chain (each step only relies on one previous step), the sequence of hidden states can be expressed as:

$$P(\mathbf{h}) = P(h_1) \prod_{t=1}^{T-1} P(h_{t+1}|h_t). \tag{2.11}$$

For a sequence of observations $\mathbf{O} = \{O_1, ..., O_T\}$ that emitted from the hidden states h_t, an HMM (Figure 2.5B) is modelled as follows:

$$P(\mathbf{O}, \mathbf{h}) = P(h_1) \prod_{t=1}^{T-1} P(h_{t+1}|h_t) \prod_{t=1}^{T} P(O_t|h_t). \tag{2.12}$$

Assume that the observation O_t and the hidden state h_t are categorical random variables with N and M possible categories each. Then, the transition probability matrix $\mathbf{A} \in \mathbb{R}^{M^2}$ whose elements $A_{ij} = P(h_{t+1} = j|h_t = i)$ and the emission probability matrix $\mathbf{E} \in \mathbb{R}^{M \times N}$ whose elements $E_{ij} = P(O_t = j|h_t = i)$ need to be calculated for every possible combination to create an HMM. In the standard case, an HMM assumes that the transition and the emission probability mass functions (PMFs) are the same at all time steps. For the first time step, $P(h_1)$ needs to be given separately as a vector $\boldsymbol{\pi} \in \mathbb{R}^M$.

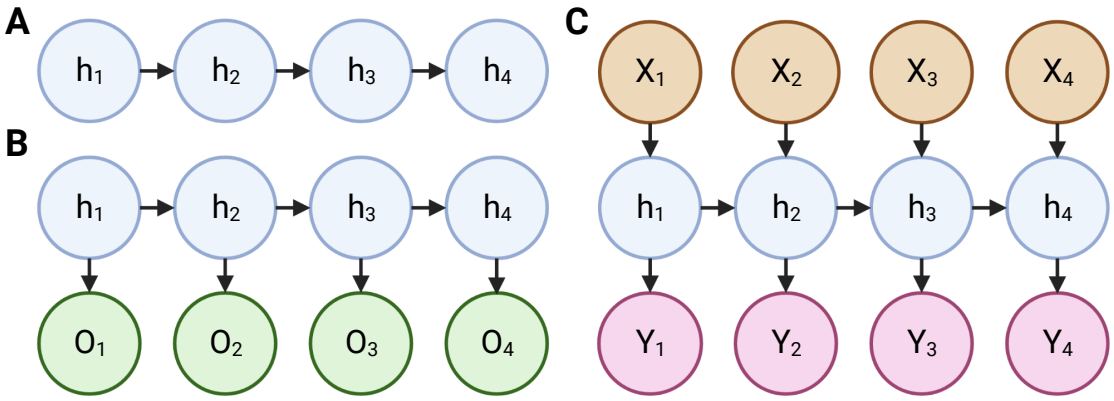


Figure 2.5: Dependency graph of sequential learning models. **(A)** Markov chain with states h_i. **(B)** Hidden Markov Model with hidden states h_i and observation O_i. **(C)** Unfolded dependency of recurrent neural network for input X_i, hidden states h_i and output Y_i.

Therefore, training an HMM is to estimate the parameter set $(\mathbf{A}, \mathbf{E}, \boldsymbol{\pi})$.

There have been several works using the Markov chain and HMM to model the dynamics of DNA [Sontag et al., 2006, Kyriakopoulos et al., 2019]. These works categorise DNAm into multiple classes describing respective stages of the DNAm establishment process. For example, considering that DNA is double-stranded, [Sontag et al., 2006] separates DNAm from different strands of DNA, thus making four states: both strands unmethylated, hemimethylated on one strand, hemimethylated on the other strand, and both strands methylated. This model was also applied to calculate the efficiency of DNA methyltransferases (DNMT), which is a type of enzyme contributing to the transfer of a methyl group to DNA molecules, in *de novo* methylation and DNAm maintenance (see Section 2.1.1). [Kyriakopoulos et al., 2019] employed an HMM to model the cytosine methylation and demethylation processes over time caused by chemical modifications such as 5-hydroxymethyl cytosine or 5-formyl cytosine.

HMM-Fisher adopts an HMM to detect unique methylation features from read-level DNAm data taking sequencing errors into account [Sun and Yu, 2016]. A transition matrix of HMM is inferred to model true CpG-wise methylation patterns correcting errors in each sample. The model yields three types of hidden state categories whose transition probability follows a multinomial distribution: N (non-methylated), P (partially methylated) and F (fully methylated). Then, it uses a truncated normal distribution for the emission probability that an observed methylation is emitted from each category. Finally, Fisher's exact test is applied to determine CpG sites showing unique methylation features by comparing the count of estimated hidden states at each site between the two groups.

In comparison with the beta-binomial and Bernoulli distributions (see Section 2.2.1), HMM

better captures the relationships between adjacent CpG methylation patterns. Nonetheless, HMM makes a strong assumption that the hidden states must form a Markov chain, which may not apply to true CpG methylation patterns. [Yoon, 2009] also pointed out that the hidden states in HMM, solely dependent on the previous steps, do not perfectly fit to explain biological sequences where molecular interactions occur in both directions.

2.2.3 Recurrent Neural Networks (RNNs)

Advancements in deep neural networks have had a great impact on performance increases in various fields [Min et al., 2017, Goh et al., 2017, Kamilaris and Prenafeta-Boldú, 2018]. Deep neural networks can exploit the entangled information behind a large data set through their high model complexity and nonlinearity. Recurrent Neural Networks (RNNs) are a class of neural networks primarily developed for sequential data. RNNs capture sequential information through recurrent training of nodes in the network over all elements in the input sequence. In standard RNNs, each hidden node $\mathbf{h}_t$ receives the current input point values $\mathbf{x}_t$ and the hidden node values in the previous state $\mathbf{h}_{t-1}$ [Lipton et al., 2015] (Figure 2.5C). RNNs for an input $\mathbf{x}_t$ and an output $\mathbf{y}_t$ at the time step t are modelled as:

$$\begin{aligned} \mathbf{h}_t &= a_h(\mathbf{W}_h\mathbf{x}_t + \mathbf{U}_h\mathbf{h}_{t-1} + \mathbf{b}_h) \\ \mathbf{y}_t &= a_y(\mathbf{W}_y\mathbf{h}_t + \mathbf{b}_y) \end{aligned} \tag{2.13}$$

and the model is trained to learn the model parameters and biases, $\mathbf{W}_h, \mathbf{W}_y, \mathbf{U}_h, \mathbf{b}_h$ and $\mathbf{b}_y$. a_h and a_y are the activation functions for calculating values of the hidden nodes at the current time point t and the output, respectively.

The recurrent structure of RNNs gives the networks the capacity to handle sequential information, but also brings the problem of exploding/vanishing gradients [Pascanu et al., 2013]. In particular, the vanishing gradient problem makes it difficult to learn long-distance dependencies [Lipton et al., 2015]. Long Short-Term Memory (LSTM) was proposed to overcome this problem by having a memory cell that memorises the essential information from previous time steps. [Cho et al., 2014] improved the LSTM in terms of model complexity by controlling the update of hidden layers using a reset gate and an update gate. The developed model is named gated recurrent unit (GRU).

RNNs have been used in a broad range of fields such as natural language processing or speech recognition [Graves et al., 2013, Yin et al., 2017]. Likewise, many studies have suggested using RNNs for modelling sequencing-based DNAm data for varying tasks [Angermueller et al., 2017, Maruyama et al., 2022].

DeepCpG is a combined model of bidirectional GRU and convolutional neural networks (CNNs) to infer missing CpG methylation patterns based on neighbouring DNAm patterns and DNA sequences [Angermueller et al., 2017]. The bidirectional GRU networks were designed to encode binary methylation patterns of neighbouring CpGs collected from multiple cells, and the CNNs encode the reference DNA sequence. In this way, the model

can infer a CpG-wise methylation pattern taking both cell-to-cell methylation variability and the sequence motifs associated with DNAm. This model has become particularly useful with the growth of scBS-seq data which often involves a sparsity problem.

[Maruyama et al., 2022] also used GRU to determine the impact of DNA sequences on DNAm inheritance in CpG islands (CGIs), which is a genomic region where a high frequency of CpG sites are found. The method processes DNA sequences at CGIs into k-mer sequences and creates embedding vectors from the k-mer sequences which are given as input to the GRU model. Then, the sigmoid-transformed output is compared with ground-truth methylation patterns interpreting the output as the probability that CGI is unmethylated given the DNA sequence.

[Angermueller et al., 2017] and [Maruyama et al., 2022] discovered that RNNs make it possible to associate sequential methylation patterns with DNA sequences. The nonlinearity and the high model complexity of neural networks seem to catch the relations between methylation patterns and DNA sequences. This also applies to Transformer-based models explained in the following section and the newly developed cell-type deconvolution model for tumour samples developed in this thesis (see Chapter 4).

2.2.4 Transformers

[Vaswani et al., 2017] proposed a new sequential learning model called *Transformer* overcoming the major limitations of the recurrent models described in Section 2.2.3. Due to the dependence between the current hidden state and the previous hidden state, recurrent models cannot exploit the benefit of parallel computing and suffer from a long training time. On the other hand, the Transformer uses *attention mechanisms* enabling bidirectional learning in sequences without a time-dependent hidden layer or state. Specifically, the Transformer applies 'self-attention' to create representations of input and output sequences.

Attention mechanism

An attention function is designed to map a query vector and a pair of key and value vectors to an output vector. In the Transformer network, 'Scaled Dot-Product Attention' is used. Scaled dot-product attention for the query, key and value vectors $\mathbf{Q} \in \mathbb{R}^{d_Q}, \mathbf{K} \in \mathbb{R}^{d_K}, \mathbf{V} \in \mathbb{R}^{d_V}$ is calculated as:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}. \tag{2.14}$$

$\frac{1}{\sqrt{d_k}}$ is applied as a scaling factor to prevent a large magnitude of $Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ value due to a large size of the vector. The large magnitude can eventually cause the vanishing gradient problem.

In order to find different projections of query, key and value vectors, the Transformer employs multiple attention functions, the so-called 'Multi-head attention'. Multi-head attention concatenates H attention matrices computed from the weighted query, key and value vectors with $\mathbf{W}^Q, \mathbf{W}^K$, and $\mathbf{W}^V$, respectively:

$$\begin{aligned} \text{Multi-head attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concatenate(\mathbf{A}_1, ..., \mathbf{A}_H)\mathbf{W}^O, \\ \text{where } \mathbf{A}_i = Attention_i(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V). \end{aligned} \tag{2.15}$$

$\mathbf{W}^O$ is a projection matrix of the concatenated attention matrices yielding the final output of the multi-head attention. The three input vectors of the attention function can be three different sequences, but it can be also one sequence. The latter is called 'self-attention'. Self-attention relates different positions in the same sequence and finds positions considered to be important by other positions.

The Transformer architecture consists of two parts: an encoder and a decoder. The encoder applies multi-head self-attention to the input sequence and extracts encoded features, whereas the decoder involves both a multi-head self-attention layer for the output sequence and a multi-head attention layer applied to the encoded input and output.

BERT

Taking the superior performance of Transformers, there have been diverse strategies proposed to pre-train Transformer-based models for NLP tasks [Sarzynska-Wawer et al., 2021, Radford et al., 2018]. Bidirectional Encoder Representations from Transformers (BERT) is one of the Transformer-based models, which is mainly comprised of the Transformer encoder part [Devlin et al., 2018]. It is pre-trained in an unsupervised manner using masked language model (MLM) and next sentence prediction (NSP). Once pre-training is complete, the model can be fine-tuned for a specific task.

Prior to the pre-training, the input consisting of two sentences needs to be processed into three embeddings containing different information (Figure 2.6). Token embeddings are created referring to a lookup table where individual words are matched with a unique number. Sentence embeddings indicate which sentence each token belongs to, and position embeddings are given to store the order of tokens within the input.

After the embedding step, the three embedded vectors are summed up into one final embedding vector. For MLM, some of the input tokens are randomly masked and the BERT model predicts a token at the masked positions. The first token of any input is [CLS] to inform the beginning of the input. The output at [CLS] position is used for NSP to predict whether sentences A and B are connected to each other. The pre-training makes the BERT model understand the context of the language so that model fine-tuning can be performed more easily for specific tasks, such as language translation.

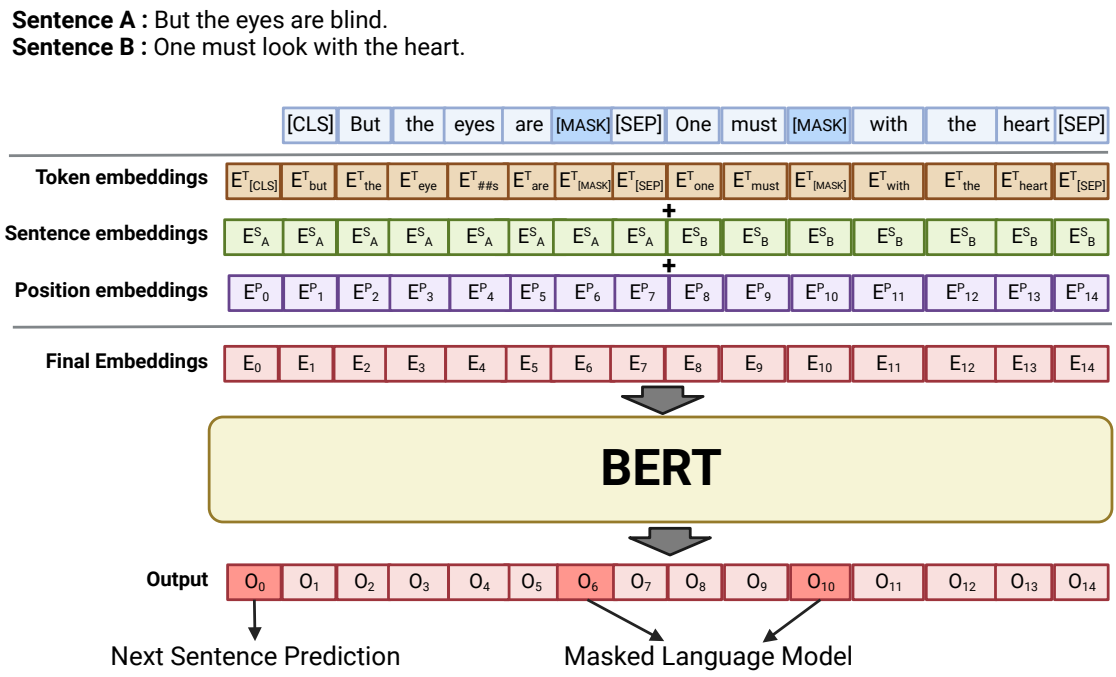


Figure 2.6: BERT pre-training. Two input sentences are converted into three embeddings. BERT uses next sentence prediction and masked language model for pre-training.

BERT is able to associate input tokens within a sentence using the multi-head self-attention mechanism via the encoders from Transformers. Using BERT for sequential data has two major benefits: global context learning and non-directional training. For training with sequential data, CNNs calculate features using a convolution operation in a fixed-size window, usually much shorter than the input sequence. On the other hand, a long training time caused by the dependence on the previous hidden state often becomes a problem in RNNs. BERT addresses these issues by calculating attention matrices over whole sequences, not assuming a certain direction of flow in the sequence, and not focusing on a specific part of the sequence.

Transformers and BERT have achieved groundbreaking performance in many tasks involving sequential data such as NLP, computer vision, speech recognition, and so on [Subakan et al., 2021, Lin et al., 2022, Wang et al., 2019, Wagner and Rohr, 2022]. The application of Transformer-based models in bioinformatics has also solved unanswered questions in biology by discovering complicated hidden relationships in molecule-level systems [Jumper et al., 2021, Ji et al., 2021, Gwak and Rho, 2022]. There have also been studies suggesting the use of Transformers or BERT for sequencing-based methylation data, [Yu et al., 2021, De Waele et al., 2022], which are described below.

[Yu et al., 2021] applied BERT to identify different types of DNAm in multiple species

using DNA sequences surrounding methylated nucleotides. In their model, methylation patterns are not encoded using the BERT model but instead, long-distance dependencies on DNA sequences are found to determine the DNAm type. As input data, DNA sequences are encoded into token and position embeddings. Sentence embeddings are not used.

[De Waele et al., 2022] developed a model called CpG Transformer which infers missing methylation patterns based on Transformers. CpG Transformer and DeepCpG introduced in Section 2.2.3 aim at the same task: methylation imputation, and both combine a sequential model with CNNs whose input are embeddings of DNA sequences. CpG Transformer merges the embedded DNA sequences and a cell-by-CpGs DNAm matrix, then calculates attention matrices over cells and CpG sites in separate steps.

Both models outperformed previous works, however, they are still not applicable to address cell type-specific patterns in read-level methylomes. While the model suggested by [Yu et al., 2021] relies solely on DNA sequences as input, [De Waele et al., 2022] did not take read-level information into account.

2.3 Cell-type deconvolution models using DNA methylomes

2.3.1 Cell-type deconvolution

Despite the recent rise in the popularity of single-cell data, bulk profiling of DNA methylation is still commonly applied to samples with multiple populations and cell types. Therefore, knowing cell type-specific signals and the composition of cell types is crucial to analysing the impact of individual cell types in bulks. However, dissecting cell type-specific signals from bulk DNA methylomes is challenging because of confounding factors. DNA methylation is also related to gender, age, and environmental influences [Zhang et al., 2011, Boks et al., 2009]. Additionally, *in vitro* experiments conducted while generating DNAm profiles may introduce technical artefacts perplexing cell type-specific signals.

Cell-type deconvolution, also known as cellular deconvolution, is an *in silico* method to estimate the cell-type proportions within bulk samples. The estimated cell-type proportions are strongly related to phenotypic characteristics and can be a biomarker in clinical applications [Prince et al., 2007, Wen et al., 2017]. A robust and accurate cell-type deconvolution method enables cellular-level analyses on a large cohort of data. Single-cell sequencing would not be a realistic option for such a large cohort due to high cost. Moreover, cell-type deconvolution methods are also not affected by technical biases like the error rate caused by high dropout events and high sparsity. The high dropout and high sparsity problems have been regarded as a challenge in single-cell technologies. Cell sorting and cell enrichment are available as *in vitro* technologies for dissecting a cell mixture. However, these technologies also induce additional costs and have the risk of introducing technical variations in the data. Another benefit of cell-type deconvolution is the application for analysing stored bulk data. In many biological studies, the stored (old) data is still

worthwhile for a longitudinal study. In addition, such data can lead to new conclusions when a novel analysis method is applied. Since the biological samples for the old data are often not available, single-cell technologies cannot be used to generate new data for the same samples. In this case, cell-type deconvolution is significantly advantageous.

Due to the cellular heterogeneity observed in multiple layers of the omic profile, cell-type deconvolution methods can target different types of data. For example, CIBERSORT and MuSiC infer cell-type proportions based on gene expression profiling using RNA sequencing data [Wang et al., 2019, Newman et al., 2019]. It is not very common, but there have been a few studies to perform cell-type deconvolution using proteomic data [Wang et al., 2022, Petralia et al., 2021]. The following sections exclusively focus on DNAm-based cell-type deconvolution methods, which are the main topic of this thesis. Reference-based and reference-free cell-type deconvolution methods are described in separate sections.

2.3.2 Reference-based methods

Reference-based cell-type deconvolution methods require reference data to train a supervised model for cell-type composition estimation. Different types of information can be provided as reference data. DNAm profiles of pure cell types are commonly acquired from either scBS-seq data or bulk methylomes from cell lines made of specific cell or tissue types. When it comes to cell line-originated bulk data, cell lines can be generated using one specific cell type. As an alternative, another set of bulk samples with ground-truth compositions can be given as reference data, particularly for regression-based models, but it is usually not easy to obtain.

For array-based data, regression is one of the most common approaches for cell-type deconvolution [Houseman et al., 2012, Arneson et al., 2020, Chakravarthy et al., 2018]. The majority of regression-based models estimating the composition of C cell types within N bulk samples based on methylation pattern of M CpG sites assume:

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \mathbf{e}, \tag{2.16}$$

when bulk DNAm profiles $\mathbf{Y} \in [0,1]^{N \times M}$ and pure cell-type DNAm profiles $\mathbf{X} \in [0,1]^{C \times M}$ are given. Then the approaches infer the cell-type composition matrix $\mathbf{W} \in [0,1]^{N \times C}$. $\mathbf{e}$ is an error vector of the model, which can be estimated either from the data or sampled from a designated distribution.

Reference-based cell-type deconvolution models for read-level DNAm data vary more in terms of methodological design compared to the array-based methods. MethylFlow is based on a minimum-cost flow problem on the network of reads where edges connect regions with equal methylation patterns on overlapping CpGs [Dorri et al., 2016]. ClubCpG employs a regression-based model to estimate cell-type compositions using extracted principal components from the bulk DNAm profiles. The method requires training data, which

consists of bulk samples containing the same cell types to build a regression model [Scott et al., 2020]. BED is a probabilistic model to infer cell-type compositions via the maximum a posteriori method [Barrett et al., 2017].

Recently, reference-based cell-type deconvolution using read-level methylomes has been used for ctDNA analysis. CancerDetector assumes a beta-binomial distribution on the CpG-wise methylation values, and it separately calculates a probability of tumour and healthy cell types given a read [Li et al., 2018]. DISMIR is an RNN-based model yielding a 'd-score,' which is a sigmoid-calculated output [Li et al., 2021]. Both methods employ maximum likelihood estimation (MLE) to infer the tumour purity using the calculated probability or d-score.

CancerDetector and DISMIR, however, have limitations as reference-based cell-type deconvolution methods for read-level methylomes. CancerDetector averages the methylation values on a read, disregarding single molecule level methylation patterns. On the other hand, DISMIR restricts the maximum read length to 66 bps, which is shorter than half of the common read length in BS-seq data (150 bps). Consequently, both models cannot fully utilise the benefits of read-level methylation patterns due to either the averaging of values or the short read length.

2.3.3 Reference-free methods

Reference-free cell-type deconvolution methods are unsupervised learning methods for estimating the cell-type compositions. The absence of reference data involves difficulties not only in the inference of cell type-specific methylation patterns, but also in choosing genomic regions to explore, since detecting DMRs also requires DNAm profiles of purified cell types. Therefore, many reference-free methods recommend providing DNAm profiles on promoter or enhancer regions, or CpGs showing the highest variability of methylation patterns over multiple samples.

Non-negative matrix factorisation (NMF) is a representative algorithm used for reference-free cell-type deconvolution methods developed for array-based DNAm data. NMF factorises a matrix $\mathbf{V}$ into two different matrices $\mathbf{W}$ and $\mathbf{H}$ with the assumption that all matrices do not have a negative value:

$$\mathbf{V} \approx \mathbf{WH}. \tag{2.17}$$

In NMF for DNAm-based cell-type deconvolution, $\mathbf{V} \in [0,1]^{M \times N}$ is a methylation beta-value matrix at M CpGs including N bulk samples. $\mathbf{W} \in [0,1]^{M \times C}$ and $\mathbf{H} \in [0,1]^{C \times N}$ refer to a DNAm profile of the estimated number of cell types C and the composition of C cell types in N bulk samples, respectively. [Titus et al., 2017b] applied the NMF algorithm to estimate tumour purity in breast cancer patients. [Lutsik et al., 2017] improved the NMF by introducing a regularisation term forcing the inferred subpopulation-wise DNAm

profiles (equal to cell type-wise DNAm profiles in this thesis) to the range of $[0, 1]$.

Reference-free cell-type deconvolution models for sequencing-based data commonly summarise read-level methylomes per cell type in multiple genomic regions and estimate the cell-type composition based on the summarised information. The method Prism corrects outlier read-level methylation patterns using HMM and fits a beta-binomial distribution to region-wise fully methylated and unmethylated reads distributions in order to estimate the cell-type composition [Lee et al., 2019]. [Yin et al., 2019] also used an NMF-based algorithm to estimate the cell-type composition from a summarised methylation pattern matrix from read-level methylation patterns in multiple samples.

Reference-free methods are preferred when suitable reference data are not available, but they are bound to the cell types associated with phenotypic variations [Houseman et al., 2016]. Additionally, the estimated proportions have to be annotated by users and this usually requires another source of data to compare the inferred methylation profiles of cell types.

Chapter 3

Systematic Evaluation of Cell-Type Deconvolution Methods for DNA methylomes

3.1 Introduction

In the past decade, cell-type deconvolution methods have been extensively applied to estimate the cell-type composition within *array-based* DNA methylation (DNAm) data. Following the popularity, there have been several benchmarking studies about array-based cell-type deconvolution for DNAm data. [Titus et al., 2017a] tested array-based cell-type deconvolution methods using blood-derived DNAm data and analysed the relations between the estimated cell-type proportions and risk factors of multiple cancer types. [Decamps et al., 2020] particularly evaluated reference-free cell-type deconvolution methods using array-based DNAm data. They elaborated on the importance of feature selection and confounding factor removal for accurate cell-type proportion estimation. [Song and Kuan, 2022] more recently performed a benchmarking study of array-based cell-type deconvolution methods specifically for blood samples.

Next-generation sequencing[1] technologies allow for the profiling of read-level methylomes and provide deeper insights into cellular heterogeneity within DNAm. With the advent of next-generation sequencing, several cell-type deconvolution methods tailored for *sequencing-based* DNAm data have been published [Zheng et al., 2014, Barrett et al., 2017, Lee et al., 2019, Yin et al., 2019, Scott et al., 2020, Fong et al., 2021]. However, the authors tested their methods with different data sets and criteria. In addition, method-specific data preprocessing pipelines and different formats of output make it hard to assess

[1] Next-generation sequencing is a sequencing technology which can process millions of DNA fragments in parallel [Behjati and Tarpey, 2013].

the published methods.

In this work, we systematically evaluate current cell-type deconvolution methods for sequencing-based DNAm data and perform an unbiased and thorough comparison. We establish a benchmarking strategy to assess the unique properties of sequencing-based cell-type deconvolution methods. In addition, we compare the performance of previous methods to each other, as well as to array-based methods, in varying aspects. Our benchmarking study specifically addresses three questions as follows:

- What are the commonly shared characteristics in the algorithmic design of sequencing-based cell-type deconvolution methods?
- What is the current state-of-the-art of sequencing-based cell-type deconvolution methods and how do they perform compared to array-based methods?
- What are the limitations of current methods?

Compared to the previous benchmarking studies [Titus et al., 2017a, Decamps et al., 2020, Song and Kuan, 2022], not only does our work provide a systematic evaluation targeting sequencing-based cell-type deconvolution methods, but the evaluation also encompasses a broad range of biological scenarios including normal tissues, tumour tissues, and circulating tumour DNA (ctDNA) samples. Furthermore, we conduct a thorough comparison between sequencing-based and array-based cell-type deconvolution methods which has not been attempted by other studies. Therefore, our benchmarking study provides important information for the field of bioinformatics by covering unexamined aspects of sequencing-based cell-type deconvolution. This work was published in [Jeong et al., 2022].

Below, in Section 3.2, an overview of our benchmarking study is given. In Sections 3.3 and 3.4, the data sets used for the evaluation and algorithmic details of the evaluated methods are described, respectively. Section 3.5 explains the performance evaluation metrics. In the following sections, we present the comparison results for the informative region selection step (Section 3.6) and the cell-type composition estimation step (Section 3.7). Finally, Section 3.8 studies the influential factors affecting cell-type deconvolution performance.

3.2 Overview of benchmarking procedures

Figure 3.1 shows the overall scheme of our benchmarking study. We compare six sequencing-based cell-type deconvolution methods: Bayesian epiallele detection (BED) [Barrett et al., 2017], ClubCpG [Scott et al., 2020], csmFinder + coMethy [Yin et al., 2019], DXM [Fong et al., 2021], MethylPurify [Zheng et al., 2014] and PRISM [Lee et al., 2019]. Two array-based methods, Houseman's method [Houseman et al., 2012] and MeDeCom [Lutsik et al., 2017], are added to the evaluation as a comparison group.

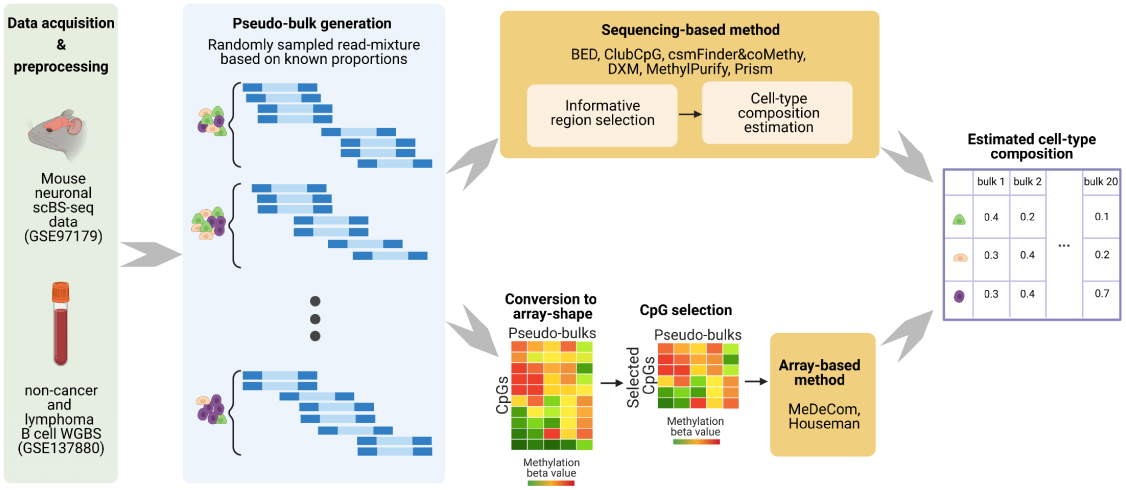


Figure 3.1: Overall scheme of cell-type deconvolution benchmarking. For the evaluation, synthetic mixtures of reads were generated. The read-level methylomes were acquired from the mouse neuronal single-cell bisulfite sequencing (scBS-seq) data set or tumour whole genome bisulfite sequencing (WGBS) data set. Synthetic mixtures are referred to as *pseudo-bulks*. Sequencing-based cell-type deconvolution methods can use the read-level methylomes without further processing and select informative regions for cell-type specific signals (upper part). Then, the cell-type composition is estimated within these regions. On the other hand, we converted the read-level methylomes into a matrix shape for the array-based cell-type deconvolution methods (bottom part). Pre-designated CpGs are used for the conversion, and the array-based methods infer cell-type distributions from the matrix-shaped data.

The evaluated sequencing-based methods consist of two steps: *informative region selection* and *cell-type composition estimation*. Approximately 80% of CpGs are methylated in mammal genomes [Acharjee et al., 2023], and many of these regions have identical methylation patterns regardless of cell types. Only a minor number of genomic regions are detected as differentially methylated regions for human normal cell types [Loyfer et al., 2023], which means that it is excessive to investigate the entire genome for cellular heterogeneity analysis. Furthermore, confounding factors can perturb methylation patterns in some genomic regions. Therefore, selecting informative regions in the entire genome can avoid infeasible computational complexity and remove irrelevant signals, perplexing the modelling of cell-type composition estimation.

During the informative region selection step, each method defines a 'region' as a group of CpGs closely located to each other or covering a certain genomic region. Afterwards, the methods filter out regions that do not satisfy certain criteria. The remaining regions are considered 'informative regions' to deconvolute the mixture of cell type-specific signals shown in bulk samples. Three criteria to select informative regions for each benchmarked method are specified in Table 3.1. We followed the hyperparameter values mentioned in

the original publications, but modified some values when a method performed poorly for the data used in our benchmarking study. Based on these criteria, only regions overlapping with a sufficient number of CpGs within a certain length of region and covered by enough number of reads are selected.

Houseman's method and MeDeCom do not have a specific informative region selection step, since array-based methods require array-shaped data containing methylation patterns only at designated CpGs. Thus, in our benchmarking study, we pre-selected CpGs to convert the sequencing-based DNAm data to an array-shaped format for these methods. For Houseman's method, which is reference-based, cell-type differentially methylated regions (DMRs) calculated by comparing reference cell types were chosen for the conversion of sequencing-based data. For MeDeCom, which is designed to be a reference-free method, 20,000 CpGs with the highest variance of methylation patterns were selected. *methrix* package [Mayakonda et al., 2020] was used to convert the data and the details are described in Section 3.3.2.

Once CpG-wise methylation patterns are collected from the selected informative or predefined regions, the cell-type deconvolution methods estimate cell-type compositions based on the methylation patterns in these regions. There are three criteria to characterise the cell-type composition estimation step: reference methylomes prerequisite, number of identifiable subpopulations, and estimation scope (Table 3.1).

Reference methylomes prerequisite. For supervised (reference-based) cell-type deconvolution models, reference methylomes are required as data to train a cell-type composition estimation model. Among the benchmarked methods, BED, Houseman's method and ClubCpG are classified as reference-based methods. In the BED and Houseman's method algorithms, pure cell-type methylation profiles are required as reference data. On the other hand, ClubCpG trains a regression model on a training set of bulk samples given with ground-truth cell-type composition.

Number of identifiable subpopulations. Cell-type deconvolution methods specifically targeting tumour samples presume that the given cell mixtures have binary components of healthy and tumour stroma[2]. Such methods are often named '*tumour purity estimation*' methods and output the proportion of two subpopulations. On the other hand, standard cell-type deconvolution methods do not make an assumption about the number of subpopulations to detect. In our benchmarking study, BED and MethylPurify are considered tumour purity estimation methods.

Estimation scope. A common approach to cell-type composition estimation is to calculate final estimates from the summarised methylation patterns across the informative regions. Depending on the scope of estimation, we categorise the methods into 'local' and 'global' methods. Local methods calculate statistics in each informative region and

[2]Stroma means the area of an organ or a tissue which gives structural support.

Table 3.1: Methodological comparison of the evaluated cell-type deconvolution methods. Informative region selection is used to decide which CpGs present cell type-specific methylation patterns based on three criteria: minimum number of CpGs in the region, region size and read coverage. We altered some of the parameters from the original papers, because the default parameters given by the authors did not perform reasonable cell-type deconvolution for our data sets. The cell-type composition estimation step also has three common criteria to categorise the cell-type deconvolution methods: requirements of reference data, number of components that the method can detect, and estimation scope.

Method	Main data type	Informative region selection			Cell-type composition estimation		
		# CpG	Region Size (bp)	Coverage	Class	# Components	Estimation Scope
BED	RRBS	>4	*NA****	>20	ref-based	2	local
ClubCpG	WGBS	>4*	100	>20*	ref-based	2 or more	global
csmFinder + coMethy	WGBS	>4	*NA****	>10	ref-free	2 or more	global
DXM	Any kinds of BS-seq	Promoter and CpG island regions		>4	ref-free	2 or more	global
MethylPurify	WGBS	>10	300/200**	>10	ref-free	2	local
Prism	RRBS	>4	*NA****	>20	ref-free	2 or more	local
Houseman	Methylation microarrays	CpGs overlapping with ctDMRs			ref-based	2 or more	global
MeDeCom	Methylation microarrays	CpGs showing high methylation variance over covering reads			ref-free	2 or more	global

* The original setup in [Scott et al., 2020] uses 2 for the minimum number of CpG and 10 for the minimum read coverage.
** The original setup in [Zheng et al., 2014] uses 300 bp for the region size. In tumour-normal pseudo-bulk analysis, we changed the region size parameter value to 200 bp.
*** BED, csmFinder and Prism do not require specific region sizes.

establish a distribution of the region-wise statistics. Then, the final estimation of cell-type proportions is decided by choosing the peak or the best value of the distribution. BED, MethylPurify and PRISM belong to this category. The other methods, however, globally estimate the cell-type proportions directly from all selected regions.

3.3 Data set

3.3.1 Pseudo-bulk generation

For the evaluation of cell-type deconvolution methods, ground-truth cell-type proportions are compulsory for test bulk samples. However, as mentioned in Section 2.3, obtaining very accurate cell-type proportions from biological samples *in vitro* is not easy, and it potentially introduces technical variations that affect cell-type proportions. Therefore, we generated a virtual cell mixture *in silico*, called 'pseudo-bulk', by mixing randomly sampled sequencing reads from single cell or cell line[3] samples. We regard the ratio of reads sampled from each cell type as ground-truth cell-type proportions that cell-type deconvolution methods are supposed to infer.

In order to simulate various biological scenarios, three pseudo-bulk data sets were created (Table 3.2). 20 pseudo-bulk samples were generated to create two different data sets containing two and five different mouse neuronal cell types. These data sets represent normal (non-tumour) tissue samples including a different number of subpopulations. Pure mouse neuronal cell-type samples were collected from a publicly accessible single-nucleus mouse brain DNA methylation data set [Luo et al., 2017]. The data set was downloaded from Gene Expression Omnibus (GEO) with the accession number GSE97179. Since the downloaded data set contains 16 different mouse neuronal cell types, five cell types that have adequate single-cell samples and constitute clear clusters on the t-SNE visualisation made by the authors were chosen. Both inhibitory (mPv) and excitatory (mDL-2, mL2-3, mL5-1 and mL6-2) neuron types were included in the five cell types.

The nature of tumour methylomes certainly differs from normal cell-type methylomes. For example, partially methylated domain (PMD) and increasing allele-specific methylation (ASM) in tumours form a greater number of anomalies in tumour tissue methylation patterns [Do et al., 2020, Nishiyama and Nakanishi, 2021]. For this reason, another pseudo-bulk data set is prepared for evaluating the cell-type deconvolution methods in terms of tumour-derived abnormal DNAm patterns. The tumour-normal pseudo-bulk data set also has 20 bulk samples generated by merging reads from diffuse large B-cell lymphoma (DLBCL) and normal healthy B-cell samples downloaded with the GEO accession number GSE137880 [Do et al., 2020].

The ground-truth cell-type proportions in pseudo-bulk data sets were sampled from Dirich-

[3]Cell line refers to as cultures of a specific population of cells [Farrell, 2011].

Table 3.2: Specification of generated pseudo-bulk data sets. Three different sets of pseudo-bulks were created for our benchmarking study. Mouse neuronal scBS-seq data set was used as a resource for two and five cell-type mouse neuronal pseudo-bulk samples, and one more data set was created from DLBCL and normal B cell WGBS data. Each pseudo-bulk data set contains 20 samples.

	Resource (GEO accession)	#bulks	Cell types	Biological sample	Sequencing protocol
2 cell-type mouse neuronal pseudo-bulk	GSE97179	20	mL6-2, mPv	Mouse neuron	scBS-seq (Illumina HiSeq 4000)
5 cell-type mouse neuronal pseudo-bulk	GSE97179	20	mL6-2, mPv, mDL-2, mL2-3, mL5-1	Mouse neuron	scBS-seq (Illumina HiSeq 4000)
Tumour-normal pseudo-bulk	GSE137880	20	Normal B-cell, DLBCL	B-cell	WGBS (Illumina NovaSeq 6000)

let distributions using the *generateExample* function implemented in R package MeDeCom (`https://rdrr.io/github/lutsik/MeDeCom/src/R/utilities.R`). Most parameters were set up as default values, but we changed *proportion.var.factor* value to 10 and the number of genomic features value to 1 million. The simulated cell-type proportions are listed in Table 3.3 and the pipeline can be found on `https://github.com/CompEpigen/SeqDeconv_Pipeline/blob/main/Pseudo_bulk_generation_pipeline.md`.

3.3.2 Conversion of BS-seq data into an array data

In order to compare the sequencing-based methods with the array-based methods, we reshaped the read-level methylomes into an array shape. The conversion was done by *methrix* R package [Mayakonda et al., 2020]. Methrix is a toolkit for WGBS data analysis which provides various functions including genomic strand collapse, quality control, and read coverage filtering. It creates a matrix of methylation beta-values from a tab-delimited file (e.g., bedGraph file) containing read coverage, and the number of methylated reads at CpGs. As input data of methrix, a bedGraph file from each pseudo-bulk sample was generated using *MethylDackel*[4].

[4] `https://github.com/dpryan79/MethylDackel`

Table 3.3: Cell-type proportions for pseudo-bulk samples used in Chapter 3

	2 cell-type mouse neuron		**5 cell-type mouse neuron**					**Tumour-normal**	
Samples	mL6-2	mPv	mDL-2	mL2-3	mL5-1	mL6-2	mPv	B cell non-cancer	B cell Lymphoma
Bulk 1	0.731	0.269	0.350	0.148	0.242	0.091	0.169	0.151	0.849
Bulk 2	0.445	0.555	0.149	0.096	0.451	0.106	0.197	0.945	0.055
Bulk 3	0.810	0.190	0.444	0.166	0.078	0.293	0.020	0.152	0.848
Bulk 4	0.658	0.342	0.376	0.176	0.245	0.091	0.112	0.190	0.810
Bulk 5	0.338	0.662	0.049	0.459	0.062	0.172	0.258	0.680	0.320
Bulk 6	0.352	0.648	0.381	0.100	0.037	0.407	0.074	0.801	0.199
Bulk 7	0.617	0.383	0.176	0.035	0.242	0.317	0.230	0.790	0.210
Bulk 8	0.591	0.409	0.141	0.075	0.124	0.249	0.410	0.496	0.504
Bulk 9	0.558	0.442	0.160	0.199	0.166	0.151	0.324	0.624	0.376
Bulk 10	0.444	0.556	0.280	0.130	0.456	0.034	0.100	0.657	0.343
Bulk 11	0.330	0.669	0.259	0.155	0.264	0.290	0.032	0.552	0.448
Bulk 12	0.377	0.623	0.081	0.446	0.140	0.166	0.166	0.963	0.037
Bulk 13	0.662	0.338	0.248	0.142	0.141	0.118	0.351	0.955	0.045
Bulk 14	0.461	0.539	0.119	0.340	0.118	0.245	0.177	0.315	0.685
Bulk 15	0.835	0.165	0.150	0.062	0.278	0.295	0.215	0.983	0.017
Bulk 16	0.694	0.306	0.201	0.451	0.088	0.241	0.019	0.804	0.196
Bulk 17	0.550	0.450	0.266	0.210	0.210	0.169	0.146	0.170	0.830
Bulk 18	0.624	0.376	0.350	0.027	0.340	0.225	0.058	0.738	0.262
Bulk 19	0.671	0.329	0.271	0.159	0.334	0.198	0.039	0.673	0.327
Bulk 20	0.539	0.461	0.053	0.390	0.055	0.112	0.390	0.926	0.074

3.4 Algorithmic details of evaluated methods

In this section, we describe the details of six sequencing-based and two array-based cell-type deconvolution algorithms benchmarked in this thesis. Although the study aims to evaluate sequencing-based methods, also array-based cell-type deconvolution algorithms were included as a comparison group.

3.4.1 Sequencing-based methods

Bayesian epiallele detection (BED)

[Barrett et al., 2017] suggested using Bayesian modelling for the distribution of all possible methylation patterns in specified genomic regions. The method is named BED and the authors define 'epialleles' as all possible cases of methylation patterns in a region. Over the modelled distribution, BED estimates tumour purity. The BED algorithm involves two inferences performed in every region: epialleles and the epiallele class of individual reads. These inferences are done via maximum a posteriori (MAP). Tumour purity is finally estimated by taking the peak of estimated region-wise tumour purity. In a region i that has Q epiallele classes $\{q_1, ..., q_Q\}$, the tumour purity is approximately estimated as follows:

$$\frac{1}{2}\sum_{j=1}^{Q}|P(q_j|b,i) - P(q_j|n,i)|, \tag{3.1}$$

where b and n are a bulk sample to deconvolute and normal tissue reference data, respectively. Preprocessing of the input data and cell-type composition estimation were performed by a pipeline uploaded on *bed-beta* Github page[5].

ClubCpG

Density-based spatial clustering of applications with noise (DBSCAN) [Ester et al., 1996] is used for clustering reads in ClubCpG algorithm [Scott et al., 2020]. Only reads that fully cover the selected regions are used to perform the clustering. We modified the parameter values for the informative region selection step from the default setup to more suitable values for the data set used in this study (Table 3.1). As a cell-type composition estimation procedure, the authors suggested regression-based estimation on principal components (PCs), which we followed to obtain the final estimates. As a training set for the regression model, 100 pseudo-bulks were additionally generated. Following the paper, we fitted a multivariate linear regression model to 20 PCs extracted from the training set to predict the cell-type composition and the fitted model inferred cell-type compositions within given pseudo-bulk samples.

[5] `https://github.com/james-e-barrett/bed-beta`

csmFinder + coMethy

[Yin et al., 2019] proposed a pipeline consisting of two computational models, csmFinder and coMethy. csmFinder identifies putative cell-type specific methylated (pCSM) loci which are conceptually equal to informative regions in our benchmarking study. The design of the csmFinder algorithm assumes bipolar patterns, meaning that fully methylated and unmethylated reads exclusively exist in informative regions thus filtering out other regions. coMethy dissects a samples-by-pCSM loci methylation beta-value matrix into cell-type proportions of the samples and cell type-specific methylomes in each cell type. The coMethy algorithm is developed based on non-negative matrix factorisation (NMF). Due to the special input requirement of csmFinder, the input bulk samples were converted into methylation call files by *bismark_methylation_extractor* [Krueger and Andrews, 2011]. To determine pCSM loci, csmFinder also requires reference genomes, so *hg19* human genome and *mm10* mouse genome were provided accordingly. After detecting pCSM loci in each bulk sample, we only retained loci where all bulk samples present methylation signals.

DXM

DXM employs L1-norm minimisation to infer the number of subpopulations, the proportions, and the methylome profiles of each subpopulation [Fong et al., 2021]. The concept of subpopulation can be regarded the same as the concept of cell type in our benchmarking study. It computes the L1-norm between 10,000 randomly generated distributions and the distribution of methylation beta-values of a given bulk in every informative region. Then, the generated distribution yielding the lowest L1-norm value is chosen to represent the cell-type distribution in the bulk. Although it is a sequencing-based method, DXM demands users to provide a pre-selected genomic region set rather than finding informative regions. Therefore, we have provided promoter and CpG island regions where the read coverage is higher than 4. The UCSC genome annotation data set was used to obtain CpG island regions in mm10 and hg19 genomes (`https://hgdownload.cse.ucsc.edu/goldenpath/mm10/database/` and `https://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/`). Promoter regions were collected from TxDb objects of UCSC annotation [Team and Maintainer, 2020, Carlson and Maintainer, 2015].

MethylPurify

MethylPurify [Zheng et al., 2014] is a tumour purity estimation method based on the Expectation-Maximisation (EM) algorithm. The EM algorithm infers both methylation levels in tumour and normal cell types (denoted as m_1 and m_2 in the paper), and the tumour purity (denoted as α_1). Technically, the original paper does not explicitly mention that cell-type 1 is the tumour cell type, but here we assume so for convenience. Also, we introduce another random variable c_i for cell type, where c_1 is tumour and c_2 is normal. The authors established a likelihood function involving m_1, m_2, α_1 and, α_2 assuming that

N sequencing reads $R = \{r_1, ..., r_N\}$ are independently sampled:

$$L(m_1, m_2, \alpha_1, \alpha_2) = \prod_{r_i \in R} \sum_{j=1}^{2} \alpha_j P_{m_j}(r_i | c_j). \tag{3.2}$$

We note that α_j is the estimated proportion of cell type c_j, thus $\alpha_1 + \alpha_2 = 1$. Therefore, the parameter estimation problem has 3 degrees of freedom. Parameters are estimated via the EM algorithm by maximising the likelihood value. Regarding hyperparameters, the mouse neuronal data set and tumour data set had 300 bp and 200 bp as bin size values, but other values were given, as stated in the paper. The original implementation determined informative regions only in CpG islands, but, for our benchmarking study, it was modified to not restrict the range of discovering informative regions due to the insufficient number of informative regions detected in CpG islands.

PRISM

PRISM [Lee et al., 2019] mainly focuses on only fully methylated and fully unmethylated reads to disregard non-dichotomous methylation patterns while estimating tumour purity using the EM algorithm. Prior to the estimation, it employs a hidden Markov model (HMM) to amend possibly erroneous methylation patterns. The HMM model in the PRISM algorithm resembles the principle of DNMT1 enzyme[6]. DNMT1 methylates hemimethylated CpGs[7] Associating the DNMT1 enzyme status with observed methylation patterns, [Lee et al., 2019] designed an HMM whose hidden state has two categories: whether DNMT1 is attached to DNA or not. The observation of the HMM is the methylation pattern. After correcting methylation patterns, PRISM assumes a beta-binomial mixture model for region-wise binary pattern distributions. The parameters of the beta-binomial model, α and β (described in Section 2.2), and the weight of cell types (referred to as subclones in their paper) in the mixture model are estimated via the EM algorithm. The weight values can be regarded as cell-type proportions. Although the authors stated that PRISM can identify multiple cell types within a bulk, we found that PRISM can only detect two cell types for the pseudo-bulk data set used in our benchmarking study.

3.4.2 Array-based methods

Houseman's method

Houseman's method uses regression calibration to predict cell-type distribution within array-based DNAm data [Houseman et al., 2012]. To convert bisulfite sequencing data into an array shape, CpGs that are located in DMRs were chosen following the marker-CpG selection part in the original paper. Then, the conversion from read-level methylomes

[6]DNMT1 maintains the methylation on the cytosine molecule after the DNA replication as explained in Section 2.1

[7]Hemimethylation means CpGs are only methylated at one strand.

to the array shape has been done as described in Section 3.3.2. Although the original implementation selects only 500 CpGs, we increased the number to 1,000 because of the broader CpG coverage of the pseudo-bulk data compared to the real microarray data used in the original paper.

MeDeCom

MeDeCom is an unsupervised array-based cell-type deconvolution method based on NMF [Lutsik et al., 2017]. As mentioned in Section 2.3, MeDeCom includes an additional regularisation term in the optimisation to ensure that the inferred cell type-specific methylation profiles are in the range between zero and one. For the informative region, 20,000 CpG sites showing the highest variance of methylation level across all samples were chosen. The following parameter specifications were used: 500 for regularisation and 30 for the random initialisation number.

3.5 Performance metrics

In this section, we explain the metrics used for measuring informative region selection and cell-type composition estimation performances of benchmarked cell-type deconvolution methods.

3.5.1 Performance metrics for informative region selection

Overlaps with DMRs

Ideally, selected informative regions should involve clearly different methylation patterns between cell types. Once DNAm data from different types of samples are given, calculating differentially methylated regions (DMRs) is a gold-standard analysis. DMRs can be calculated by comparing methylation profiles between given sample types. In the case of cell-type deconvolution, DMRs calculated between different cell types can be regarded as an ideal selection of genomic regions involving cell type-specific methylation patterns. Therefore, we established the hypothesis that more clear cell type-specific signals are obtained from selected informative regions with higher similarity to DMRs. DMRs for each cell type were called as described in Section 2.1.4. As measurements of the similarity, we use the number of overlaps between DMRs and the selected regions to measure how accurately a method can identify DMRs.

For counting overlaps, we used the *findOverlaps* function in the GenomicRanges R package [Lawrence et al., 2013]. The findOverlaps function identifies overlaps between query and reference sets of genomic ranges based on a red-black tree T whose node i contains an individual genomic range in the reference set [Cormen et al., 2022]. The contained genomic range (interval) is written as $i.int$. $i.int.start$ and $i.int.end$ indicate the start and end

positions of the region. Then, the overlap of two regions $i.int$ and $i'.int$ is defined as:

$$i.int.start \leq i'.int.end \text{ AND } i'.int.start \leq i.int.end \tag{3.3}$$

with an assumption that both regions are located on the same chromosome.

For each query genomic range j, the findOverlaps function identifies a node whose region overlaps with j by using the *Interval-Search(T, j)* operation as described in Algorithm 1 [Cormen et al., 2022].

Algorithm 1 The *Interval-Search(T, j)* algorithm

Input: T a red-black tree whose nodes contain regions in the reference set
j Query region
Output: i The node whose region overlaps with query region j

```
1: i = T.root
2: while i ≠ nil AND j does not overlap with i.int do
3:     if i.left ≠  AND i.left.max ≤ j.start then
4:         i = i.left
5:     else
6:         i = i.right
```

$T.root$ indicates the root node of tree T
nil means an empty tree
$i.left$ and $i.right$ mean the left and right subtrees of the node i
$i.left.max$ is the max value in {$i.end$, the largest end within $i.left$, the largest end within $i.right$}

Genomic correlation

'Genomic correlation' was suggested by [Favorov et al., 2012] to determine the difference between two region sets based on a distribution of distances. In our benchmarking study, we used it to measure the similarity between DMRs and selected informative regions.

For calculating genomic correlation, the 'relative distance' needs to be defined to form a distribution of distances between two region sets. Here, we define a genomic region as $[chr, s, e]$, where chr, s, e refer to the chromosome, start and end of the region. When a selected informative region $q_i \coloneqq [chr_{q_i},\ s_{q_i},\ e_{q_i}]$ and a set of D DMRs ordered by genomic position $d_k \coloneqq [chr_{d_k},\ s_{d_k},\ e_{d_k}] \in \{d_1, ..., d_D\}$, the relative distance η_i is defined as:

$$m_{q_i} = \lceil \frac{s_{q_i} + e_{q_i}}{2} \rceil,\ m_{d_k} = \lceil \frac{s_{d_k} + e_{d_k}}{2} \rceil \tag{3.4}$$

$$\bar{k} = \arg\min_k (m_{q_i} - m_{d_k}),\ \text{where } chr_{q_i} = chr_{d_k} \text{ and } m_{d_k} < m_{q_i} \tag{3.5}$$

$$\eta_i = \frac{min(|m_{q_i} - m_{d_{\bar{k}}}|, |m_{q_i} - m_{d_{\bar{k}+1}}|)}{|m_{d_{\bar{k}+1}} - m_{d_{\bar{k}}}|}. \tag{3.6}$$

η_i is a normalised distance between q_i and the closest DMR in the range $[0, \frac{1}{2}]$. Then the genomic correlation is calculated using the empirical cumulative distribution function defined as:

$$ECDF_\eta(x) = \frac{1}{Q} \sum_{\eta_i \in \{\eta_1 \dots \eta_Q\}} \mathbb{1}_{\eta_i < x} \tag{3.7}$$

when Q selected informative regions are given. The genomic correlation is calculated by comparing $ECDF_\eta$ with $ECDF_{ideal}$. $ECDF_{ideal}$ is defined as the ECDF under a null hypothesis. Here, the null hypothesis is that the relative distances have a uniform distribution between 0 and 0.5 when the two region sets are independent. The comparison between two $ECDF$s is done based on the area between two functions:

$$GenomicCorr \coloneqq \frac{\int_0^{\frac{1}{2}} (ECDF_\eta(x) - ECDF_{ideal}(x))dx}{\int_0^{\frac{1}{2}} ECDF_{ideal}(x)dx}. \tag{3.8}$$

The genomic correlation is interval-bounded by [-1, 1]. Two independent region sets yield zero and identical region sets yield one. Although the genomic correlation can be -1 if all selected regions fall into the middle between two DMRs, all selected regions in our benchmarking study had values ≥ 0 (see Section 3.6). Therefore, we consider the magnitude of genomic correlation as the overall proximity measure between informative region selection results and DMRs.

3.5.2 Performance metrics for cell-type composition estimation

Absolute error and percentage absolute error

The major criterion used for the cell-type composition estimation step is the median absolute error (MAE) between the predicted and ground-truth cell-type proportions. The two NMF-based methods benchmarked in this chapter, coMethy and MeDeCom, do not assign cell types automatically to the estimated proportions. Therefore, we assigned cell types to the estimated proportions by comparing the MAE of all possible combinations of cell types.

In the cell-type deconvolution analysis for very low cell-type proportions, the median absolute error may become harder to interpret. Therefore, for this analysis, we used the mean absolute percentage error (MAPE), which is known to be scale-independent [Kim and Kim, 2016]. MAPE calculated over N cell types is defined as:

$$MAPE \coloneqq \frac{1}{N} \sum_{n=1}^{N} |\frac{y_n - \hat{y}_n}{y_n}|, \tag{3.9}$$

where y_n and $\hat{y}_n$ are the ground-truth and estimated cell-type proportions for a cell type n.

Entropy of cell-type distribution

For a random variable, the entropy measures the uncertainty or the amount of carried information of the possible results. It is also commonly used for quantifying the irregularity of proportion or distribution [Humeau-Heurtier, 2018, Inouye et al., 1991]. Here, we determine the irregularity (or the uniformity) of the cell-type distribution within a cell mixture using entropy to test if a biased cell-type distribution makes the cell-type deconvolution intractable. The entropy of the cell-type distribution $H(C)$ in a bulk sample whose N cell-type proportions are given as $C = \{c_1, ..., c_N\}$ is calculated as:

$$H(C) = -\sum_{i=1}^{N} c_i log(c_i). \tag{3.10}$$

The entropy value increases, when cell types are more uniformly distributed.

3.6 Informative region selection result comparison

In this section, we evaluate the informative region selection step of the benchmarked cell-type deconvolution methods. For the overlap between the selected informative regions and DMRs, BED and csmFinder yielded the highest number for the two cell-type mouse neuronal and tumour-normal pseudo-bulks (Figure 3.2A and B). For the five cell-type pseudo-bulks, ClubCpG detected the largest number of regions overlapping with DMRs for all cell types (Figure 3.2C). However, measuring the total number of overlaps cannot be the only metric to compare the informative region selection step, because the methods detected different sizes of informative region sets, and the number of detected regions is highly correlated with the number of overlaps in all pseudo-bulk samples (Figure 3.3). Therefore, even if the selected regions have a lot of overlapping regions with DMRs, there might be many more regions not overlapping with DMRs in the selected region set. This can eventually hinder extracting sufficient cell type-specific methylation patterns for cell-type deconvolution due to the majority of uninformative regions.

To complement the limitations of the number of overlaps for informative region selection evaluation, the genomic correlation was calculated as another score to measure the similarity. In both two cell-type mouse neuronal and tumour-normal pseudo-bulk analyses, a high genomic correlation value is achieved by csmFinder, Prism and MethylPurify (Figure 3.4A and C). On the other hand, BED - which yields the largest number of overlaps for two cell-type mouse neuronal pseudo-bulks (Figure 3.2A) - had the lowest genomic correlation for both cell-type DMR sets. csmFinder achieved the highest genomic correlation for all five cell-type DMRs in the five cell-type pseudo-bulk analysis (Figure 3.4B).

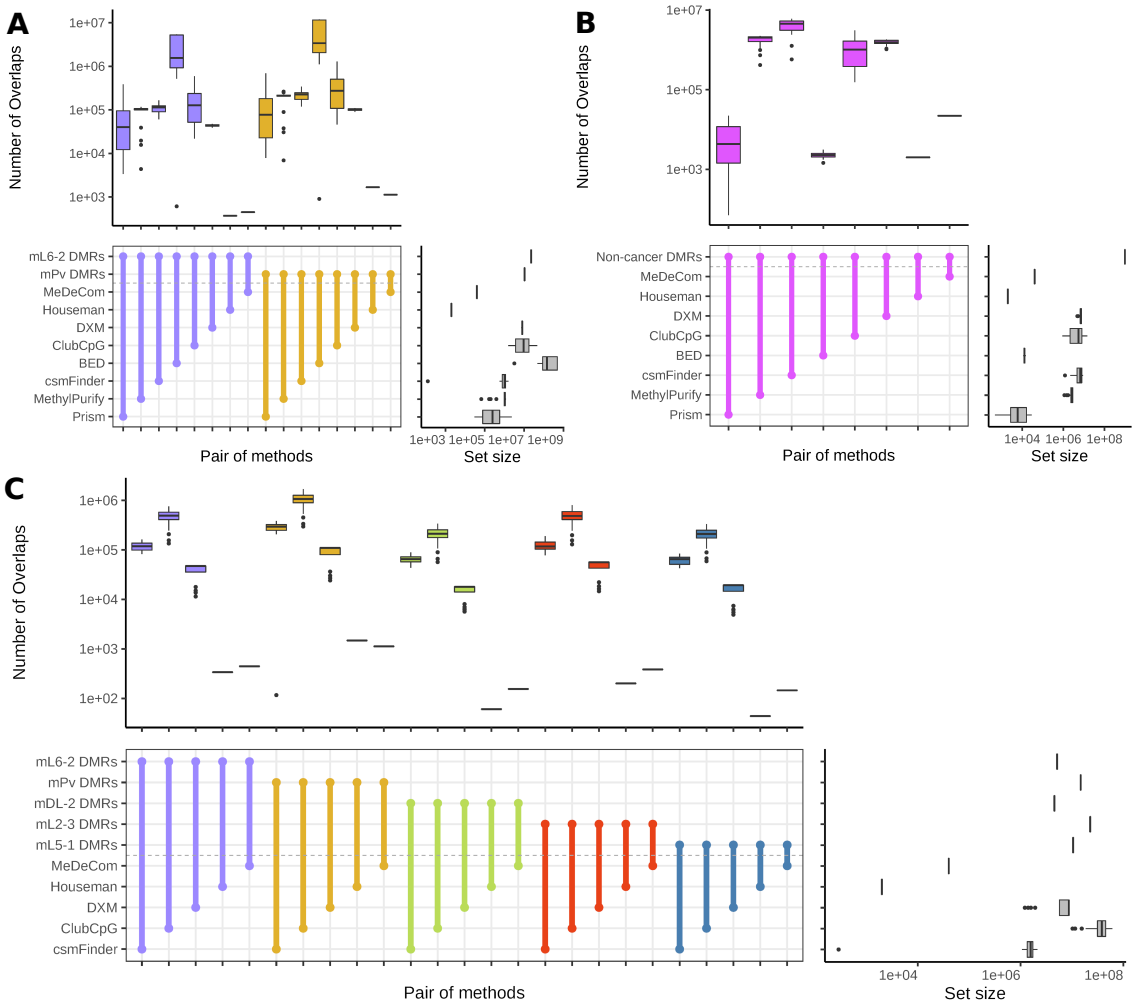


Figure 3.2: Overlaps between DMRs and selected informative regions. (A) 2 cell-type mouse neuronal pseudo-bulks (B) Tumour-normal pseudo-bulks (C) 5 cell-type mouse neuronal pseudo-bulks. The coloured boxplots at the top present the number of overlapping informative regions with DMRs over all pseudo-bulk samples. Two groups where overlaps were calculated are connected by a line in the middle. The right grey boxplot shows the size of the region set in each method or DMR. We mark the result for different cell types by different colours.

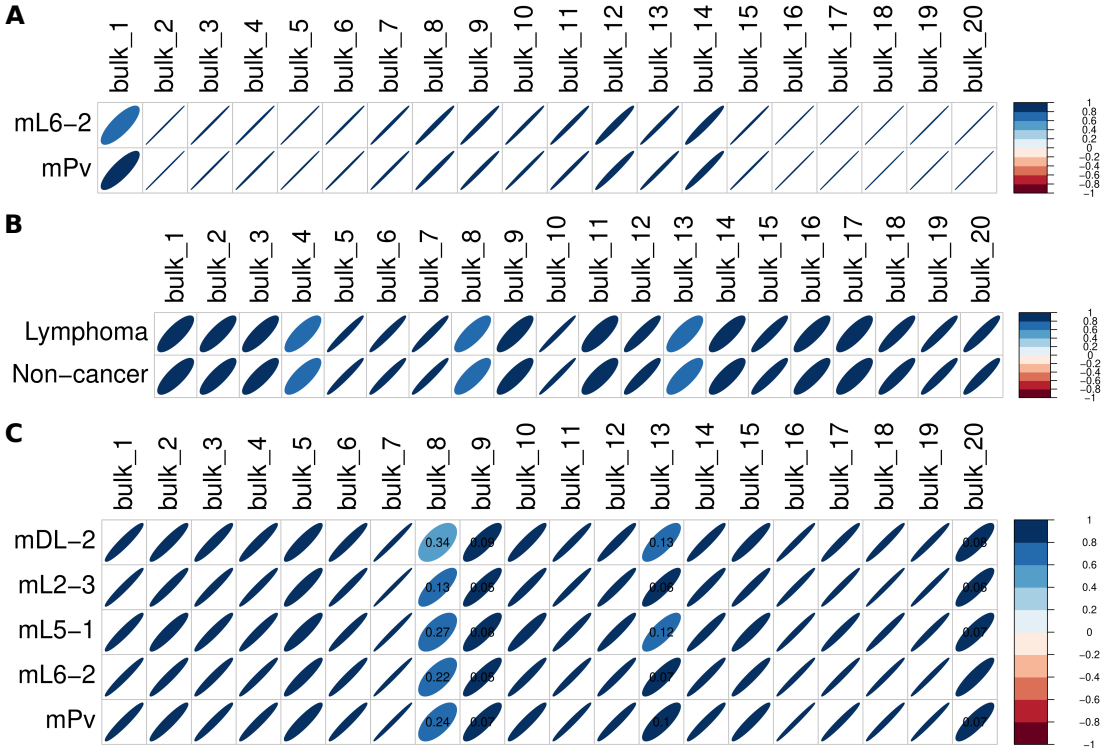


Figure 3.3: Correlation between the number of overlapping regions with DMRs and the size of region set. (A) 2 cell-type mouse neuronal pseudo-bulks (B) Tumour-normal pseudo-bulk (3) 5 cell-type mouse neuronal pseudo-bulk. For every pair of cell-type and pseudo-bulk samples, the correlation was calculated for all benchmarked methods. Non-significant correlations (p-value > 0.05) have a p-value written on the ellipse.

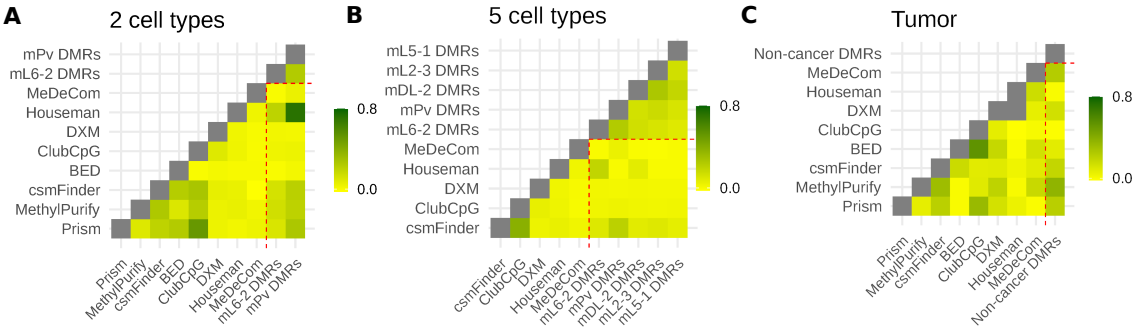


Figure 3.4: Genomic correlation between the selected informative regions and DMRs. **(A)** 2 cell-type mouse neuronal pseudo-bulks. **(B)** 5 cell-type mouse neuronal pseudo-bulks. **(C)** Tumour-normal pseudo-bulks. A higher score is indicated with a darker green colour.

Genome annotation is another broadly used analysis method in bioinformatics to investigate functional elements over genomic sequences such as promoters[8] or enhancers[9]. Relating methylation patterns with specific genome annotations can explain the mechanism of gene expression controlled by epigenetic modifications. Many studies have associated DNAm in promoter regions with gene silencing and cell-type specificity [Phillips et al., 2008, Suzuki and Bird, 2008]. Negative correlations between the gene expression level and DNAm in the first intron regions also have been reported [Anastasiadi et al., 2018].

Figure 3.5 shows the genomic annotation results of the selected informative regions for each method. The ratio of promoter regions is higher in the selected informative regions than in DMRs. In particular, regions detected by DXM include the highest ratio of promoters. This corresponds to the previous finding that promoter methylation plays a role as a gene regulator forming cell-type identity. Furthermore, both selected regions and DMRs involve a much larger number of distal intergenic regions in the tumour-normal pseudo-bulk data set than the other two data sets created from mouse neuronal cells. This concurs with studies showing that dominant aberrant methylation patterns are present in intergenic regions in lymphoma and leukaemia [Kretzmer et al., 2015, Almamun et al., 2017].

[8]Promoter is a genomic region where proteins bind to start transcription of a gene
[9]Enhancer is a genomic region where proteins bind to enhance transcription of a gene

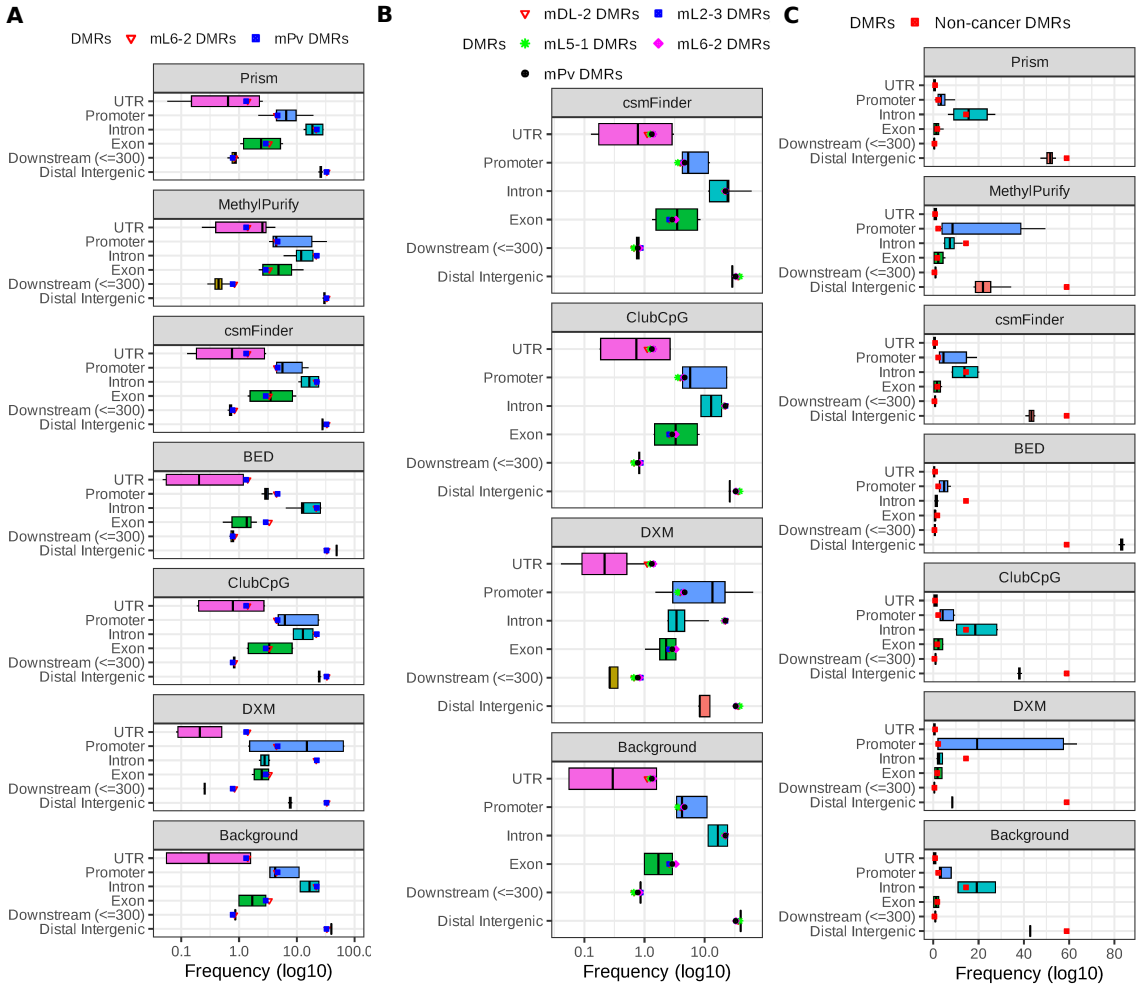


Figure 3.5: Genome annotation of selected informative regions and DMRs. The annotation was conducted in individual bulk samples. The box plot depicts the distribution of annotation frequencies over the bulk samples. The median, the first, and third quantiles are shown as the middle bar, at each end of the box. Annotation of DMRs are marked as different shapes of dots in the box plots. The same analysis was done for all CpGs of each bulk sample, which is referred to as 'background' to present the original distribution of annotations.

Furthermore, we examined CpG-wise methylation differences in the selected informative regions and DMRs (Figure 3.6). The analysis was done by comparing methylation beta-values between pure cell-type samples at examined CpG sites. This analysis is necessary because, even if DMRs are selected based on the CpGs showing distinct methylation patterns between two cell types, all CpGs in DMRs do not necessarily present cell-type specific methylation patterns. DSS, which is the DMR calling method used for our benchmarking study, selects regions based on the number of CpGs with a statistically significant methylation difference [Feng et al., 2014]. The statistics are calculated by a hypothesis test with a null hypothesis that a CpG methylation pattern difference is zero between two cell types (see Section 2.1.4). Other methods, such as MethylKit or DMRcaller, select regions where region-wise methylation level differs between two cell types, not CpG-wise methylation level [Akalin et al., 2012, Catoni et al., 2018].

For bi-component pseudo-bulks (2 cell-type mouse neuronal and tumour-normal pseudo-bulk samples), we subtracted the beta-value of one cell type from the other. Then, a difference > 0 means hypermethylation of the cell type whereas a difference < 0 means hypomethylation of the cell type. A difference value of 0 means that there is no difference between the two cell types in methylation beta-value. Figure 3.6A and B show the distributions of beta-value difference in the bi-component pseudo-bulks. Although CpGs covered by DMRs mostly have hyper- or hypomethylation in the corresponding cell type, a noticeable number of CpGs were also detected not to have a significant methylation difference between the two cell types. DLBCL DMRs tend to be hypermethylated, while both mouse neuronal cell-type DMRs have much more hypomethylated CpGs. This result agrees with the exceeding number of hypermethylations discovered specifically in cancer [Das and Singal, 2004]. Array-based methods also involve the majority of CpGs with cell-type specific methylation patterns, since the regions were selected from DMRs or based on methylation variance. However, sequencing-based methods cannot detect as many CpGs with cell-type specific signals compared to DMRs or array-based methods. In all sequencing-based method results, a peak was observed at zero in the methylation beta-value difference distribution, which means that most CpGs do not present a methylation difference between two compared cell types.

For five cell-type mouse neuronal bulks, in the ideal case of informative regions, a single cell type is supposed to present a unique pattern and all others cannot be distinguished from each other. Therefore, we calculated the absolute difference of minimum and maximum among five methylation beta-values at individual CpG sites (Figure 3.6C). The absolute value closer to 1 implies a clearer cell-type specific methylation pattern. The largest number of CpGs in DMRs and the regions for array-based methods have a highly different methylation pattern in one cell type. However, within the regions detected by sequencing-based methods, far more CpGs have low methylation beta-value difference (<0.3).

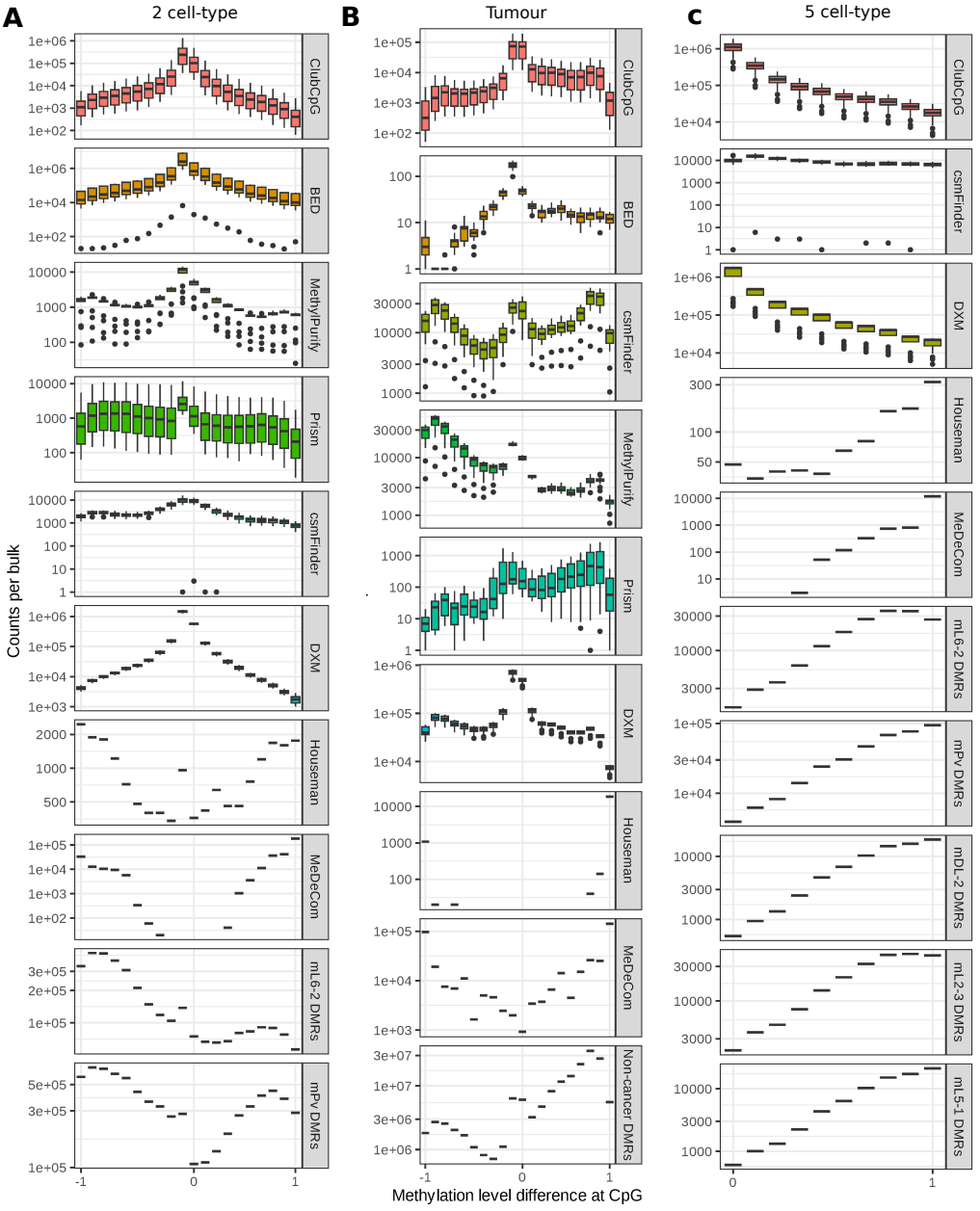


Figure 3.6: Methylation beta-value difference at CpGs in the selected regions. (A) 2 cell-type mouse neuronal pseudo-bulk result. (B) Tumour-normal pseudo-bulk result. (C) 5 cell-type mouse neuronal pseudo-bulk result. For the bi-component bulks (2 cell-type mouse and tumour-normal samples), the difference between two pure cell-type methylomes was calculated at CpGs located in the regions. For the five cell-type mouse samples, the difference was calculated between min and max methylation beta-values among five pure cell-type methylomes. The difference values are binned by 0.1, and the boxplots contain the results of 20 bulks in each data set.

3.7 Cell-type composition estimation

Based on the selected informative regions in the previous step, sequencing-based cell-type deconvolution methods estimate global cell-type compositions within given bulk samples. In this section, we evaluate the cell-type composition estimation performance for all benchmarked methods. For a fair comparison, reference-based and reference-free methods are assessed separately.

Mouse neuronal cell-type deconvolution

For two cell-type pseudo-bulk analysis shown in Figure 3.7, only mPv cell-type results are shown because two cell-type proportions sum up to one, and the cell-type deconvolution performance measured for both cell types makes it redundant. Houseman's method and coMethy yield the lowest median absolute error among reference-based and reference-free methods each (Figure 3.7A). It is noted that, for the high percentage of mPv cell type in the bulks, MethylPurify and Prism performed better than coMethy which overestimates the proportion for those samples (Figure 3.7B).

In the analysis of five cell-type pseudo-bulks shown in Figure 3.8, Houseman's method again performed best among the reference-based methods, but DXM showed the best performance among reference-free methods (Figure 3.8A). According to the performance comparison per cell type, the mL2-3 cell-type proportion is the most difficult to estimate for coMethy, while ClubCpG and MeDecom are the most inaccurate in estimating mDL-2 cell-type proportion (Figures 3.8B). Most of the methods struggled more with estimating cell-type compositions in five cell-type bulks, but Houseman's method and DXM achieved lower median absolute error compared to the result for two cell-type bulks.

Overall, reference-based methods (excluding BED) yielded lower median absolute error values than reference-free methods for two cell-type pseudo-bulks (Figure 3.7A). For five cell-type pseudo-bulks, the reference-free method DXM outperformed the reference-based method ClubCpG (Figure 3.8A). From the individual bulk result analysis, we have discovered that ClubCpG can produce an estimation below 0 or above 1 in the case of extremely low or high ground-truth cell-type proportion because ClubCpG uses a linear regression model without limiting a range of estimated values (Figure 3.7B and 3.8B).

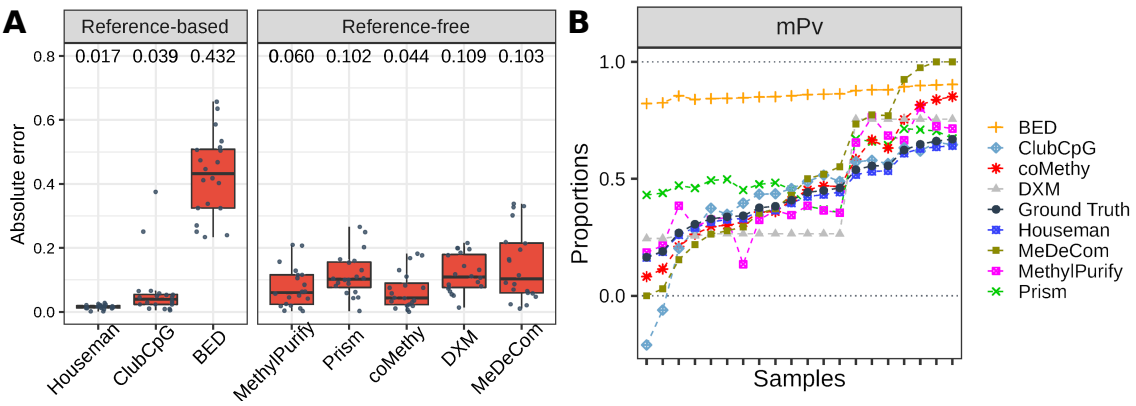


Figure 3.7: Mouse neuronal two cell-type pseudo-bulk composition estimation. (A) Absolute error calculated between ground-truth and predicted cell-type proportion. The number above the box plot shows the median value. Error values were computed only in the mPv cell type to prevent redundancy (B) Inferred proportions by cell-type deconvolution methods and the ground-truth proportion (black). Bulk samples are ordered by the ground-truth cell-type proportion.

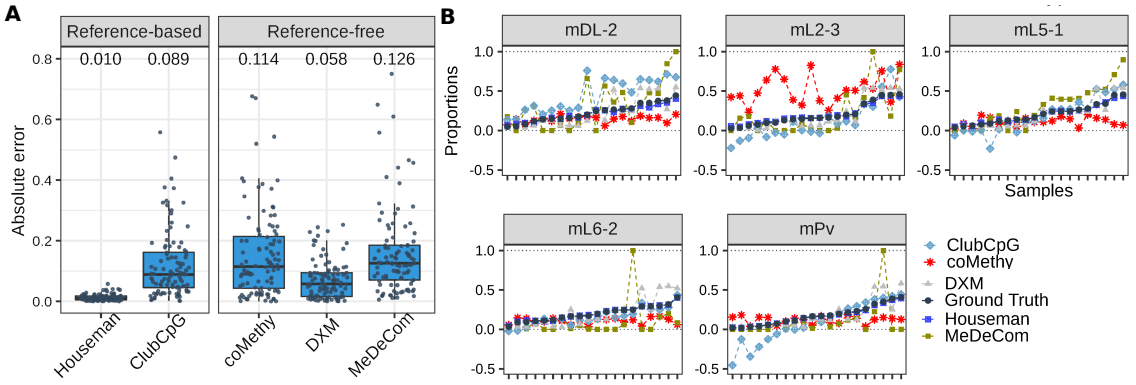


Figure 3.8: Mouse neuronal five cell-type pseudo-bulk composition estimation. (A) Absolute error calculated between the ground-truth and predicted cell-type proportions. The number above the box plot shows the median value. The error values were obtained for all five cell types. (B) Inferred proportions by cell-type deconvolution methods and the ground-truth proportion (black) for different cell types. Bulk samples are ordered by the ground-truth cell-type proportion.

Tumour purity estimation

Tumour tissues contain not only tumour cells, but also normal cells associated with tumour such as epithelial, immune, or stromal cells [Yoshihara et al., 2013]. Therefore, purifying tumour signals from tumour tissue methylomes makes it possible to obtain pure tumour DNAm profiles and measure the percentage of tumour cells within. Once the contamination of non-targeted normal cells is high, clinical analysis of tumour tissues gets more erroneous [de Ridder et al., 2005]. In conclusion, estimating accurate tumour purity is crucial for tumour studies. Out of the benchmarked methods, BED and MethylPurify are particularly designed for such bulk samples originating from tumour and normal subpopulations. Yet, we assessed all methods with tumour-normal pseudo-bulks to see if these methods actually perform better than others in tumour purity estimation (Figure 3.9).

The results of reference-based methods had the same tendency as the mouse neuronal cell-type deconvolution results: Houseman's method outperformed all other methods. It even achieves a lower error value in tumour cell-type deconvolution than in mouse neuronal cell-type deconvolution, despite the high complexity of tumour methylomes as described in Section 2.1.2. In the reference-free method analysis, MeDeCom performed tumour purity estimation most accurately.

In Section 2.1.2, we described that many of the recent studies in liquid biopsy have clarified the usage of ctDNA in blood plasma samples for non-invasive early cancer diagnosis [Martins et al., 2021, Egyud et al., 2019]. The main challenge in ctDNA analysis is the extremely low percentage of ctDNA occurring, especially at the early stage of tumour development, which requires a highly sensitive cell-type deconvolution model.

Therefore, we assessed the benchmarked methods with tumour-normal pseudo-bulks including extremely low percentage of tumour-derived reads. Ten more bulk samples were newly created by adding reads from DLBCL increasing the percentage by 0.1% from 0.1% to 1%. As a performance metric, we used MAPE, rather than MAE, to show more comparable performance differences within the excessively small range of ground-truth values.

Consistently, the best performance is achieved by Houseman's method among reference-based methods (Figure 3.10A). In the reference-free method analysis, MethylPurify outperforms all other methods. Notwithstanding, the estimation of tumour purity was done more inaccurately compared to other pseudo-bulk analyses and only Houseman's method achieved a prediction below 1% (Figure 3.10B). Particularly, Prism, MeDeCom, and coMethy are not able to purify rare tumour signals and often yield a tumour purity estimate above 20%.

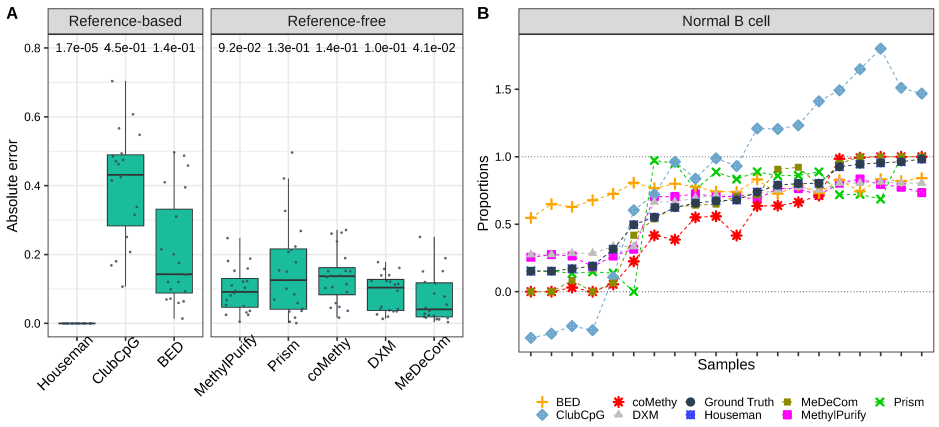


Figure 3.9: Tumour-normal pseudo-bulk composition estimation. (A) Absolute error calculated between ground-truth and predicted cell-type proportion. The number above the box plot shows the median value. Again, the error values were calculated only in the normal B cell type because of the redundant estimated proportions of the two cell types. (B) Inferred proportions by cell-type deconvolution methods and the ground-truth proportion (black). Bulk samples are ordered by the ground-truth cell-type proportion.

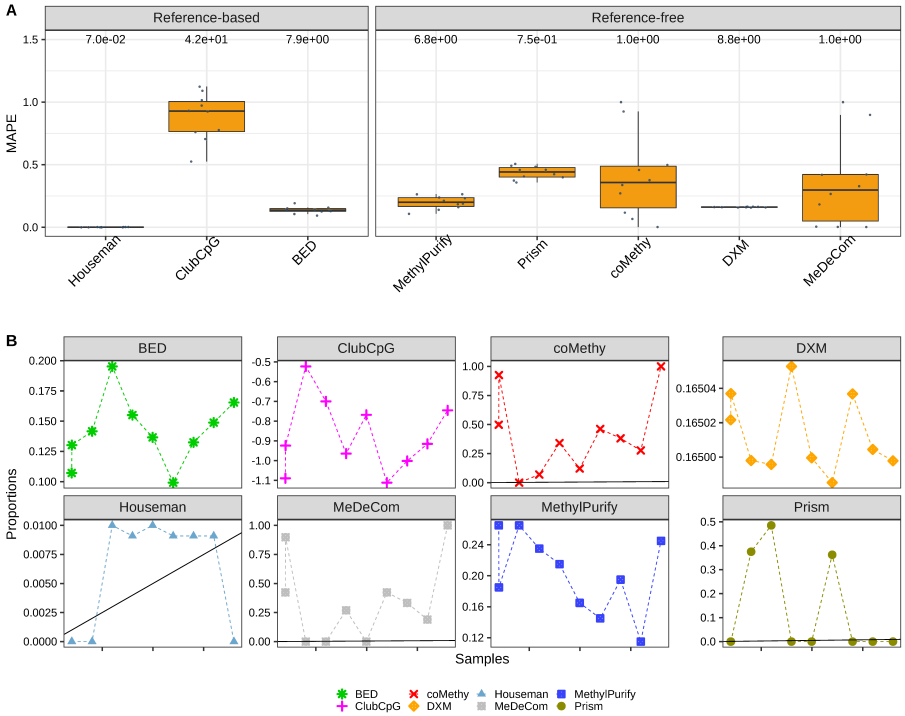


Figure 3.10: Rare cell-type pseudo-bulk composition estimation. (A) Mean absolute percentage error calculated between ground-truth and predicted cell-type proportion. The number above box plot shows the median value. (B) Inferred proportions by cell-type deconvolution methods. The ground-truth values are given as a black line in each facet. Pseudo-bulk samples are ordered by ground-truth tumour purity.

3.8 Influential factors of cell-type deconvolution performance

Ultimately, we examine the influence of informative region selection on cell-type composition estimation in sequencing-based cell-type deconvolution methods. To cope with different scales of values, the comparison is done by performance rank rather than the value itself. For the considered methods, the cell-type composition estimation error and genomic correlation values were averaged over all bulk samples, then the mean values were ranked. Here, array-based methods whose pipeline does not involve the informative region selection step have been excluded from the comparison.

For the two bi-component pseudo-bulk data sets, an inverse correlation has been found between the mean absolute error rank and mean genomic correlation rank (Figure 3.11). This indicates that when there are two cell types supposed to be discovered, the capability of the informative region selection step in recognising CpGs overlapping with DMRs is important to determine cell-type composition estimation performance. Yet, in five cell-type pseudo-bulks, an opposite tendency has been discovered.

The distribution of cell types within a bulk can be another factor affecting the cell-type

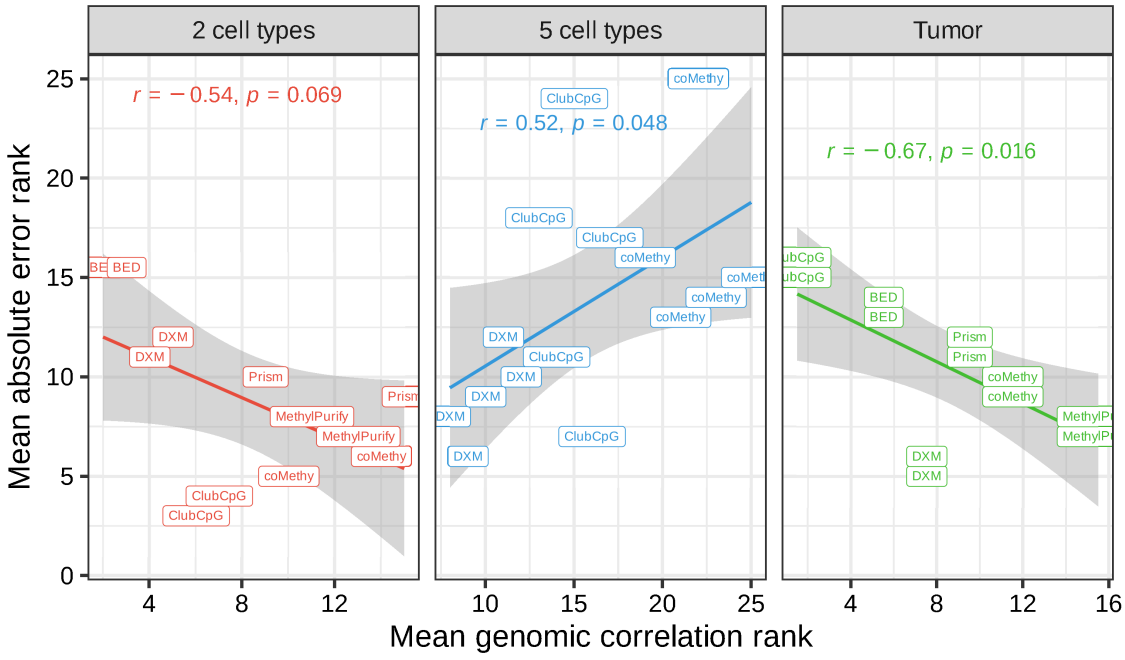


Figure 3.11: Influence of genomic correlation in cell-type composition estimation. Pearson correlation coefficient and p-value were calculated between the rank of mean absolute error and the rank of mean genomic correlation. The line with a grey background signifies the fitted linear model and the confidence interval of 0.95. Tumour-normal pseudo-bulk result has the same rank for two different cell types because DMRs were calculated by comparing only two cell types, DLBCL and non-cancer B cell.

deconvolution performance. A rare cell type may not present sufficient cell type-specific signals for accurate cell-type deconvolution, whereas equally distributed cell types may increase the complexity of overall methylation patterns across DMRs. Therefore, we investigate the relationship between cell-type composition estimation performance and the entropy of cell-type proportions. The mean absolute error value is calculated over all cell types in each bulk. As a result, the cell-type proportion entropy has been found as a factor commonly affecting cell-type deconvolution performance in the five cell-type mouse neuronal pseudo-bulk data set (Figure 3.12). With the exception of DXM, the entropy of cell-type proportions is negatively correlated with the mean absolute error between the predicted and the ground-truth values. Namely, cell-type deconvolution is conducted more accurately when the composition of five cell types is more uniformly distributed. This can be interpreted that a biased distribution of cell types, which yields a lower entropy, may not have enough cell-type specific signals from the minor cell type. However, a reverse result is obtained for DXM. We presume that the DXM algorithm, which determines the best fit among multiple random distributions without regularisation, may be more suitable for a biased cell-type distribution than other methods whose computational model is iteratively optimised. This trend is not necessarily observed for bi-component pseudo-bulk analyses. For example, Houseman's method has a positive correlation between the entropy and mean absolute error for two mouse neuronal cell-type bulks. The same tendency is observed for coMethy and Prism results for tumour-normal pseudo-bulk samples. This can be understood that two cell types uniformly distributed result in two similar distributions of methylation patterns in each DMR, which could confuse a model to find a correct match of cell type and methylation pattern distribution.

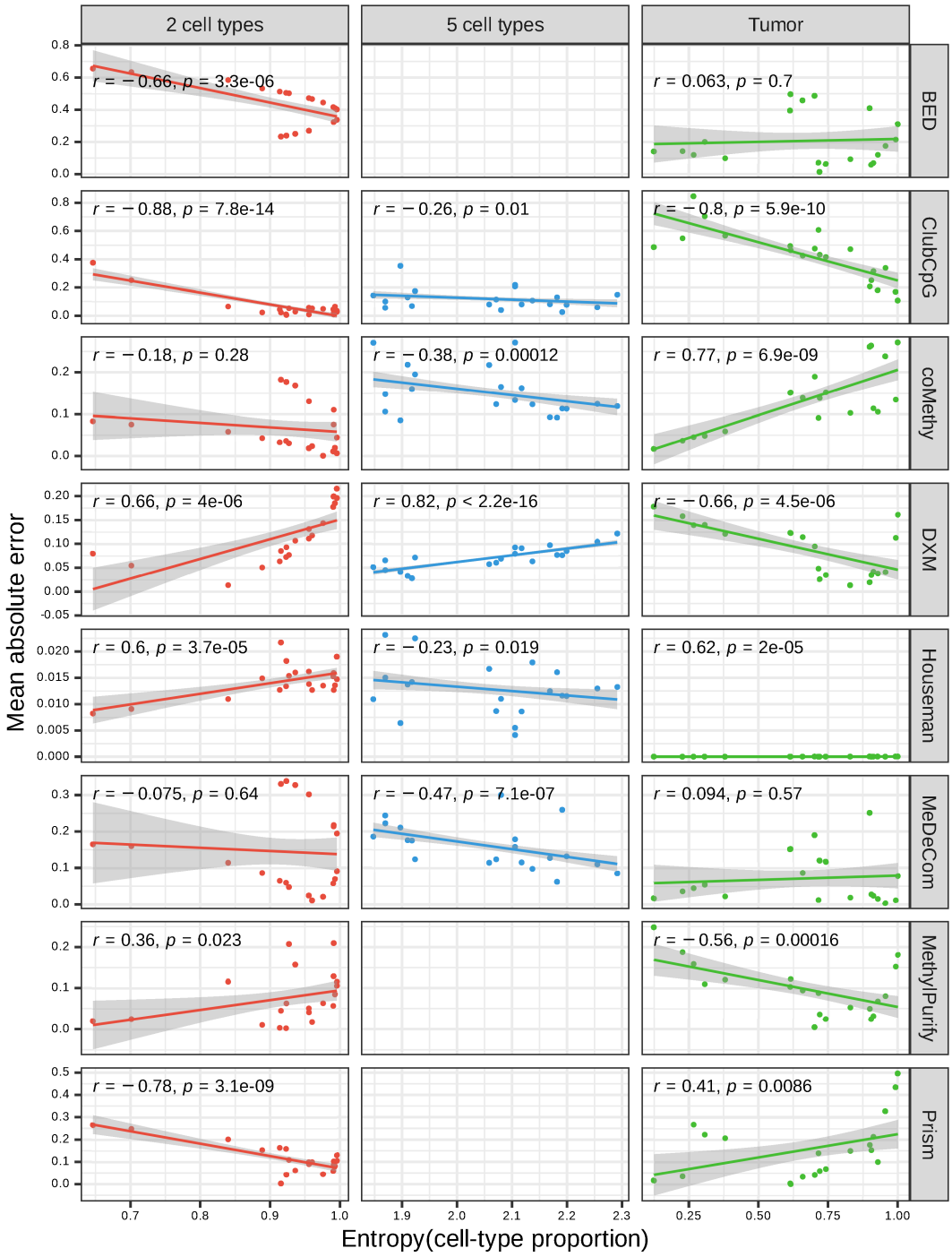


Figure 3.12: Correlation between cell-type proportion entropy and mean absolute error. Pearson correlation coefficient and p-value are given in each plot. Individual points indicate different bulk samples. A fitted linear function is shown as a line with a grey background indicating a confidence interval of 0.95.

3.9 Discussion

In this chapter, six existing sequencing-based cell-type deconvolution methods have been systematically evaluated together with two array-based methods as a comparison group. The array-based methods have been included particularly for assessing the capability of sequencing-based methods to leverage the unique features of read-level methylation patterns.

In order to cover the variability of biological data, three sets of pseudo-bulk samples were simulated mimicking different biological scenarios. Two and five cell-type mouse neuronal pseudo-bulk data sets represent the different compositions of subpopulations within bulk samples, whereas the tumour-normal pseudo-bulk data set generated from DLBCL samples was used for testing the methods on tumour bulk samples. Pseudo-bulk samples were generated by merging randomly sampled reads from pure cell-type samples with known proportions.

Regarding the algorithmic design, all benchmarked sequencing-based methods have two major steps: informative region selection and cell-type composition estimation. During the informative region selection step, the methods pre-filter genomic regions where cell type-specific signals are not presented in the methylation patterns. Then, cell-type composition estimation is conducted only using the DNAm profiles in the remaining regions. Therefore, the assessment has been done separately for each step, and we investigated whether the performance of informative region selection affects the final cell-type deconvolution result.

For the evaluation of informative region selection, DMRs were considered the gold-standard genomic region involving cell type-specific signals. According to the comparison between DMRs and selected genomic regions by individual methods, ClubCpG yielded the largest number of overlaps with DMRs for the mouse neuronal pseudo-bulk samples, though csmFinder achieved the highest genomic correlation value. ClubCpG did not have a high genomic correlation because a large number of overlapping regions is rather caused by a large region set size (Figure 3.2).

The assessment of the cell-type composition estimation step was done mainly based on absolute error. Considering the prior knowledge provided in a reference-based manner, the methods have been separately evaluated according to the requirement of reference data. Houseman's method clearly outperformed the others in the reference-based method evaluation. Among the reference-free methods, coMethy most accurately estimated cell-type composition for the mouse neuronal pseudo-bulk data set.

Cancerous bulks comprised of normal and tumour cell types generally contain more complex structures of subpopulations. In addition, DNAm patterns gain abnormalities over tumour development [McCabe et al., 2009]. To inspect the applicability of the benchmarked methods to tumour samples, we conducted the same evaluation with another pseudo-bulk

data set made of normal B cell and DLBCL cell types. In the informative region selection result, csmFinder detects the largest overlapping regions with DMRs, as well as accomplishes the highest genomic correlation. However, for the cell-type composition estimation step, two array-based methods, Houseman's method and MeDeCom, inferred the most accurate cell-type composition in reference-based and reference-free method evaluations, respectively. In rare cell-type pseudo-bulk analysis, none of the benchmarked methods achieved a reasonably positive correlation between the ground-truth and estimated cell-type proportions, which implies that these methods cannot be used for the ctDNA analysis which requires to deal with a very low percentage of tumour cells.

Finally, the analysis verified that selecting valid informative regions matters for accurate cell-type composition estimation. The negative rank correlation between mean absolute error and genomic correlation in bi-component pseudo-bulk samples emphasises that methods detecting more genomic regions overlapping with DMRs better estimate the cell-type composition (Figure 3.11). However, when there are more cell types, this is not always the case. For five mouse neuronal cell-type pseudo-bulks, the entropy of cell-type distribution rather showed a negative correlation with the absolute error in most cases (Figure 3.12). This result indicates that the cell-type distribution is more impactful in deciding the performance of cell-type deconvolution for a larger number of cell types.

Consequently, the benchmarked methods overall inferred reasonable cell-type proportions, but we have discovered that sequencing-based cell-type deconvolution methods do not perform significantly better than array-based deconvolution methods in terms of cell-type composition estimation. Houseman's method undoubtedly outperformed other sequencing-based methods and MeDeCom performed best among reference-free methods in the tumour-normal pseudo-bulk analysis. The biggest challenge of the sequencing-based approach is addressing the high complexity of methylation patterns caused by the possibly disparate methylation states in different DNA molecules at the same CpG site. This information is simply averaged yielding one beta-value in array-based profiling, which makes computational/statistical modelling smooth, yet loses the single-molecule resolution of DNAm pattern. Removing redundant information by selecting valid informative regions is also crucial to perform accurate cell-type deconvolution. Therefore, sequencing-based methods need to be able to eliminate uninformative methylation patterns to prevent a model bias towards non-cell type-specific methylation patterns, while at the same time, preserving informative methylation patterns.

Another problem that arose is that the benchmarked sequencing-based methods do not exploit the advantages of read-level methylomes for accurate inference. For example, csmFinder + coMethy converts methylation patterns obtained from sequencing data into a matrix, whereas DXM only finds the best fit to given data out of a thousand randomly generated distributions, rather than performing regression, which would improve model optimisation. Moreover, Prism retains only fully methylated and unmethylated reads

during the informative region selection. Not involving partially methylated reads is critical for tumour analysis where PMDs and ASM are not avoidable. These approaches over-simplify the obtained methylation patterns in common which results in a less accurate estimation of cell-type compositions. Thus, better modelling for cell-type deconvolution needs to be designed to take advantage of read-level methylomes.

To sum up, our benchmarking study conveys an apparent paradigm of sequencing-based cell-type deconvolution. Not only do the results provide a systematic comparison of currently available sequencing-based cell-type deconvolution methods, but the study also suggests that there is room for methodological improvement. The intrinsic benefit of sequencing-based methylation profiling should allow more accurate cell-type composition estimation based on read-level information compared with array-based profiling. Taken all together, we do see the necessity of a new sequencing-based cell-type deconvolution method designed to extract appropriate cell type-specific signals, while handling confounding factors and outliers, which make read-level methylome analysis challenging. Therefore, the following chapters of this thesis focus on the development of a novel sequencing-based cell-type deconvolution method and the assessment of the developed method.

Chapter 4

Transformer-based cell-type deconvolution model for tumour read-level methylomes

4.1 Introduction

Transformers were proposed by [Vaswani et al., 2017] as a deep learning model which associates tokenised words with each other in given sentences. As explained in Section 2.2.4, the multi-head self-attention mechanism enables such associations by calculating the scaled dot-product of the query, key, and value vectors, see Equation (2.14). The recent advancements in natural language processing (NLP) have been predominantly achieved by Transformer-based models. [Dong et al., 2018] replaced the 1D attention mechanism in Transformers with a 2D attention mechanism for speech feature sequences which are represented in a time×frequency space and achieved competitive performances in speech recognition tasks. The BigBird model showed better performance than other large language models in the question-answering (QA) task by employing a sparse attention mechanism in the Transformer model [Zaheer et al., 2020].

Bidirectional Encoder Representations from Transformers (BERT) is one of the most commonly used Transformer-based models. BERT particularly adopts the encoders of the original Transformer model [Devlin et al., 2018]. [Devlin et al., 2018] confirmed the broad extensibility of BERT to varying NLP tasks by testing it with multiple NLP benchmarking data sets including the Stanford Question Answering Data set (SQuAD) [Rajpurkar et al., 2016, Rajpurkar et al., 2018], the SWAG data set for common sense inference [Zellers et al., 2018], and the General Language Understanding Evaluation (GLUE) benchmarking data set [Wang et al., 2018]. The bidirectional training and global context learning are significant advantages of BERT compared to other deep neural networks as described in Section

2.2.4. The bidirectional training overcomes the long training time and vanishing gradients problems of recurrent neural networks (RNNs), whereas the global context learning makes BERT more suitable for sequential data than convolutional neural networks (CNNs) mainly focusing on a local context within the kernel window.

Transformers have also been broadly applied to biological sequential data. [Ji et al., 2021] developed DNABERT based on Transformers to predict genomic features such as promoters or enhancers. [Gwak and Rho, 2022] classified metagenome sequences into virus types using a Transformer-based model. The application of Transformers has also shown superiority in DNA methylation (DNAm)-related tasks. For instance, methBERT successfully detected DNAm in long-read sequencing reads by using Transformers [Wang et al., 2023], while [De Waele et al., 2022] suggested a Transformer-based model combined with CNNs for single-cell DNAm pattern imputation.

In this chapter, we introduce *MethylBERT*, a new cell-type deconvolution model based on Transformers for tumour read-level methylomes. Despite its successful application to DNAm-related tasks, BERT has not been used for cell-type deconvolution for DNAm data. Table 4.1 gives an overview of previous methods for sequencing-based DNAm data categorised by purposes and baseline models. Transformers have been utilised for methylation site prediction, nanopore (long-read) methylation calling, and DNAm imputation, but not for cell-type deconvolution. On the other hand, although there have been various classical machine learning methods such as the hidden Markov model (HMM) or RNNs suggested for cell-type deconvolution, Transformers have not been used for this purpose. Thus, to our best knowledge, MethylBERT is the first application of Transformers for cell-type deconvolution using DNAm data. This work is published in [Jeong et al., 2023a].

This chapter explains the MethylBERT model in the following three sections. Section 4.2 describes the BERT architecture employed in MethylBERT and training strategies separated into pre-training and fine-tuning. Section 4.3 explains maximum likelihood estimation applied to the probability of cell types given reads estimated by the MethylBERT network. Section 4.4 describes the general training schemes for MethylBERT and implementation details for computational speed improvement.

Table 4.1: Overview of methods for sequencing-based DNAm data analysis

	HMM/Beta-binomial/ Bernoulli-based model	RNN-based model	Transformer-based model
Methylation site prediction	iDNA-MS [Lv et al., 2020] SOMM4mC [Yang et al., 2020b]	DeepTorrent [Liu et al., 2021]	iDNA-ABT [Yu et al., 2021] MuLan-Methyl [Zeng et al., 2023]
Nanopore methylation calling	Nanopolish [Loman et al., 2015] SignalAlign [Rand et al., 2017]	DeepMod [Liu et al., 2019] DeepSignal [Ni et al., 2019]	Rockfish [Stanojević et al., 2022] methBERT [Wang et al., 2023]
DNA methylation imputation	Melissa [Kapourani and Sanguinetti, 2019] METHimpute [Taudt et al., 2018]	DeepCpG [Angermueller et al., 2017]	CpG Transformer [De Waele et al., 2022]
Cell-type deconvolution for tumour	CancerDetector [Li et al., 2018] DXM [Fong et al., 2021]	DISMIR [Li et al., 2021]	**MethylBERT** [Jeong et al., 2023a]

4.2 MethylBERT: BERT-based read classification

We developed a novel Transformer-based model, MethylBERT, by modifying the original BERT model [Devlin et al., 2018]. MethylBERT embeds read-level methylation patterns and uses the embedding for classifying reads into cell types.

A main difference between MethylBERT and the original BERT is the used input data and corresponding embeddings (Table 4.2). Unlike BERT requiring word, sentence and position embeddings as described in Section 2.2.4, MethylBERT needs the following information from an individual read: a reference DNA sequence, methylation patterns, and a DMR label indicating in which DMR the read is located. Therefore, the word embeddings in the BERT model are replaced with 3-mer DNA sequence embeddings in MethylBERT. Also, methylation embeddings, instead of sentence embeddings, are incorporated in MethylBERT. Methylation embeddings have three values: 0 for unmethylated CpG, 1 for methylated CpG, and 2 for non-CpG. Since MethylBERT is designed for both paired-end and single-end aligned reads, we removed the sentence embeddings that indicate to which sentence individual tokens belong out of two input sentences, so the MethylBERT model can handle single-end aligned reads which do not have a designated pair. Position embeddings remain the same in MethylBERT.

Table 4.2: Comparison of input embeddings between the original BERT and MethylBERT.

Original BERT	MethylBERT
Word embeddings	DNA sequence 3-mer embeddings
Sentence embeddings	CpG methylation embeddings
Position embeddings	Position embeddings

MethylBERT consists of three steps (Figure 4.1). The first step is the pre-training of the MethylBERT network using reference genome data. The reference genome data is processed into 3-mer DNA sequences. Second, the model is fine-tuned for the read classification task. The input data of the fine-tuning step are DNAm sequencing reads whose DNA sequences and methylation patterns are processed into 3-mer and binary sequences, respectively. In the third step, we calculate the tumour purity of a tumour-normal bulk sample based on the class probability of individual reads calculated by the MethylBERT network. The last step also includes the estimation of the model precision using the Fisher information and the adjustment of estimated tumour purity based on the skewness of the purity distribution. The details of each step except for tumour purity estimation are described in the following subsections. The tumour purity estimation step is described in Section 4.3.

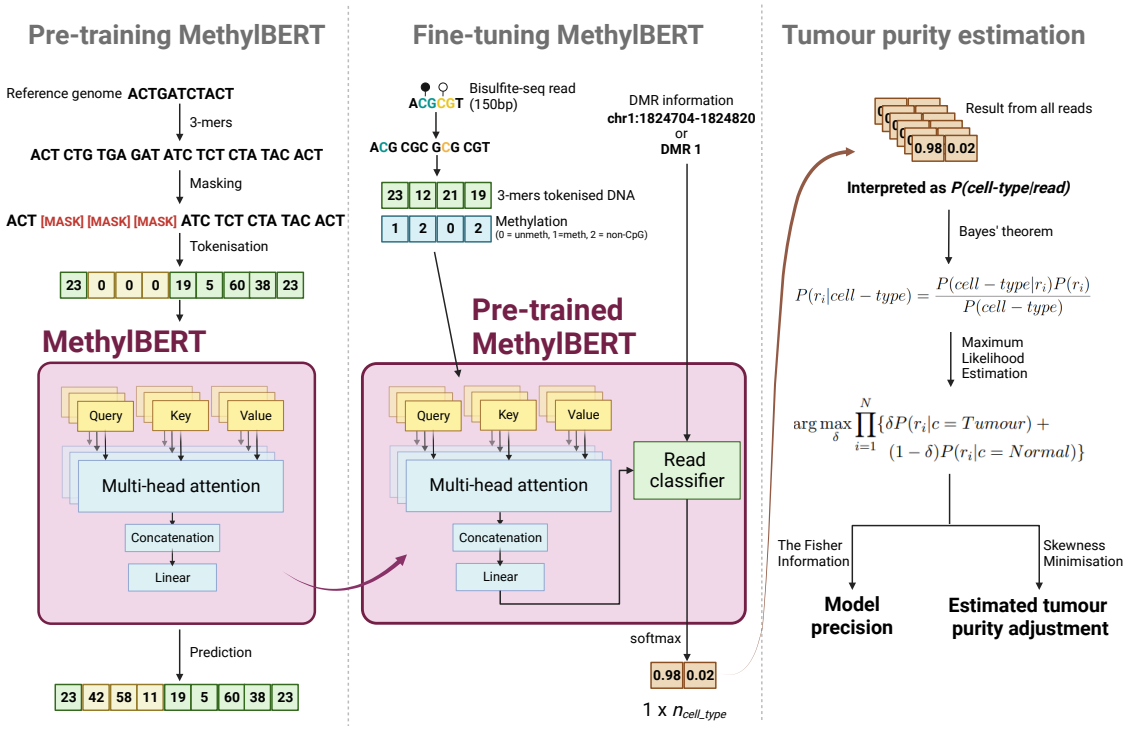


Figure 4.1: Overview of MethylBERT comprised of three steps.

4.2.1 Pre-training of MethylBERT

As explained in Section 2.2.4, pre-training of BERT is done in an unsupervised manner via masked language model (MLM) and next sentence prediction (NSP). The unsupervised pre-training makes it easier to train the BERT model with a huge amount of unlabelled data and diminishes the necessity for task-specific neural network architectures [Devlin et al., 2018]. MethylBERT pre-training is particularly inspired by DNABERT, which showed a high performance in various predictions of genomic features like promoters or enhancers on DNA sequences [Ji et al., 2021].

The input data of pre-training are DNA sequences from the entire genome. The genome is divided into 510 bps segments, and 3-mer sequences are generated from the split genome sequences. Thereafter, each 3-mer segment is converted to a token using a look-up table. We note that the methylation embeddings, which are only available when there are methylation patterns, are filled with zeros during pre-training, since the reference genome does not involve any DNAm. Although the original DNABERT pre-training uses a randomly sampled sequence length in the range between 5 and 510, it is fixed as 510 in MethylBERT because the random sampling of sequence length did not show a significant performance change. Following the MLM scheme of the original BERT model [Devlin et al., 2018], 15% of tokens in a 3-mer sequence are masked during pre-training. Within the 15%, different tokens are used for masking. 80% of the chosen 15% tokens are masked with a [MASK] token and 10% are replaced with randomly selected tokens. The remaining 10% of the chosen 15% tokens remain unchanged. Since the input sequences are processed into 3-mer tokens, the left and right neighbouring tokens of the selected token are also masked, so that the MethylBERT model avoids predicting the masked token just by looking at the neighbouring tokens.

In total, we use 69 labels for tokens (64 tokens for 3-mer tokens with four DNA nucleotides adenine, cytosine, guanine and thymine and 5 special tokens [PAD], [UNK], [EOS], [SOS] and [MASK] listed in Table 4.3). The pre-training was done with the categorical cross-

Table 4.3: Special tokens in BERT modeling

Token	Meaning
[PAD]	Padding token
[UNK]	Unknown token (when the given token is not in the look-up table)
[EOS]	End of sequence
[SOS]	Start of sequence
[MASK]	Masked token

entropy loss $L_{pre-training}$ calculated over all masked tokens $t_i \in \{t_1 ... t_T\}$:

$$L_{pre-training} = -\sum_{i=1}^{T}\sum_{l=1}^{69} y_l^{t_i} \cdot log(\hat{y}_l^{t_i}), \tag{4.1}$$

where $y_l^{t_i}$ and $\hat{y}_l^{t_i}$ indicate the logit and the one-hot encoded ground truth of label l for token t_i.

4.2.2 Read-level methylation pattern classification in MethylBERT

The fine-tuning of MethylBERT is conducted for the read classification task (Figure 4.2). The input data for the fine-tuning consists of read-level methylation patterns, a DNA sequence, and the label of DMR where the read originated from. For the DNA sequence, we use the reference genome sequence at the genomic position of each read rather than the aligned DNA sequences when each read is processed, so that the variations in nucleotides can be disregarded. Therefore, DNA sequences do not contain tumour-related information, yet they still can be an indicator of position in the region. Furthermore, in this way, the model can catch DNA context or motifs[1], which might be related to DNAm. Several studies have argued that using both DNA sequence and DNAm together can create a more sophisticated model for DNAm analysis [Li et al., 2021, Angermueller et al., 2017]. For the position embedding, we use absolute position embedding same as the original BERT model [Devlin et al., 2018].

The DNA sequence and the read-level methylation pattern are fed into the encoder after being embedded individually. The Transformer encoder in the MethylBERT network contains multiple encoder blocks. Each encoder block has a self-attention layer followed by three fully connected layers (denoted 'linear' in Figure 4.2). Although there is a pooling block in the BERT model, we did not use it because the pooling layer generates outputs for the NSP task which is not included in the MethylBERT pipeline.

The vector output by the encoder is then concatenated with an embedded DMR label. The DMR label is separately embedded outside of the encoder because it is not sequential. The region information is provided to handle the region-specific tumour methylome profile. The concatenated vector is given to the cell-type classifier network comprised of two fully connected layers. The classifier for K cell types outputs a K-dimensional vector that each value can be interpreted as the posterior probability of cell type given read $P(\text{cell type}|\text{read})$. Here, the cell type is either tumour (T) or normal (N), thus a 2-dimensional vector is the final output. During the fine-tuning process, the cross-entropy loss ($L_{fine-tuning}$) for softmax-normalised activation values for given methylation sequencing reads r_i and the groud-truth cell type $c_i \in \{T, N\}$ is calculated as:

$$L_{fine-tuning}(r_i, c_i) = -\sum_{c \in \{T,N\}} \mathbb{1}_{c_i=c} \cdot log \frac{exp(a_c(r_i))}{\sum_{c' \in \{T,N\}} exp(a_{c'}(r_i))} \tag{4.2}$$

where $a_c(r_i)$ is the final activation function of the MethylBERT network for cell type c given the read r_i.

[1]DNA motif is a short but recurring DNA sequence in the genome. It is known to be associated with gene regulation [D'haeseleer, 2006].

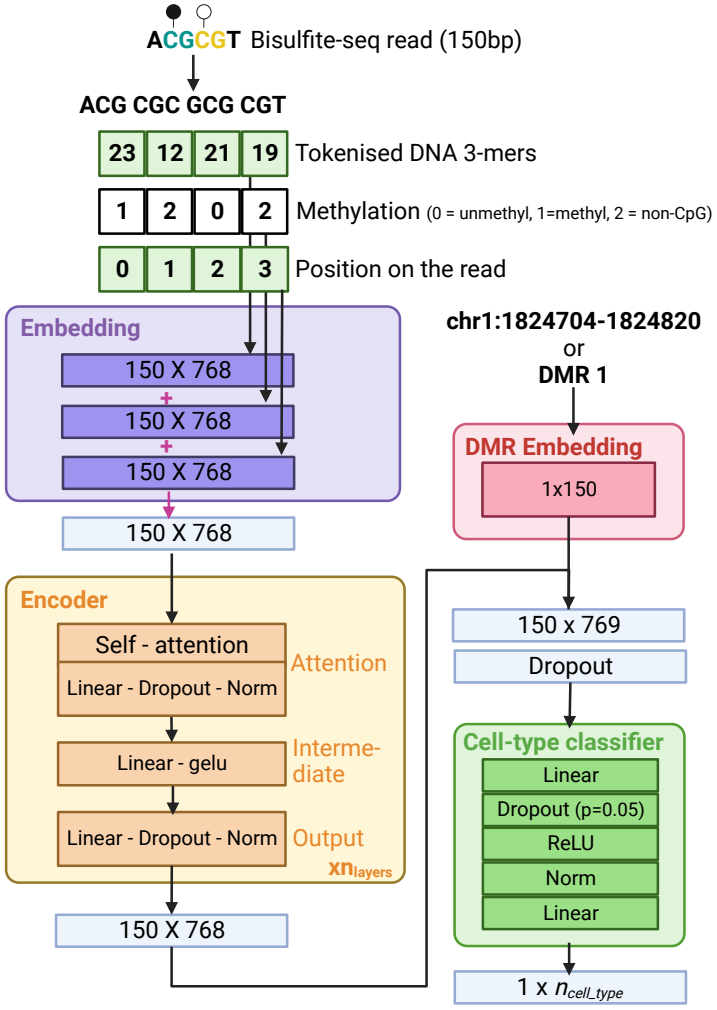


Figure 4.2: MethylBERT model architecture. Three different pieces of information (DNA 3-mer tokens, methylation patterns and the position on a read) are projected into separate embedding spaces before being feed-forwarded into the network. Then, the Transformer encoder part encodes the concatenated embeddings. The final embeddings of the read are given to the cell-type classifier together with the DMR embedding, so that the read is classified into cell types based on the provided information and encoded values.

Interpretation of MethylBERT output

To use the final output of MethylBERT for tumour purity estimation through MLE, we show that *the final output of MethylBERT represents the posterior probability of cell types given a read.* In this section, a simplified description of the MethylBERT network in Figure 4.3 is used.

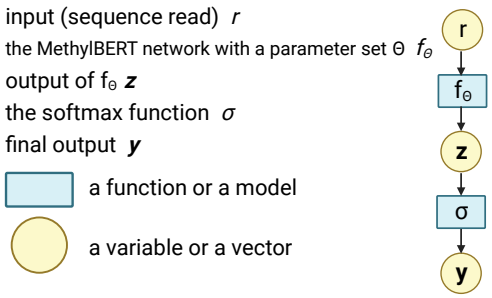


Figure 4.3: Simplified description of the MethylBERT network.

Let Θ be the parameter set of the MethylBERT network $f_\Theta(r)$ classifying the input r (read) into one of C labels (cell types). For such classification, f_Θ is designed to output $\mathbf{z} \in \mathbb{R}^C$ and the final output $\mathbf{y} = \{y_1, ..., y_C\} \in (0,1)^C$ is yielded via the softmax function (non-parametric) applied to $\mathbf{z}$. The standard softmax function $\sigma : \mathbb{R}^C \to (0,1]^C$ is defined as:

$$\sigma_i(\mathbf{z}) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \text{ for } \mathbf{z} = \{z_1, ..., z_C\} \in \mathbb{R}^C \tag{4.3}$$

where σ_i is the i^{th} member of the output vector. We show that $\mathbf{y}$ can be considered the posterior probability distribution of cell type given a read by confirming the following statements:

1. $\mathbf{y}$ represents a (discrete) probability distribution of P(cell type).

2. $y_i \in \mathbf{y}$ outputted from the MethylBERT network trained using a cross-entropy loss function described in Equation (4.2) can be an approximation of the posterior probability, $P(\text{cell type} = i|\text{read})$.

Statement 1. To clarify that the softmax function outputs probability values, we use the *probability axioms* introduced by [Kolmogorov, 1933]:

> **Probability axioms.** Let (Ω, F, P) be a probability space where Ω, F, and P denote a sample space, an event space, and a probability function of an event E, respectively. Then, the three probability axioms are
> 1. $P(E) \in \mathbb{R}$ and $P(E) \geq 0 \quad \forall E \in F$
> 2. $P(\Omega) = 1$
> 3. $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$ for any countable sequence of mutually exclusive event sets $\{E_i\}$.

If a function satisfies these conditions, it characterises a probability distribution. In the case of the softmax function used for C labels of classification, $\sigma : \mathbb{R}^C \to (0,1]^C$, we assume

that input is classified into only one of C labels (mutually exclusive), thus:

$$P(E_i \cap E_j) = 0, \text{ for all } i \neq j \text{ where } i, j \in \{1, ..., C\} \tag{4.4}$$

where E_i is an event that the label i is observed. [Bishop, 1995] clarified that the softmax function satisfies the probability axioms 1 and 2. With the assumption of mutual exclusivity for the events that each label is observed, the probability axiom 3 is also obviously satisfied by the softmax function, but we include the proof for clarity in the Supplementary. [Bridle, 1989] also explained that applying the softmax function to the output of a neural network makes the final result all positive and sum to 1 for any input, thus the final result can be interpreted as a probability distribution.

Therefore, for the MethylBERT network designed for tumour (T) and normal (N) cell-type outputs, the output of the network gives an estimate of the probability of cell type c given read r:

$$P(c|r, \Theta) = \sigma_c(f_\Theta(r)) = \frac{e^{\mathbf{z}_c}}{\sum_{c' \in \{T,N\}} e^{\mathbf{z}_{c'}}}. \tag{4.5}$$

Although the input of MethylBERT technically consists of three pieces of information (DNA sequence, read-level methylation patterns, and DMR label for every read), here we regard the input as one random variable, read r by assuming that methylation patterns depend on DMR label and DNA sequence. It is a broadly applied design to have one random variable representing read in sequencing-based genomic data analysis models. For example, [Wu et al., 2017] also regard sequencing reads as a random variable and use a binomial distribution as the best-fitting model of sequencing reads.

Statement 2. Even if statement 1 is true, it is important to obtain a well-approximated posterior probability P(cell type|read) from the MethylBERT network to use the output for the MLE-based tumour purity estimation. [Richard and Lippmann, 1991] showed that a neural network model trained for classification using either the squared error or the cross-entropy loss function can estimate a Bayesian posterior probability, via simulated experiments. [Saerens et al., 2002] also theoretically proved that the outputs of a classification neural network model trained with a reasonable loss function can be always mapped to a Bayesian posterior probability. Both [Richard and Lippmann, 1991] and [Saerens et al., 2002] clarified that the following conditions need to be satisfied for the output of a classification model to be an approximation of the posterior probability:

- The model complexity of the neural network is sufficiently high for the given data and the classification task.
- [Richard and Lippmann, 1991]: Sufficient amount of training data needs to be given.
- [Saerens et al., 2002]: At least, a local minimum of the loss function is achieved after

the training.

We train the MethylBERT network until the loss value converges by minimising the cross-entropy function. We assume that the converged loss value is the local minimum. During the MethylBERT fine-tuning, both training and validation loss curves decrease and we use the model at the lowest validation loss. Therefore, we assume that the MethylBERT model has a sufficient model complexity not causing an underfitting. For the training data, we use all available sequencing reads which makes an average read coverage of 1,241. In Section 5.3.2, we will show that the MethylBERT read classification accuracy converges at the best performance for read coverage ≥ 110 in the training data set. Thus, we assume that we have a sufficient amount of training data for the MethylBERT fine-tuning. Consequently, the MethylBERT model can be regarded as satisfying these conditions. We consider the calculated softmax value after the MethylBERT fine-tuning to be a well-estimated posterior probability of cell types, $P(\text{cell type}|\text{read})$.

4.3 Tumour purity estimation

4.3.1 Maximum likelihood estimation

The third step of MethylBERT is tumour purity estimation based on the calculated posterior cell-type probability. Maximum likelihood estimation (MLE) is chosen as a method. The likelihood function has the tumour purity of bulk as a parameter.

Let a bulk sample, whose tumour purity is supposed to be inferred, be the test data. Then, the tumour purity of the bulk sample is denoted as $\delta \coloneqq P_{test}(c = Tumour)$. We clarify that the tumour purity $P_{test}(c = Tumour)$ must be distinguished from the proportion of tumour reads in the training data set used for fine-tuning MethylBERT, which is denoted as $P_{train}(c = Tumour)$ here. $\hat{\delta}$, which is the optimal estimation of δ, is inferred by MLE using the likelihood function of δ, $\mathcal{L}(\delta|\mathbf{R})$, when reads $\mathbf{R} = \{r_1, ..., r_N\}$ are observed in the test data. Assuming the reads are independently sampled and the read-level methylomes only depend on the cell type c, the likelihood function $\mathcal{L}(\delta|\mathbf{R})$ is calculated as:

$$\begin{aligned}
\mathcal{L}(\delta|\mathbf{R}) &= P_\delta(\mathbf{R}) && (4.6)\\
&= \prod_{i=1}^{N} P_\delta(r_i) && (4.7)\\
&= \prod_{i=1}^{N} \sum_{c\in\{Tumour,Normal\}} P(r_i|c)P_{test}(c) && (4.8)\\
&= \prod_{i=1}^{N} \{\delta P(r_i|c = Tumour) + (1-\delta)P(r_i|c = Normal)\}, && (4.9)
\end{aligned}$$

where $P_{test}(c = Normal) = 1 - P_{test}(c = Tumour) = 1 - \delta$.

However, in the previous (second) step, only $P(c|r_i)$ is calculated by the MethylBERT network (Section 4.2.2). Therefore, Bayes' theorem is applied to re-express $P(r_i|c)$ with $P(c|r_i)$, assuming that the tumour and normal sequencing reads in the bulk (test data) and in the training data set originated from the same domain.

$$\hat{\delta} = \arg\max_{\delta} \mathcal{L}(\delta|\mathbf{R}) \tag{4.10}$$

$$= \arg\max_{\delta} \prod_{i=1}^{N} \{\delta P(r_i|c = Tumour) + (1-\delta)P(r_i|c = Normal)\} \tag{4.11}$$

$$= \arg\max_{\delta} \prod_{i=1}^{N} \{\delta \frac{P(c = Tumour|r_i)P(r_i)}{P_{train}(c = Tumour)} + (1-\delta)\frac{P(c = Normal|r_i)P(r_i)}{P_{train}(c = Normal)}\}. \tag{4.12}$$

The prior probability of cell types, $P_{train}(c = Tumour)$ and $P_{train}(c = Normal)$, are calculated by taking the ratio of reads from each cell type over the number of all reads in the training data set. Here, we again note that $\delta = P_{test}(c = Tumour)$, the estimated tumour purity of bulk, is not the same as $P_{train}(c = Tumour)$, the ratio of tumour reads in the training data. For the calculation, we assume the probability of reads $P(r_i)$ are equal. A grid-search algorithm is used to infer the optimal tumour purity $\hat{\delta}$.

4.3.2 Fisher information

The Fisher information is the information about a model parameter carried by observations. The Fisher information is calculated to estimate the precision of the model [Fujita et al., 2022]. When the model has only one parameter, the Fisher information is one value whereas, for a model with K parameters, the Fisher information is a $K \times K$ matrix which is also known as the Fisher information matrix. In the case of a likelihood estimation model, the variance of the derivative of the log-likelihood function with respect to the model parameter is the Fisher information. Therefore, the Fisher information (FI) of MethylBERT tumour purity estimation likelihood calculated using Equation (4.6) is calculated as:

$$FI(\delta) = Var[\frac{\partial}{\partial\delta} log\mathcal{L}(\delta; \mathbf{R})]. \tag{4.13}$$

$FI(\delta)$ indicates the model precision of MethylBERT tumour purity estimation which can be used as a comparison metric when multiple bulk samples are deconvoluted.

4.3.3 Adjustment of estimated tumour purity

Analysing pseudo-bulk samples used in the evaluation of MethylBERT (Section 5.4), we found that the median of region-wise tumour purity values is farther away from the ground-truth tumour purity when the ground-truth tumour purity of bulk is extremely low or high (Figure 4.4A). Here, we call region-wise tumour purity and bulk tumour purity *local* and *global* purity, respectively. The MLE algorithm for estimating tumour purity in Section

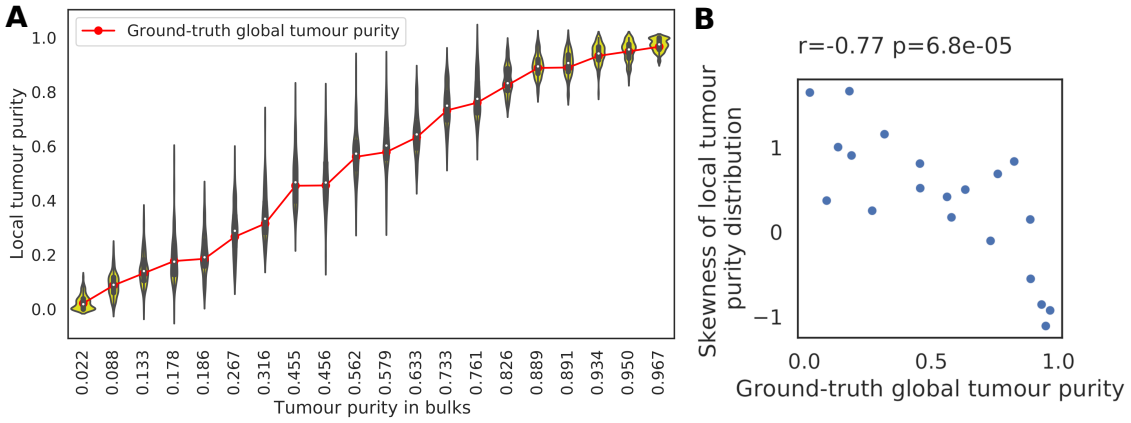


Figure 4.4: Skewness of region-wise tumour purity distribution. The local purity value was calculated for 100 DMRs used in the evaluation of MethylBERT (Section 5.4), and the skewness value was calculated over the local purity values using Equation (4.14). **(A)** Distribution of region-wise tumour purity ordered by ground-truth tumour purity within pseudo-bulk samples. Ground-truth tumour purity is presented by the red line plot. **(B)** Correlation between ground-truth tumour purity and skewness of region-wise tumour purity distribution.

4.3.1 is, however, established with the assumption that all regions have the same distribution of cell types as the global cell-type distribution, therefore the global estimation does not take the local purity into account. As a result, it disregards the case that some regions do not have sufficient sequencing reads derived from the minor cell type to estimate the global purity. Figure 4.4B confirms this by showing that ground-truth global tumour purity is negatively correlated with the skewness of local tumour purities in the pseudo-bulk data set. In other words, local purity distribution is left-skewed, having most purity values around zero when the ground-truth global purity is very low, and vice versa. We calculate the skewness of given N values $\mathbf{x} = \{x_1, ..., x_N\}$ using the adjusted Fisher-Pearson standardised moment:

$$G_1(\mathbf{x}) = \frac{m_3(\mathbf{x})\sqrt{N(N-1)}}{m_2(\mathbf{x})^{3/2}(N-2)}, \quad m_i(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^i \tag{4.14}$$

where $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, the sample mean of given values. The adjusted Fisher-Pearson standardised moment can adjust the bias in the sample distribution when calculating the skewness [Joanes and Gill, 1998, Doane and Seward, 2011].

In order to alleviate the problem that the local tumour purity distribution can be skewed, we applied the Expectation-Maximisation (EM) algorithm to adjust the estimated global tumour purity taking local purity estimates into account (Algorithm 2 and Figure 4.5). The idea is to optimise a mapping $g : \mathbb{R}^K \rightarrow \mathbb{R}^K$ to minimise the skewness of K local

purities $\boldsymbol{\delta} = \{\delta_1, ..., \delta_K\} \in \mathbb{R}^K$. We define the mapping g as:

$$g(\boldsymbol{\delta}) = \mathbf{W} \circ \boldsymbol{\delta} = \begin{pmatrix} w_1\delta_1 \\ \vdots \\ w_K\delta_K \end{pmatrix} \tag{4.15}$$

where $\mathbf{W} = \{w_1, ..., w_K\} \in \mathbb{R}^K$ contains parameters of the mapping g and $\mathbf{W} \circ \boldsymbol{\delta}$ means element-wise multiplication of two vectors $\mathbf{W}$ and $\boldsymbol{\delta}$. Since a perfectly symmetric distribution has a skewness value of zero, we assume that the mapped local purities have a symmetric distribution whose mean value is the global purity. During the M-step of the EM algorithm, the parameters $\mathbf{W}$ are optimised to minimise $G_1(\mathbf{W} \circ \boldsymbol{\delta})$, which is the skewness value of mapped local purities. The E-step estimates the new global tumour purity denoted as δ_{global} in Algorithm 2. The optimised function parameters $\hat{\mathbf{W}}$ yielded by the EM algorithm are used for determining the final adjusted tumour purity $\hat{\delta}_{global}$:

$$\hat{\delta}_{global} = \frac{1}{K}\hat{\mathbf{W}}^\intercal \boldsymbol{\delta} \tag{4.16}$$

where $\hat{\mathbf{W}}^\intercal$ means the transpose of $\hat{\mathbf{W}}$. The effect of the adjustment will be analysed in Section 5.6.

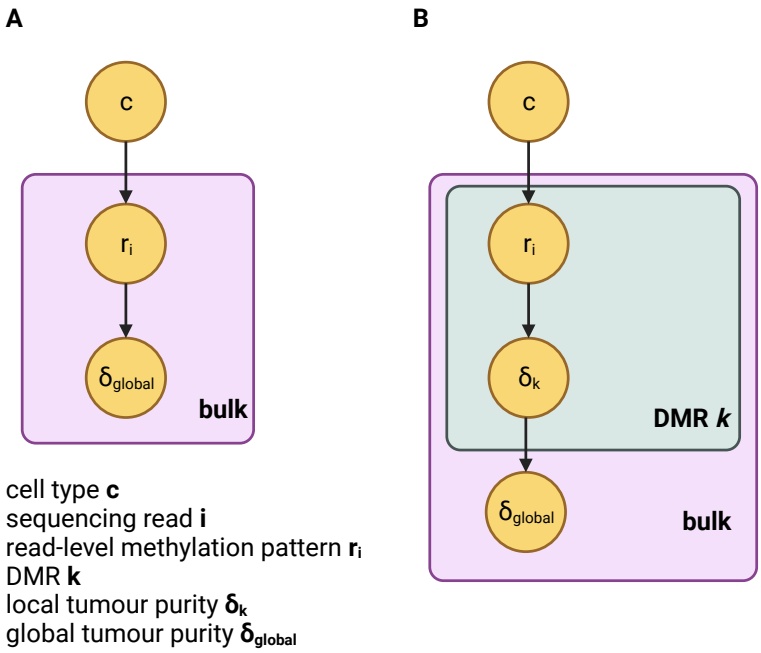


Figure 4.5: Probabilistic graphical model of tumour purity estimation in MethylBERT. **(A)** Original global tumour purity estimation. **(B)** Adjusted estimation considering local tumour purity.

Algorithm 2 Adjustment of MethylBERT tumour purity estimation

---

Input: $\mathbf{P}^t = \{p_1^t ... p_R^t\}$, $\mathbf{P}^n = \{p_1^n ... p_R^n\}$ Calculated $P(\text{read}|\text{cell type})$ for all R reads
$\boldsymbol{\mathcal{M}} = \{m_1 ... m_K\}$ K DMRs
θ Threshold for EM algorithm iteration

Output: δ_{global} Estimated tumour purity

for $m_k \in \boldsymbol{\mathcal{M}}$ **do** ▷ Estimate region-wise tumour purity
  $\delta_k \leftarrow \arg\max_{\delta} \sum_{r_i \in m_k} log(\delta p_{r_i}^t + (1-\delta) p_{r_i}^n)$
$\delta_{global} \leftarrow \frac{1}{K} \sum_{k=1}^{K} \delta_k$
$\delta_{prev} \leftarrow \infty$
$\mathbf{W} \leftarrow \mathbf{1}_K$ ▷ Initialise the parameters with 1^a
while $|\delta_{global} - \delta_{prev}| > \theta$ **do**
  $\delta_{prev} \leftarrow \delta_{global}$
  $\mathbf{W} \leftarrow \arg\min_{\mathbf{W}} G_1(\mathbf{W} \circ \boldsymbol{\delta})$ ▷ M-step
  $\delta_{global} \leftarrow \frac{1}{K} \mathbf{W}^\intercal \boldsymbol{\delta}$ ▷ E-step

---

[a] $\mathbf{1}_N$: Vector of N ones

4.4 MethylBERT Training

Determining suitable schemes and hyperparameters for training deep learning models is not trivial since the training performance significantly relies on those. In this section, we introduce general training schemes for MethylBERT pre-training and fine-tuning.

The large number of parameters in the BERT model accompanies long training time and consumption of large hardware memory. Therefore, we study the impact of the number of model parameters for the fine-tuning performance to decrease the parameter number in the MethylBERT model. To reduce the training time, we introduce two techniques to handle it: mixed precision and multi-GPU.

4.4.1 MethylBERT training scheme

In deep learning training, *epoch* means one cycle of the training data set and consists of multiple *steps* which indicate an iteration of a batch. Due to the large size of the training data set, the BERT training is conducted in multiple steps [Devlin et al., 2018]. We also use multiple steps instead of epochs for the MethylBERT training.

The pre-training was done for 120,000 steps which involve 10,000 warm-up steps at the beginning and 20,000 learning rate decrease steps at the end. $4e^{-4}$ was chosen for the learning rate. The batch size and the gradient accumulation steps were set to 256 and 4, respectively. For the optimisation, we used the AdamW optimiser whose weight decay rate was 0.01. β_1, and β_2 values in the AdamW optimiser [Loshchilov and Hutter, 2017] were set to 0.9, and 0.98, respectively.

The baseline BERT network has 12 hidden layers whose dimension is 768 and each hidden layer has 12 attention heads. This setup is the same as the $\text{BERT}_{\text{BASE}}$ model in the original BERT paper [Devlin et al., 2018]. Therefore, we used this setup for all experiments conducted in Chapter 5. However, because of the large model size exceeding the GPU

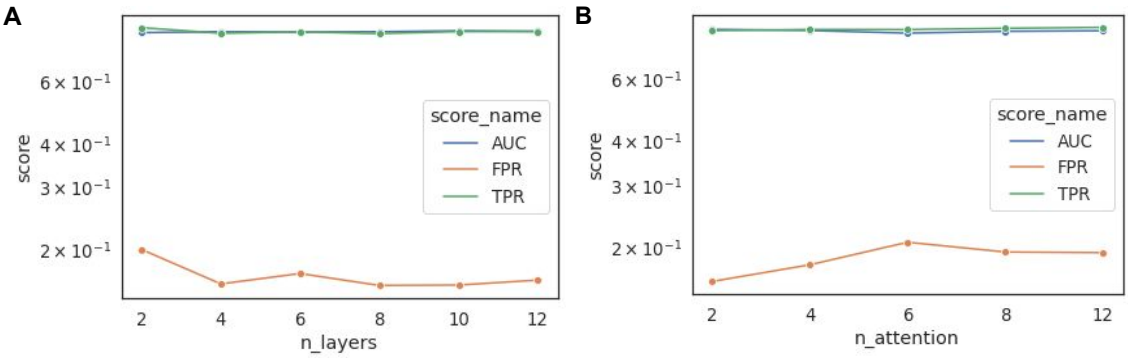


Figure 4.6: Comparison of read-level methylome classification (fine-tuning) performance for different numbers of layers (A) and different numbers of attention heads (B).

memory, 6 layers instead of 12 layers were assigned for the MethylBERT model used in the simulated long-read data analysis in Section 5.3. We confirmed that the smaller number of layers does not make a critical performance loss (Figure 4.6).

4.4.2 Mixed precision

In the classical way of neural networking training, all variables are 32-bit floating-point type (float32). Mixed precision replaces some of the variables with 16-bit floating-point type (float16) to reduce memory consumption. This is particularly beneficial for training a large neural network such as BERT which requires a large amount of memory. Moreover, many types of processors including NVIDIA GPUs run math operations faster for float16 than for float32, thus using mixed precision also improves computational speed. In the mixed precision process, float16 is specifically used for the variables for the evaluation metrics such as accuracy, since the coarse approximation has a lesser impact on model evaluation compared to model parameters. In MethylBERT, mixed precision is applied via the *autocast* function provided by PyTorch [Paszke et al., 2019].

4.4.3 Multi-GPU

MethylBERT supports multi-GPU processing by distributing an input batch over multiple GPUs. This is implemented by the *DataParallel* function in PyTorch [Paszke et al., 2019]. The input reads in a batch are split into a given number of GPUs and each GPU processes a split via a replicated network on the device. For backpropagation, the model parameters need to be updated with respect to the gradient of loss values in a batch, thus the gradients from all devices are summed up.

Chapter 5

Experimental results using MethylBERT for cell-type deconvolution

5.1 Introduction

Tumour cells present distinctive methylation patterns at CpG sites [Guo et al., 2019], thus DNA methylation (DNAm) data enables the estimation of tumour purity within tumour bulk samples. The estimated tumour purity can be associated with tumour diagnostics, phenotypes, and clinical outcomes [Zhao et al., 2021, Meyer et al., 2021, Zhang et al., 2017]. Therefore, sequencing-based cell-type deconvolution models have been extensively studied for tumour purity estimation. Such models are especially important in *circulating tumour DNA (ctDNA)* analysis as explained in Section 2.1.2. The ctDNA analysis often involves a low percentage of tumour-derived DNA fragments below 5% which would not be accurately identified by array-based DNAm profiling.

Considering the crucial role of sequencing-based cell-type deconvolution in tumour bulk and ctDNA analyses, our developed method MethylBERT (see Chapter 4) needs to be thoroughly evaluated for both applications. The evaluation should be separated into *read classification* and *tumour purity estimation* to assess the performance of each step. In particular, the evaluation of the read classification step should take multiple scenarios into account because of the assorted sequencing technologies available nowadays (e.g., Illumina short-read sequencing and long-read sequencing as shown in Table 2.2) or different complexities of tumour signals in genomic regions. In addition, it is also necessary to study the performance of the tumour purity adjustment algorithm in the MethylBERT model by comparing it with other methods.

Despite previous applications of the pre-trained BERT model to genomic sequence analysis [Ji et al., 2021, Gwak and Rho, 2022], the efficacy of pre-training on DNA sequences has not been well-explored. On the other hand, when it comes to natural language processing (NLP), several studies have been published for the analysis of the learnt linguistic features by the pre-trained BERT model. [Clark et al., 2019] comprehensively investigated the linguistic knowledge learnt by the pre-trained BERT model and explained why BERT can accomplish outstanding performance in human language-related tasks. [Jawahar et al., 2019] conducted several analyses on the language structure learnt by the pre-trained BERT model and proved its capability of recognising the hierarchical structure of human language. These studies provided an in-depth understanding of BERT pre-training in NLP, whereas such analyses have not yet been done on BERT pre-training with DNA sequences.

In addition, the assessment of MethylBERT with patient samples is crucial to study its applicability in clinical applications. Yet, in biological samples, the ground-truth tumour purity is not given which makes the evaluation difficult. For cancer patient samples, the cancer stage can be a reasonable metric for evaluating tumour purity estimation. The size of tumours is one of the major standards to determine the cancer stage [Brierley et al., 2016]. Therefore, it can be expected that a tumour sample acquired from a later-stage cancer patient involves a higher purity of tumour.

In this chapter, we present experimental results using MethylBERT to comprehensively evaluate our method and compare it with previous methods. Furthermore, we study the efficacy of BERT pre-training on DNA sequences as well as the applicability of MethylBERT in ctDNA analysis using blood samples from cancer patients. First, we describe previous methods included in the evaluation (Section 5.2). Then, in Section 5.3, we compare MethylBERT to other previous sequencing-based DNAm deconvolution methods with respect to read classification. For the comparison, simulated read-level methylomes are used, so the simulation algorithm will be also described. In Section 5.4, we compare MethylBERT to the previous methods in terms of tumour purity estimation using pseudo-bulk samples. Section 5.5 provides experimental results to show why pre-training is important in MethylBERT, although the pre-training does not involve any methylation information. The adjustment of tumour purity estimation is evaluated in Section 5.6. Finally, the cell-type deconvolution results on blood plasma samples are presented in Section 5.7, demonstrating the potential of MethylBERT as an early tumour diagnosis tool. This work was published in [Jeong et al., 2023a].

5.2 Previous methods for experimental comparison

We selected three previous methods for the comparison with MethylBERT: Houseman's method [Houseman et al., 2012], CancerDetector [Li et al., 2018], and DISMIR [Li et al., 2021]. Houseman's method is chosen as the best-performing method in the benchmarking described in Chapter 3. CancerDetector and DISMIR were developed for sequencing-based

cell-type deconvolution targeting tumour samples by using the beta-binomial model and RNNs, respectively. Hence, this evaluation can be seen as a comparison of the Transformer-based method (MethylBERT) to array-based (Houseman's method), beta-binomial-based (CancerDetecotr), and RNN-based (DISMIR) cell-type deconvolution methods. Only for the read classification performance evaluation in Section 5.3, we added a hidden Markov model (HMM) as an alternative to Houseman's method which does not have a read classification step.

Houseman's method

As explained in Section 3.4.2, Houseman's method is based on regression calibration for array-based cell-type deconvolution. Therefore, as in the benchmarking study, we converted the read-level DNAm data into an array shape to test Houseman's method following Section 3.3.2.

Hidden Markov Model

We implemented an HMM to compare MethylBERT in the read classification task. The model inputs are CpG-wise methylation patterns which have binary categories: methylated and unmethylated. For the hidden state, we assumed a variable with two categories implying differentially methylated and non-differentially methylated CpGs between cell types.

CancerDetector

In the CancerDetector algorithm, the methylation alpha value is introduced as an average of methylation levels in individual reads. In a differentially methylated region (DMR) k, CancerDetector yields the probability of each read r given cell type c :

$$P(r|c,k) = P(r|B_{k,c}(\alpha,\beta)) \tag{5.1}$$

where $B(\alpha,\beta)$ represents the beta distribution parameterised by α and β. The distribution $B_{k,c}(\alpha,\beta)$ is modelled by cell type c-derived reads extracted from the DMR k in the training data using the method of momentum [Bowman and Shenton, 2007].

DISMIR

DISMIR uses one-dimensional convolutional neural networks and a bi-directional long short-term memory (LSTM) to model read-level DNAm patterns (Figure 5.1). Unlike MethylBERT encoding which the DNA sequences and the methylation patterns are embedded and encoded separately, DISMIR encodes those together using one-hot encoding. The final output, which is named 'd-score', is calculated using the sigmoid function and interpreted as the probability that the given read is derived from the cell type.

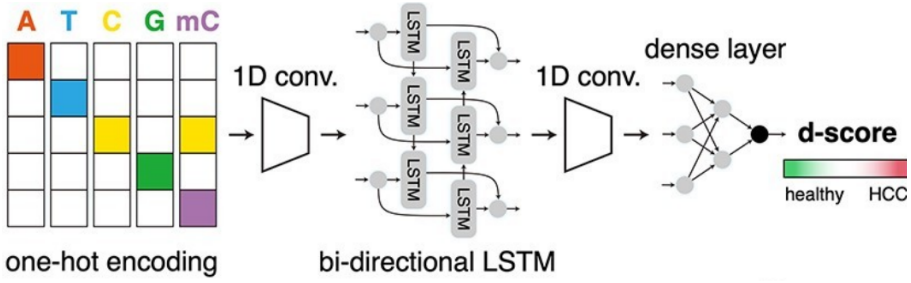


Figure 5.1: Overview of DISMIR (Figure modified from [Li et al., 2021]).

5.2.1 Methodological comparison to MethylBERT

Both CancerDetector and DISMIR employ maximum likelihood estimation (MLE) for estimating tumour purity by using the calculated probabilities or the d-score. Technically, CancerDetector and DISMIR do not mention the 'read classification' step in their model, yet we use the computed $P(\text{read}|\text{cell type} = \text{Tumour})$ probabilities from the CancerDetector algorithm and the d-score from the DISMIR algorithm for the read classification evaluation. When the probability or d-score is above 0.5, reads are classified as tumour-derived reads.

Regarding the adjustment of estimated tumour purity, only CancerDetector has a similar step, named 'removal of confounding factor'. In the removal of the confounding factor algorithm, the Expectation-Maximisation (EM) algorithm is applied to iteratively remove confounding factors. More details about the algorithm are explained in Section 5.6.

5.3 Read classification performance evaluation

In this section, we compare the performance of different read classification models using simulated read-level methylomes. Since sequencing reads from real tumour bulks are intricate to analyse due to the absence of ground-truth labels for individual reads, synthesised methylation patterns with controlled parameters are used.

5.3.1 Read-level methylation pattern simulation

The complexity of methylation patterns is known to be associated with the regulation of oncogenes and tumour suppressor genes which result in tumour heterogeneity [Wang et al., 2020, Liu et al., 2019]. Hence, it is essential to address complex methylation patterns for an accurate estimation of tumour purity.

For the read simulation, we mainly controlled the methylation pattern complexity to create varying cases of tumour-specific methylation profiles. Different complexity of region-wise tumour methylation level d_i is assumed to follow a beta distribution with a different shape

parameter α:

$$d_i \sim Beta(\alpha, \beta = 5). \tag{5.2}$$

To ensure adequate methylation signals, the top 100 CGIs with the largest number of CpGs are chosen to simulate reads. Only α value changes to simulate different distributions, but the other shape parameter β is fixed as five. For the normal cell type, we assign $1 - d_i$ as the region-wise methylation level.

Read-level methylation patterns for each cell type are sampled from a binomial distribution whose success probability is the sampled methylation level. In a region i containing K CpGs, read-level methylation patterns for tumour cell type $\mathbf{m}_i^T = \{m_{i,1}^T, ..., m_{i,K}^T\}$ are sampled as follows:

$$\mathbf{m}_i^T \sim Binomial(n = K, p = d_i), \tag{5.3}$$

where n and p are the number of trials and the probability of success of the binomial distribution. Read-level methylation patterns for normal cell type $\mathbf{m}_i^N = \{m_{i,1}^N, ..., m_{i,K}^N\}$ are also sampled in the same manner:

$$\mathbf{m}_i^N \sim Binomial(n = K, p = 1 - d_i). \tag{5.4}$$

The beta distribution is used for the simulation of different methylation pattern complexities by applying four different α values: 0.1, 1.0, 2.0 and 3.0. The simulated data set for each complexity is named a0_b5, a1_b5, a2_b5 and a3_b5. Figure 5.2 presents the sampled region-wise tumour methylation level using Equation (5.2) and Figure 5.3A shows the example of simulated reads from respective distributions. A higher value of α creates a higher mean value in the beta distribution. With a higher mean value, a higher tumour methylation level d_i is sampled more often and this makes a lower methylation level for normal cell type, $1 - d_i$. When the difference between d_i and $1 - d_i$ is smaller, the DNAm

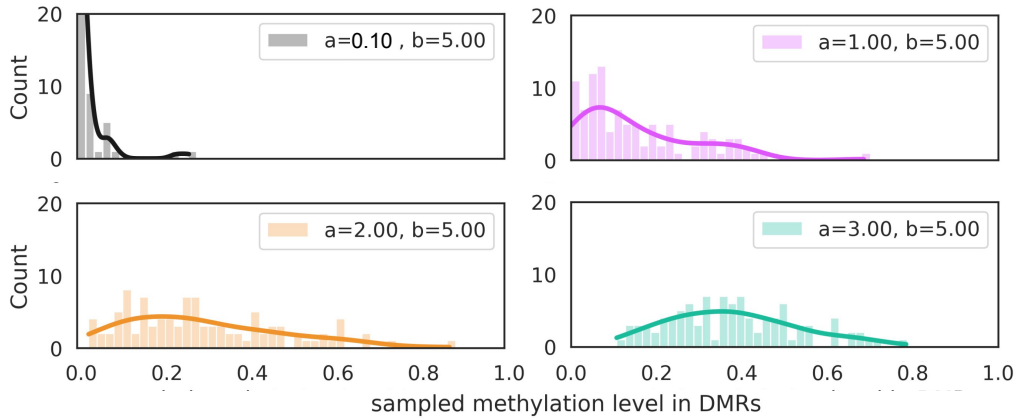


Figure 5.2: Distribution of sampled tumour mean methylation level for DMRs using beta distribution in Equation (5.2) with four α values, 0.1, 1, 2 and 3.

patterns generally get more complex to distinguish between two cell types.

The four simulations are done in two different read lengths: 150 bps and 500 bps. 150bp read length represents standard bisulfite sequencing data, whereas 500bp read length has been simulated to test the applicability of read classification models to long-read sequencing data, such as nanopore sequencing. The main challenge in long-read sequencing data is that DMRs are generally shorter than the sequencing read (Figure 5.3B). Therefore, the reads include CpGs not showing cell type-specific methylation patterns.

In biology, it is well understood that neighbouring CpGs tend to have the same methylation patterns [Affinito et al., 2020]. However, this is not always the case. There have been many studies reporting highly stochastic DNAm patterns, especially in cancer [Videtic Paska and Hudler, 2015, Saghafinia et al., 2018]. Furthermore, allele-specific and strand-specific methylation (ASM and SSM) diversify methylation patterns at CpGs, even in the same tissue- or cell-type samples. These are also considered important indicators for examining cell development and disease [Sen et al., 2021, Do et al., 2020]. Thus, we conducted a simulation of another read-level methylome data set by giving two different cell type-specific methylation patterns between odd and even indices of CpGs (Figure 5.3C). This data set is called *CpG-specific methylation* data set in the following sections. We note that region-wise DNAm level between two cell types does not have a large difference in the CpG-specific methylation data set.

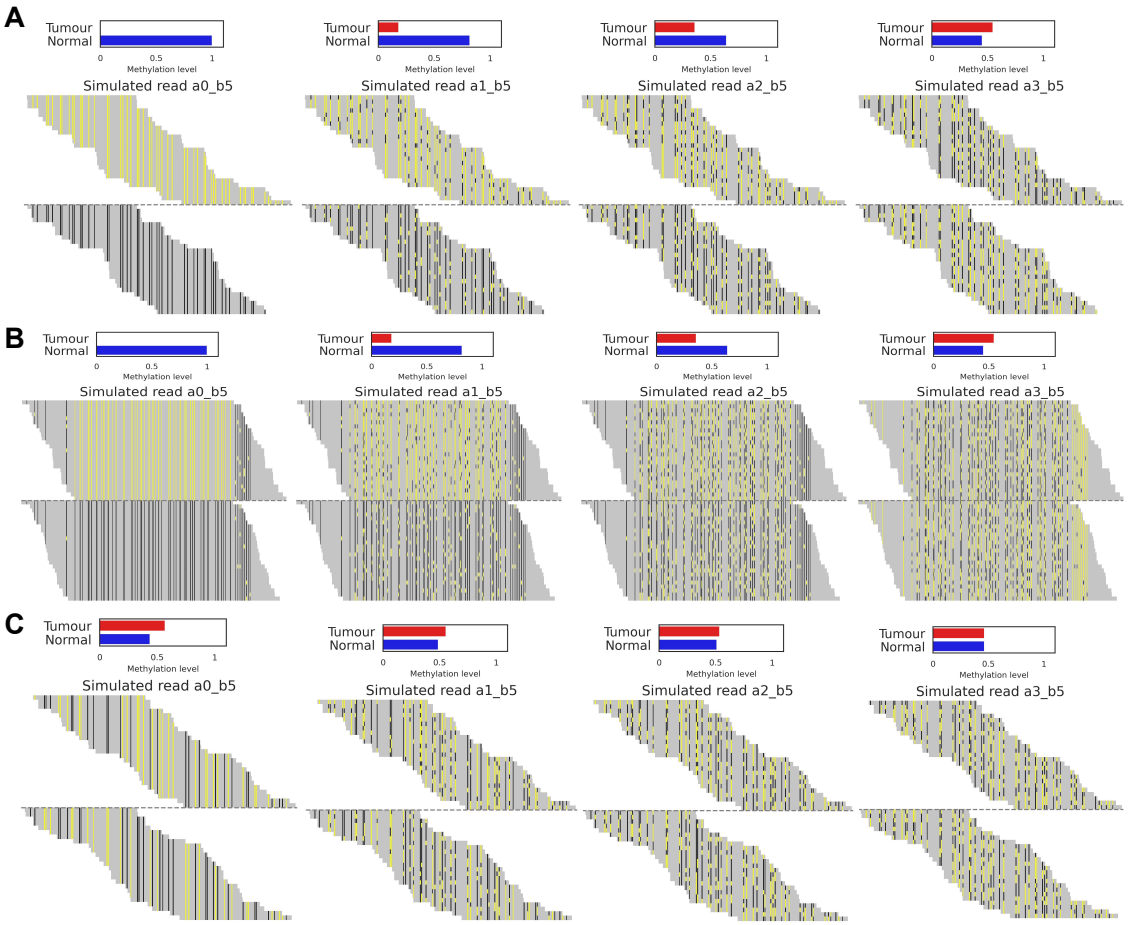


Figure 5.3: Example of simulated reads from the four different beta-binomial distributions. **(A)** Example of simulated 150bps reads. **(B)** Example of simulated reads with 500 bp read length. **(C)** Example of simulated reads with CpG-specific methylation patterns. Grey horizontal lines are sequencing reads. Yellow and black on each read mean methylated and unmethylated CpGs. Reads for two cell types are divided by a dotted line in the middle. Region-wise methylation levels in tumour and normal cell types are shown in the histogram.

5.3.2 Read classification performance comparison with simulated read-level methylomes

With the simulated data, we evaluate the performance and robustness of the MethylBERT read classification model in different scenarios of methylation patterns. We use region-wise classification accuracy as a performance metric by calculating the proportion of correctly classified reads over all reads in each DMR.

For the 150 bp read simulation, MethylBERT outperforms other methods regardless of the complexity (Figure 5.4A). All the methods classify the reads most accurately with the lowest deviation for the simplest case of simulation (a0_b5). Even though the accuracy decreases with higher complexities, MethylBERT still achieves the highest accuracy in each case. When it comes to the methods not based on deep learning, CancerDetector performs better than HMM for the simulated reads with lower complexities, a0_b5 and a1_b5, but HMM outperforms CancerDetector in the case of more complex simulations, a2_b5 and a3_b5.

Compared to the other methods, MethylBERT again performs best in every complexity of 500 bp read simulation (Figure 5.4B). In this simulation, reads are generally longer than the DMR where they are located, so most reads have both differentially and non-differentially methylated patterns (Figure 5.3B). Deep learning-based models, MethylBERT and DISMIR, show higher accuracy and lower deviation of accuracy values at the a0_b5 case. This tendency is similarly shown in the 150 bp read classification result (Figure 5.4A). However, CancerDetector and HMM yield a higher deviation of accuracy even at the lowest complexity, compared to the same complexity level in 150bp read classification results. Overall, MethylBERT and DISMIR performed better than others in the evaluation using the simulated 500 bp reads.

In the evaluation using the CpG-specific methylation data set, CancerDetector and HMM cannot correctly classify the reads for a0_b5 and a1_b5 despite a relatively clear difference in methylation patterns between two cell types (Figure 5.4C). On the other hand, deep learning-based methods still show high accuracy of read classification. In particular, MethylBERT outperforms DISMIR with the exception of the highly noisy methylation patterns, a2_b5 and a3_b5. These results imply that Methylation can handle CpG-specific methylation patterns not being biased towards the region-wise average methylation level of cell types. This is important for analysing tumour bulk samples because, in tumour cells, CpG-specific methylation patterns which differ from neighbouring CpGs are more commonly found [Videtic Paska and Hudler, 2015].

In read-level methylome analysis, read coverage is always regarded as important for reliable analysis. As described in Section 3.2, most sequencing-based deconvolution algorithms have their own informative region selection procedure taking the minimum number of CpGs and the read coverage into account. In this way, a sufficient amount of methylation

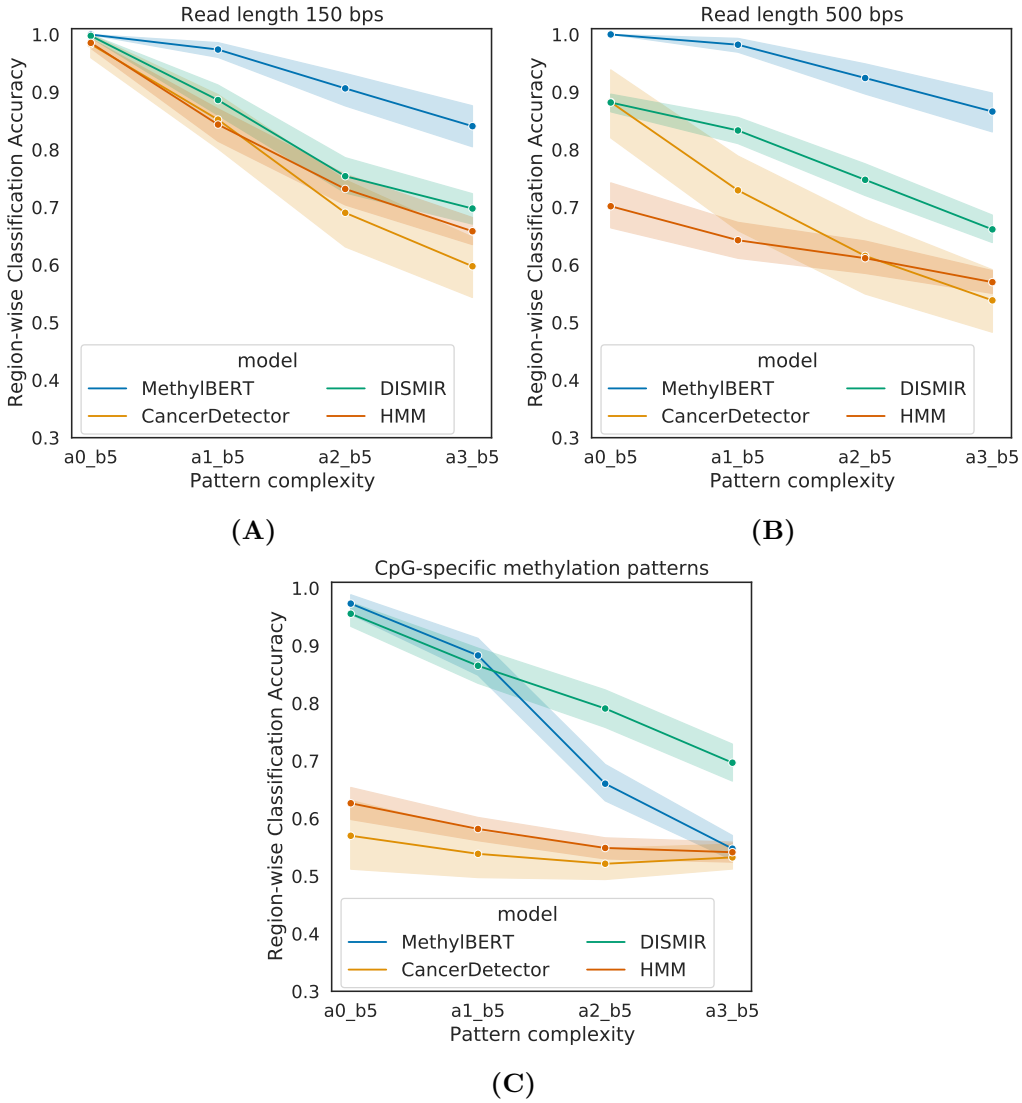


Figure 5.4: Read classification performance comparison. **(A)** Performance for simulated reads with 150 bp length. **(B)** Performance for simulated reads with 500 bp length. **(C)** Performance for simulated reads for the CpG-specific methylation data set.

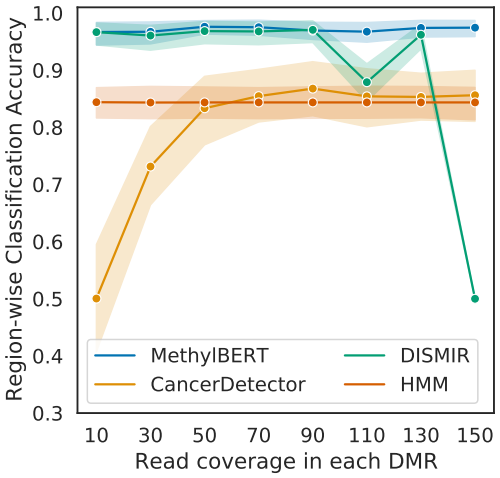


Figure 5.5: MethylBERT performance in different read coverages. Performance comparison with other read classification methods based on read classification accuracy.

patterns is ensured for deconvolution analysis.

For this reason, we compare MethylBERT to the other methods using another set of simulated read-level methylomes with different read coverage values from 10 to 150 in the training data set (Figure 5.5). The read coverage of the test set is fixed as 50. MethylBERT outperforms all other methods in terms of read classification accuracy. Particularly, for read coverage values < 50, CancerDetector performed poorly compared to other methods. We describe the reason for poor performance with the standard error of the mean (SEM) which measures how much sample means represent the underlying population mean (Figure 5.6A). The SEM was estimated in each region for the mean methylation level of simulated n reads $\boldsymbol{\mu} = \{\mu_1, ...\mu_n\}$:

$$\frac{\sigma_{\boldsymbol{\mu}}}{\sqrt{n}} \tag{5.5}$$

where $\sigma_{\boldsymbol{\mu}}$ is the standard deviation of $\boldsymbol{\mu}$. In Figure 5.6A, the lower the read coverage is, the higher the SEM value is, regardless of the cell type and complexity. MethylBERT and DISMIR still achieve an accuracy value higher than 0.95 for the read coverage < 50. However, DISMIR shows a significant performance drop when the read coverage is higher than 100.

We further evaluate MethylBERT with different read coverages for every simulation of pattern complexity (Figure 5.6B). Regardless of the coverage, MethylBERT consistently achieved an accuracy higher than 0.95 in simpler cases of read simulation, a0_b5 and a1_b5. On the other hand, for the simulation of a2_b5 and a3_b5 which have more complex patterns of methylation, the accuracy value increases over the coverage and converges at the best performance with the coverage ≥ 110. Also, the deviation of region-wise accuracy is higher for a2_b5 and a3_b5 than for a0_b5 and a1_b5. From these results, we conclude

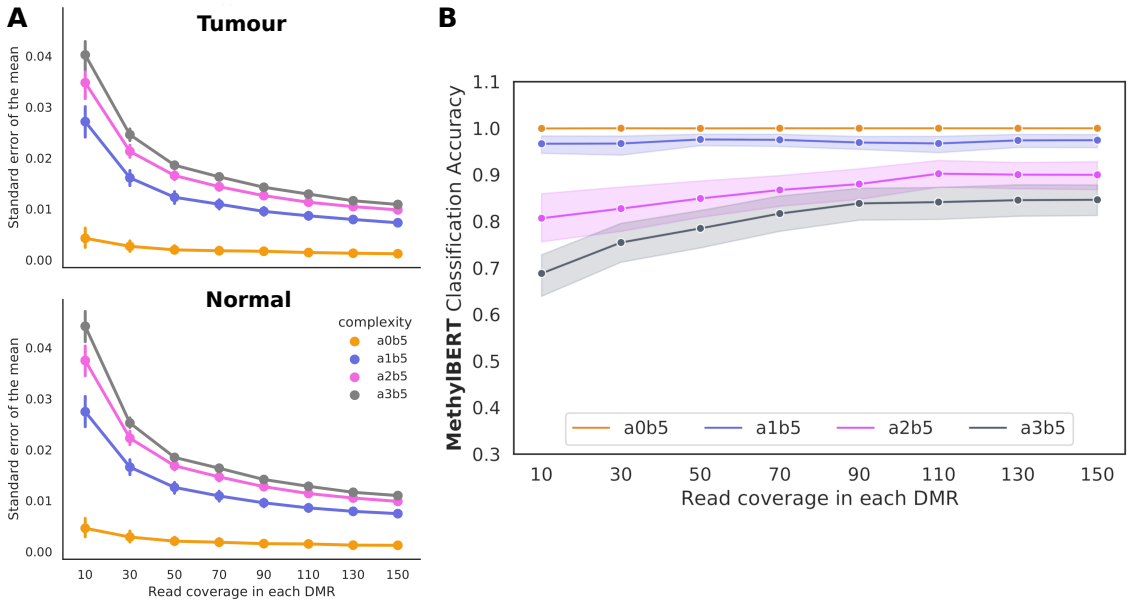


Figure 5.6: MethylBERT read classification result for different read coverages and complexities of methylation patterns. **(A)** Standard error of the mean measure in each region for different read coverages and complexities of methylation patterns. **(B)** Read classification accuracy achieved by MethylBERT for different read coverages and complexities of methylation patterns.

that when the methylation pattern is complex, the training data needs to have a relatively high coverage of reads to properly train the MethylBERT model. However, for the simpler cases of methylation patterns, the read coverage in the training data is not very impactful in determining the read classification performance.

Based on the read classification results using simulated data, we conclude that the classification performance differs depending on the methylation pattern complexity. We verify this tendency in diffuse large B-cell lymphoma (DLBCL) and normal B cell samples used in Chapter 3 (Figure 5.7).

According to the correlation between region-wise area under the curve (AUC) values calculated by MethylBERT and read coverage in DMRs, the DLBCL and normal B cell data show the same tendency that the accuracy is higher with lower coverage (Figure 5.7A). It differs from the read classification result for the simulated data shown in Figure 5.6B. For the simulated reads, the accuracy is rather positively correlated with the coverage value when the methylation patterns are complex.

In DMRs selected from biological samples, methylation pattern complexity is affected by three different factors: the number of CpGs within the region, region length and methylation level difference. The larger the region length and the number of CpGs are, the higher the methylation pattern complexity is. On the contrary, a lower methylation

level difference makes a higher complexity of methylation patterns. For quantifying the methylation level difference between normal and tumour cells, *diff.Methy* score is used. As described in Section 2.1.4, diff.Methy score is calculated as the distance between the mean methylation level of two cell types. In the analysis, the absolute value of diff.Methy score is used so we disregard the directionality in the score. Originally, the positive and negative values of diff.Methy indicates which cell type of two is hypomethylated.

The AUC value of read classification shows a negative correlation with the number of CpGs and region length, but a positive correlation with absolute *diff.Methy* (Figure 5.7B-D). This corresponds to the analysis done with the simulated data: read classification accuracy is lower when the methylation pattern complexity gets higher. To sum up, the results present that the high complexity of methylation patterns also impedes the read classification performance in actual tumour samples, as shown in the simulated data results.

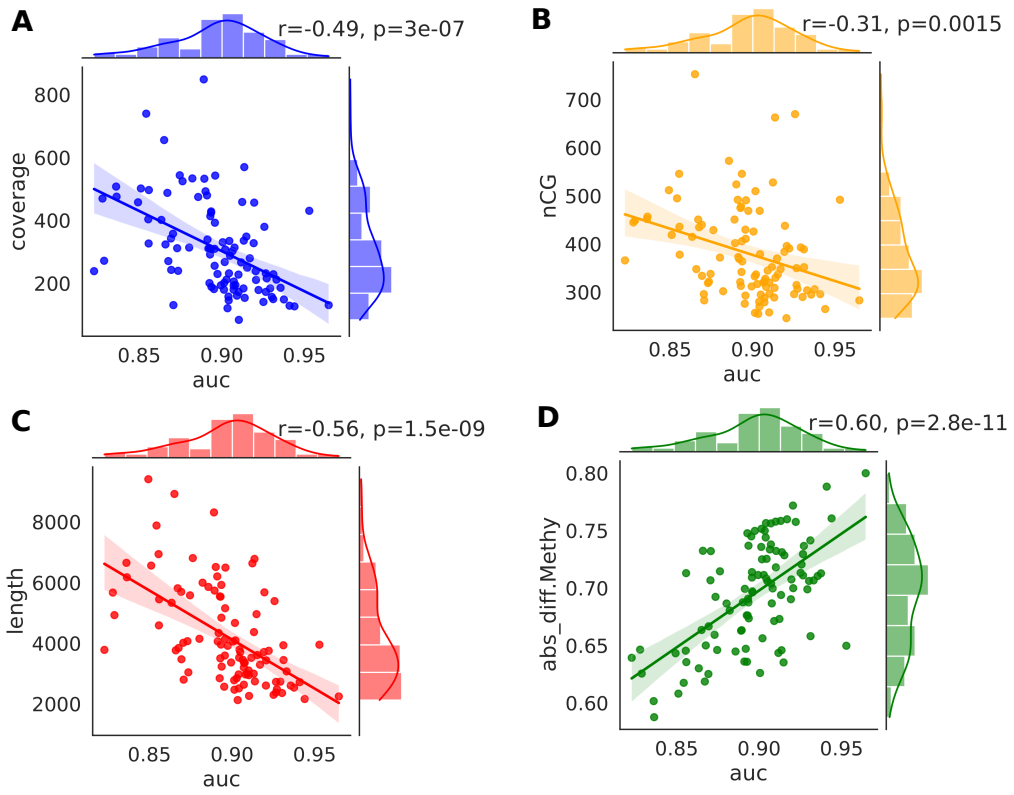


Figure 5.7: Correlation between read classification AUC and region-wise statistics: **(A)** read coverage, **(B)** number of CpGs in DMR, **(C)** length of DMR and **(D)** absolute value of diff.Methy score (described in Section 2.1.4). We note that all correlation values are statistically significant (p-value < 0.01).

5.4 Tumour purity estimation performance evaluation with pseudo-bulk samples

In this section, we describe the evaluation of the tumour purity estimation step of MethylBERT and the comparison with previous cell-type deconvolution methods mentioned in Section 5.2. For the evaluation data, pseudo-bulk samples were generated using DLBCL and normal B cell-derived methylomes in the same way as in Section 3.3.1 (Table 5.1). We only changed the distribution to simulate cell-type compositions. In Section 3.3.1, a Dirichlet distribution was used, but we used a uniform distribution whose range is $[0, 1]$ in this chapter due to the binary cell-type composition (tumour and normal cell types). In the evaluation using pseudo-bulks resembling ctDNA samples (Section 5.4.2), we used ten pseudo-bulks with a very low proportion of tumour. These ten pseudo-bulks have a ground-truth tumour purity between 0.001 and 0.01 with a difference of 0.001. For tumour purity estimation, we used the top 100 DMRs selected based on the areaStat score explained in Section 2.1.4. The areaStat score quantifies the region-wise methylation pattern differences taking the number of CpGs into account.

As a major evaluation metric, the absolute error between ground-truth and estimated tumour purity is used. Yet, in Section 5.4.2, we calculate the absolute percentage error and Spearman's correlation between ground-truth and estimated values. This is because the absolute error is not sufficient to evaluate the very low estimation values.

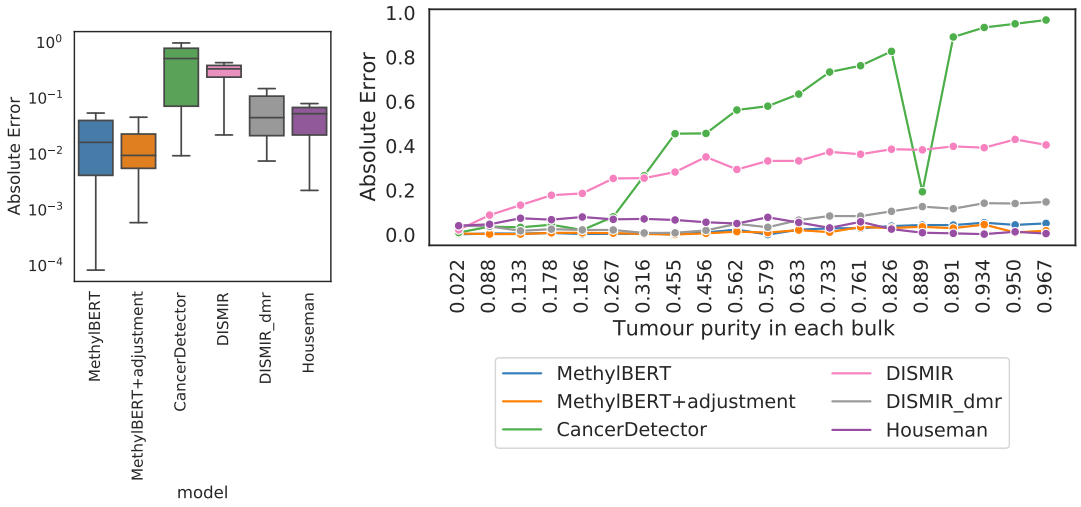
Table 5.1: Cell-type proportions for pseudo-bulk samples used in Section 5.4

	Pseudo-bulks		**Pseudo-bulks with a very low tumour proportion**	
Samples	Tumour	Normal	Tumour	Normal
Bulk 1	0.186	0.814	0.001	0.999
Bulk 2	0.826	0.174	0.002	0.998
Bulk 3	0.633	0.367	0.003	0.997
Bulk 4	0.088	0.912	0.004	0.996
Bulk 5	0.267	0.733	0.005	0.995
Bulk 6	0.562	0.438	0.006	0.994
Bulk 7	0.022	0.978	0.007	0.993
Bulk 8	0.456	0.544	0.008	0.992
Bulk 9	0.889	0.111	0.009	0.991
Bulk 10	0.455	0.545	0.01	0.99
Bulk 11	0.967	0.033	-	-
Bulk 12	0.933	0.066	-	-
Bulk 13	0.733	0.267	-	-
Bulk 14	0.178	0.822	-	-
Bulk 15	0.133	0.867	-	-
Bulk 16	0.579	0.421	-	-
Bulk 17	0.950	0.050	-	-
Bulk 18	0.891	0.109	-	-
Bulk 19	0.316	0.684	-	-
Bulk 20	0.761	0.239	-	-

5.4.1 Evaluation using tumour pseudo-bulks

As sequencing-based deconvolution models for the comparison, CancerDetector and DISMIR are included [Li et al., 2018, Li et al., 2021]. Unlike CancerDetector and MethylBERT which classify reads within pre-selected DMRs, the original model of DISMIR has an informative region selection step. However, during our experiments, DISMIR could not show reasonable deconvolution results for some pseudo-bulks with its own informative region selection algorithm. Therefore, we have included another deconvolution result yielded by DISMIR but using pre-selected DMRs as for CancerDetector and MethylBERT. These results are labelled as "DISMIR_dmr".

Figure 5.8 demonstrates that MethylBERT performs best in the tumour purity estimation for 20 pseudo-bulk samples shown in Table 5.1. Two sequencing-based deconvolution methods, CancerDetector and DISMIR, tend to have higher error values for the high proportion of tumour-derived reads. Contrarily, the array-based Houseman's method shows better performance in the bulk samples with a higher tumour purity. The performance gap between the low and high tumour purities is not as large in both MethylBERT results. Especially with the estimation adjustment, the error value decreases for the bulks whose tumour purity is higher than 0.95.



(A) **(B)**

Figure 5.8: Tumour purity estimation result for tumour pseudo-bulk samples. **(A)** The distribution of absolute error values for different methods. **(B)** Absolute error for each pseudo-bulk ordered by ground-truth tumour purity.

When investigating region-wise read classification accuracy in pseudo-bulk samples, we observe that the accuracy tends to be lower in the bulk samples with higher tumour proportion (Figure 5.9). This can be explained by two reasons: the impurity of the source of tumour data in the training set and the intra-tumour methylation heterogeneity. The

impurity is generally caused by tumour-associated stromal[1] and epithelial cells[2] commonly observed in tumour bulk samples including DLBCL [Yoshihara et al., 2013, Sangaletti et al., 2020]. Therefore, even if pure tumour bulk samples are acquired for the training set, the samples are always expected to have a very minor amount of non-tumour cells. Intra-tumour heterogeneity refers to subpopulations within a tumour presenting heterogeneity in phenotypic characteristics or molecular features including DNAm [Guo et al., 2019]. Such heterogeneity reduces the accuracy of read classification caused by outlier tumour methylation patterns. Therefore, the more tumour cells exist in a bulk sample, the lower the read classification accuracy is.

We also find a positive correlation value of 0.574 between the read classification accuracy and the tumour methylation level in selected DMRs (Figure 5.10). We explain this tendency with the example of two DMRs. As shown by the calculated mean methylation levels and the visualisation of DMR 92 and 20, the tumour methylation patterns are more inconsistent when the tumour methylation level is lower (Table 5.2, Figures 5.11 and 5.12). The partially/fully unmethylated reads which do not follow the dominant fully-methylated CpGs cause heterogeneous methylation patterns in DMR 92 and deteriorate the read classification accuracy (Figure 5.11). On the other hand, in DMR 20, there are much fewer partially/fully unmethylated reads (Figure 5.12).

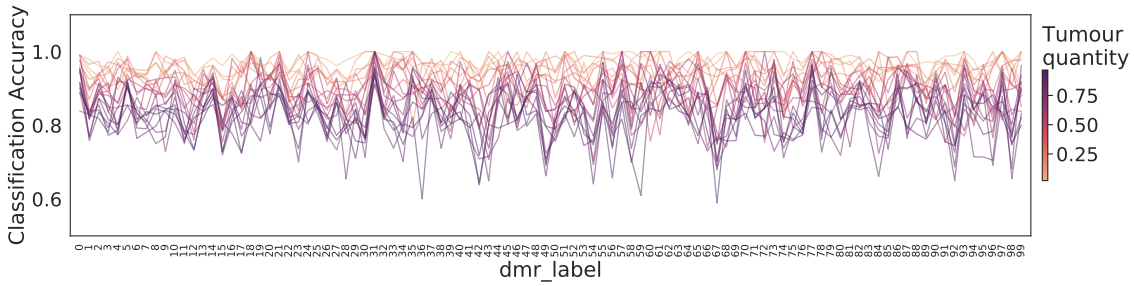


Figure 5.9: Read classification accuracy in each DMR. Individual lines show the accuracy for each bulk. A darker colour means higher ground-truth tumour purity in the bulk.

Table 5.2: Selected DMRs for the comparison

DMR ID	chr*	Start	End	Tumour methyl level	Normal methyl level	Gene	Annotation
92	chr5	32708864	32714449	0.698841	0.0527377	NPR3	Promoter
20	chr7	155249319	155253114	0.828472	0.103014	EN2	Promoter

* chromosome

[1]Stromal cells are the cells at the differentiating stage. These cells are most commonly found in bone marrow, but also observed all over the body.

[2]Epithelial cells comprise outer and inner surfaces of organs and blood vessels.

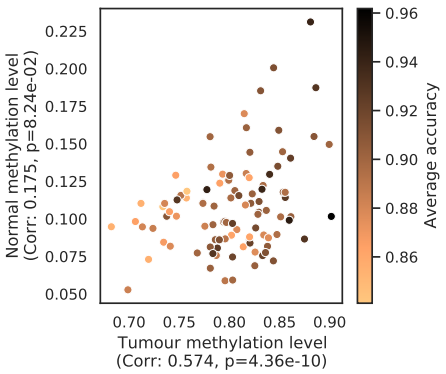


Figure 5.10: 100 DMRs plotted by tumour and normal methylation levels. The colour indicates the mean read classification accuracy in the region. The correlation value and the p-value at the x- and y-axes were calculated using the read classification accuracy and each cell-type methylation level.

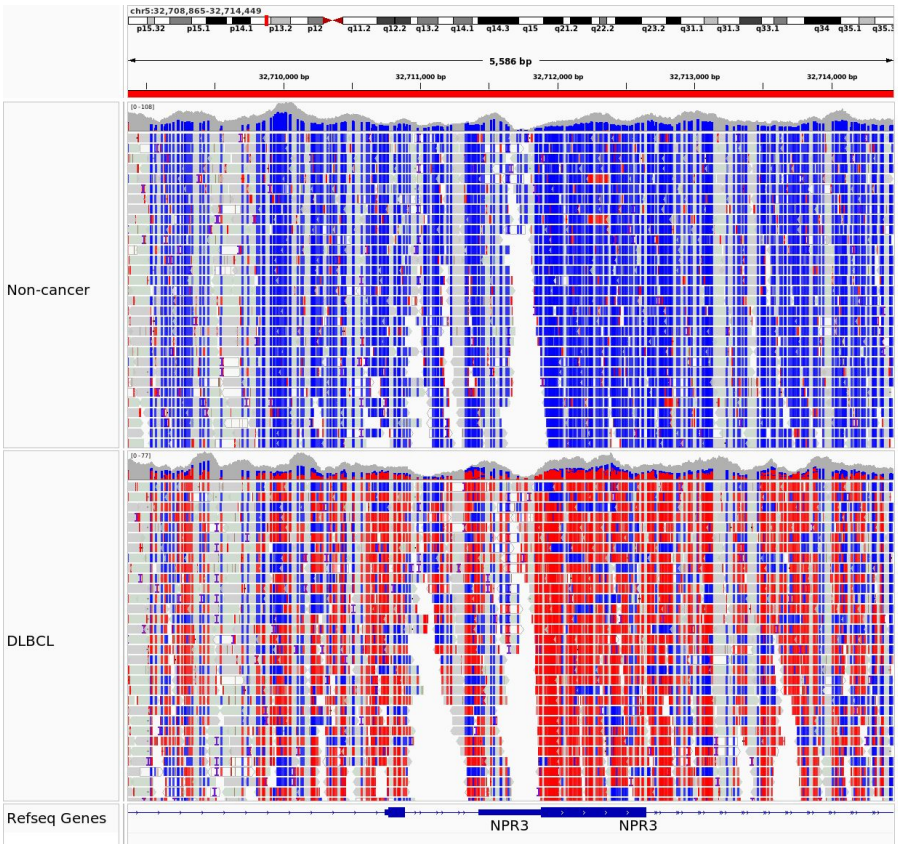


Figure 5.11: Read-level methylation patterns for normal and tumour cells in DMR 92. The blue and red mean unmethylated and methylated CpG each.

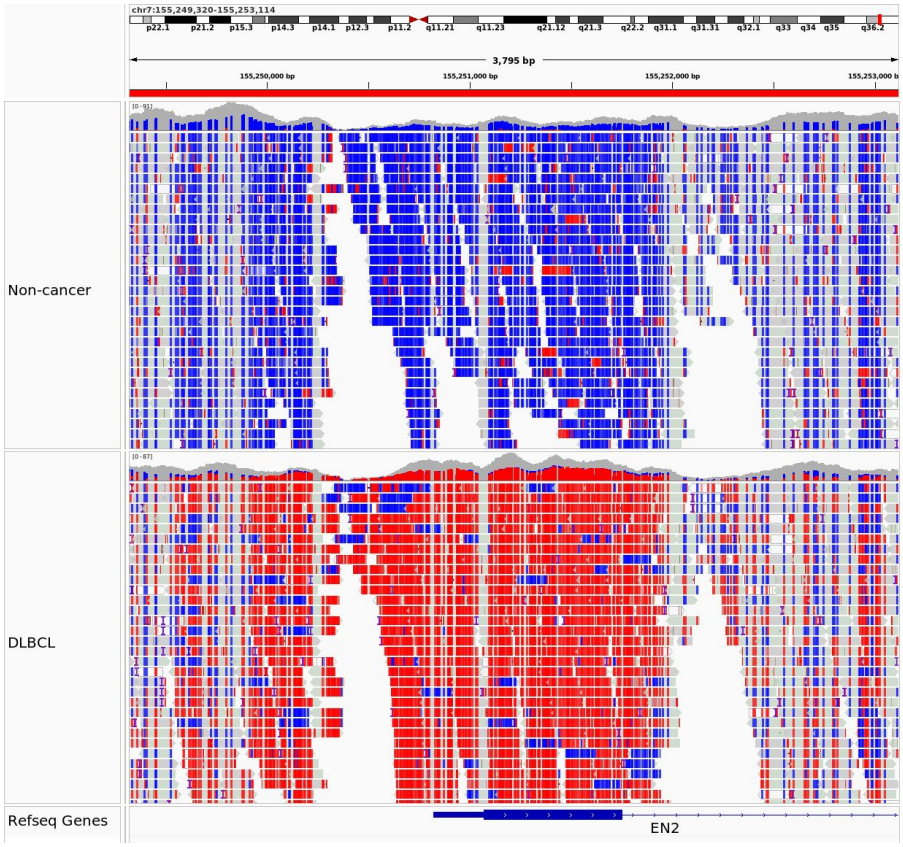
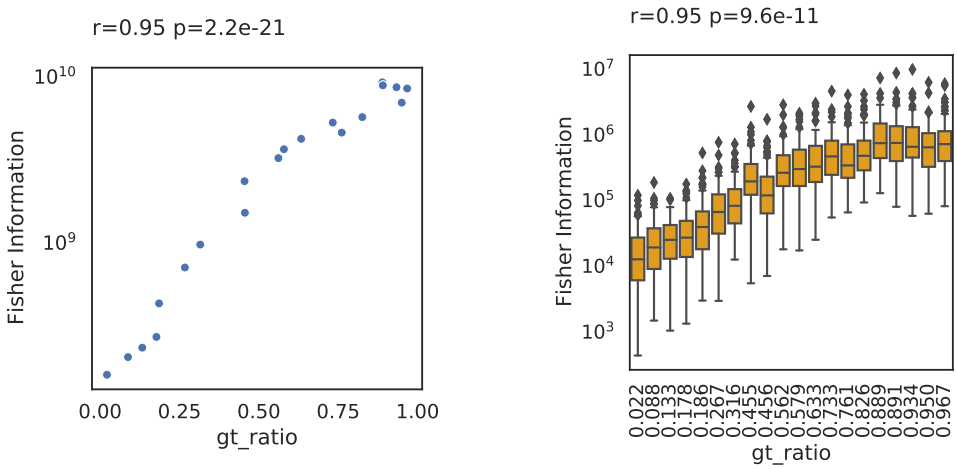


Figure 5.12: Read-level methylation patterns for normal and tumour cells in DMR 20. The blue and red mean unmethylated and methylated CpG each.



(A) **(B)**

Figure 5.13: Correlation between the ground-truth tumour purity and the Fisher information. **(A)** MethylBERT without adjustment. **(B)** MethylBERT with adjustment. When the adjustment is applied, likelihood estimation is done in individual regions, thus the Fisher information is also calculated separately.

If multiple bulk samples are given, MethylBERT not only performs deconvolution of the samples but also calculates the Fisher information (Section 4.3.2). MethylBERT, without the adjustment of estimated tumour purity, calculates the Fisher information over log-likelihood values for the global tumour purity (Figure 5.13A), and the Fisher information shows a positive correlation with the ground-truth tumour purity. When the tumour purity estimation adjustment is applied, the likelihood is calculated in each DMR. Therefore, MethylBERT outputs as many Fisher information values as the number of DMRs. For the pseudo-bulk samples, the median of region-wise Fisher information also correlates with the ground-truth tumour purity (Figure 5.13B). Both positive correlations indicate that accurate estimation of tumour purity is harder when the tumour purity is higher. We believe that this is also associated with the higher complexity of tumour methylation patterns and the impurity of the source of tumour data in the training set.

We also study whether MethylBERT can reconstruct the methylation profiles of tumour and normal cell types from a bulk sample, in addition to the accurate estimation of tumour purity. Figure 5.14 shows the comparison between the ground-truth and MethylBERT-estimated methylation profiles of tumour and normal cell types. As an example, we reconstructed the cell type-specific methylation profiles in DMR 14 for the pseudo-bulk 8 sample which has relatively balanced tumour-normal cell-type proportions (0.456 and 0.544 for tumour and normal cell types, see Table 5.1). In the CpG-wise reconstruction (Figure 5.14A), MethylBERT successfully inferred the methylation patterns corresponding

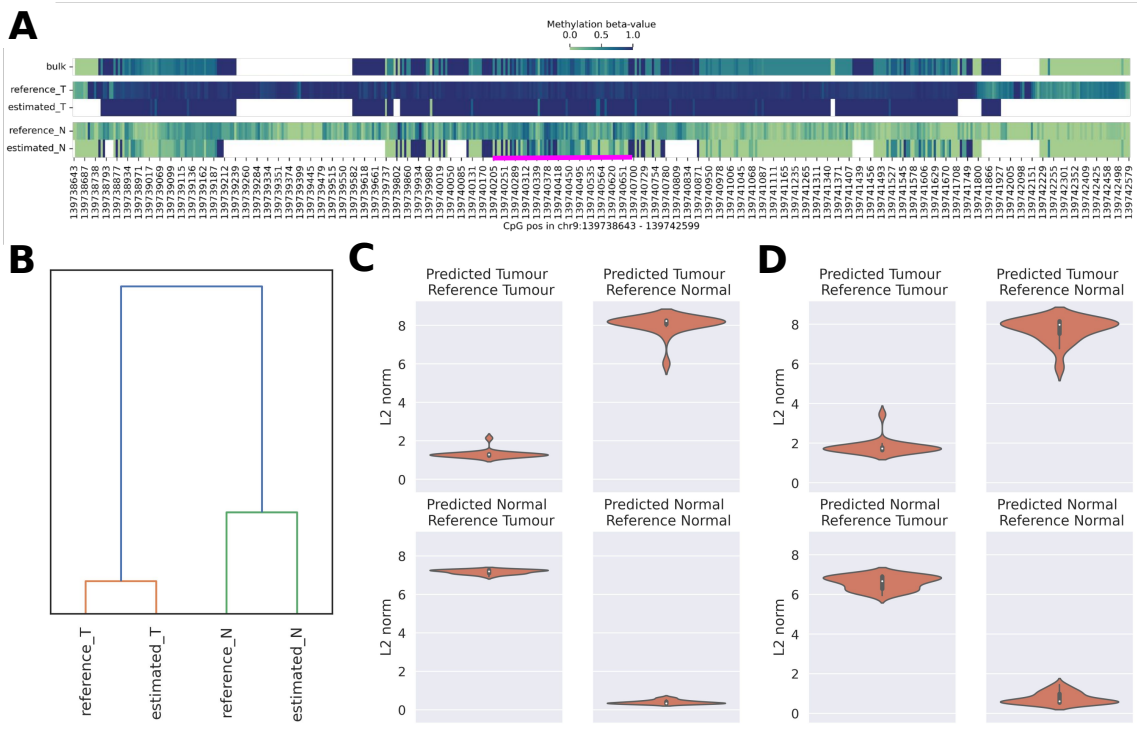


Figure 5.14: Reconstructed methylation patterns in tumour and normal cell types by MethylBERT. **(A)** Estimated tumour (T) and normal (N) CpG-wise methylation level by MethylBERT. The result was obtained for DMR 14 (chr9:139738643-139742599) in pseudo-bulk 8. The complex methylation patterns in the normal cell type are marked by the magenta line. The empty spaces mean that there are no sequencing reads in the sub-region. **(B)** The clustering of estimated and reference methylation profiles of DMR 14 in pseudo-bulk 8. We used the hierarchical clustering algorithm. **(C)** L2-norm calculated between the estimated and reference methylation levels in selected DMRs. **(D)** L2-norm calculated between the estimated and reference methylation levels in a new set of DMRs where hyper- and hypomethylated tumour methylation patterns are equally distributed.

to the ground-truth profiles. In particular, MethylBERT could reconstruct the complex methylation patterns in the normal cell type (marked by the magenta line in Figure 5.14A). Furthermore, the estimated profiles could be clustered with the reference profiles of the correct cell types (Figure 5.14B).

In addition to the CpG-wise reconstruction analysis, we compared the estimated cell type-specific methylation profiles in DMRs to the reference profile. In the 100 selected DMRs, MethylBERT successfully inferred the average methylation levels in every bulk sample except for bulk 7 (Figure 5.15A). The inference did not perform well for the tumour cell type in bulk 7 due to the lowest percentage of tumour cells. The L2-norm values between the estimated and the reference methylation profiles are also lower for a pair of the same cell types than for a pair of different cell types (Figure 5.14C). To prove that

the inference does not depend on the tumour hypermethylation dominantly shown in the selected DMRs, we did the same analysis with another 100 DMRs that 50 regions are tumour hypermethylated and the others are tumour hypomethylated, respectively. As a result, MethylBERT yielded an accurate reconstruction of both tumour and normal cell type-specific methylation profiles (Figure 5.15B). Again, the L2-norm values are lower for the same cell type than for the different cell types (Figure 5.14D).

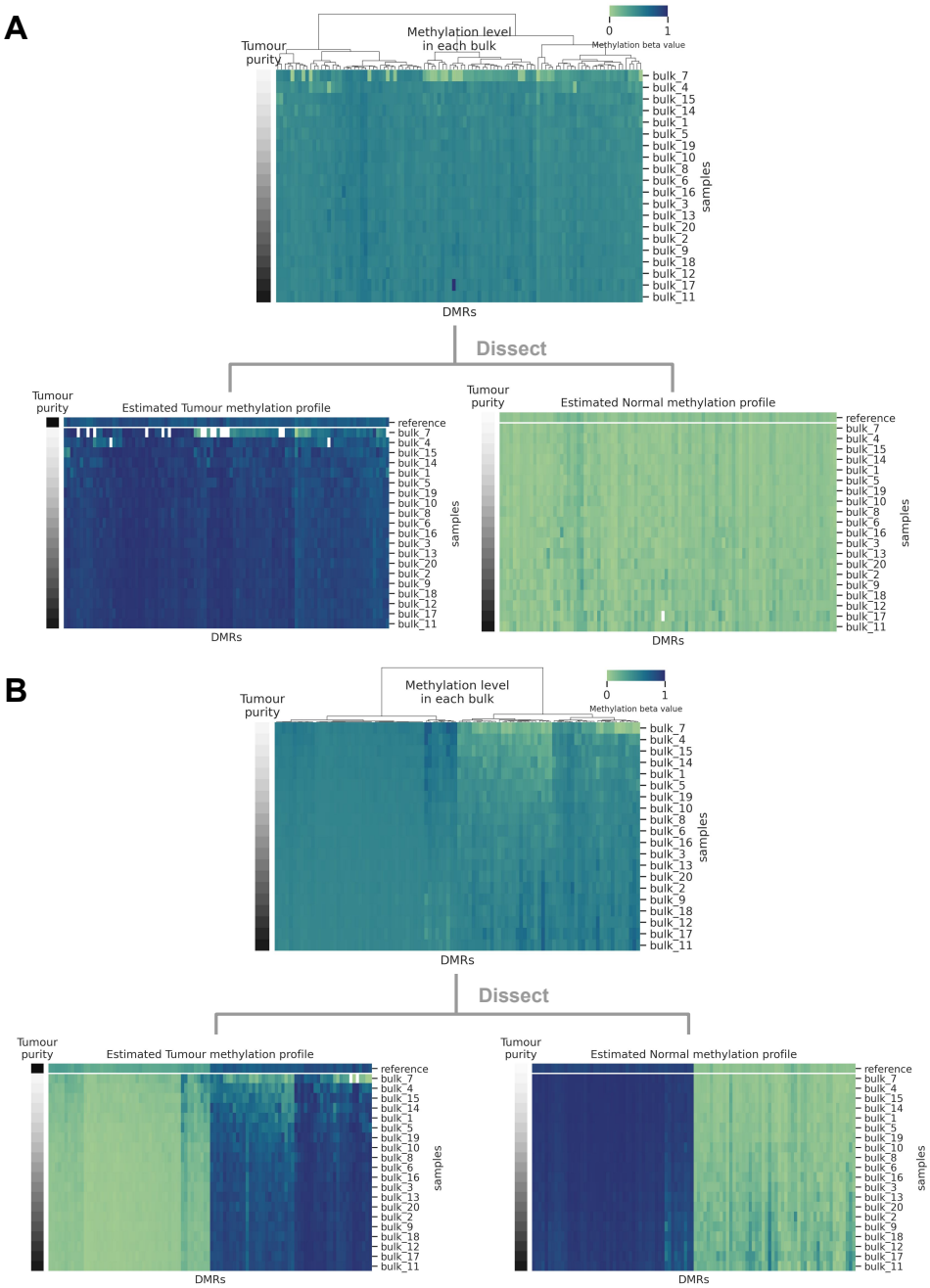


Figure 5.15: Inferred methylation profiles of each cell type from pseudo-bulks compared to the reference profile. **(A)** Inference results when 100 DMRs were selected based on areaStat. **(B)** Inference results when 100 DMRs were selected as tumour hyper- and hypomethylated patterns evenly distributed within the set of regions. In each subfigure, the top heatmap represents the methylation level in DMRs (columns) in each pseudo-bulk (rows). Two heatmaps at the bottom present inferred methylation profiles for tumour (left) and normal (right) cell types. The reference methylation profile of each cell type is shown in the first row of two bottom heatmaps. Pseudo-bulks are ordered by the ground-truth tumour purity.

5.4.2 Evaluation using tumour pseudo-bulks with a very low tumour purity

The recent progress in tumour diagnosis using ctDNA increases the demand for a highly sensitive tumour purity estimation model that is capable of handling a very low number of tumour-derived DNA fragments in blood plasma samples. As introduced in Section 2.1.2, ctDNA analysis enables early diagnosis of cancer in a non-invasive manner by using blood biopsies. Hence, we also conducted a validation of MethylBERT using 10 pseudo-bulks whose tumour purity is below 1% (Table 5.1).

MethylBERT with the adjustment performs best with the lowest median absolute percentage error (MAPE) and the highest correlation between the ground truth and the estimated values (Figure 5.16). Although CancerDetector achieves a relatively low MAPE value, it cannot achieve a high correlation value. Reversely, DISMIR either cannot yield a reasonable inference (when its own region selection method is used) or constantly overestimates the purity (when selected DMRs are used). The DISMIR_dmr results only show a strong correlation between the ground truth and the estimated values. Similarly, Houseman's method also achieves a relatively strong correlation between the estimates and the ground-truth values but achieves a higher MAPE value than MethylBERT. Hence, we conclude that MethylBERT is the most sensitive method to detect ctDNA signals in blood plasma samples in comparison to previous methods.

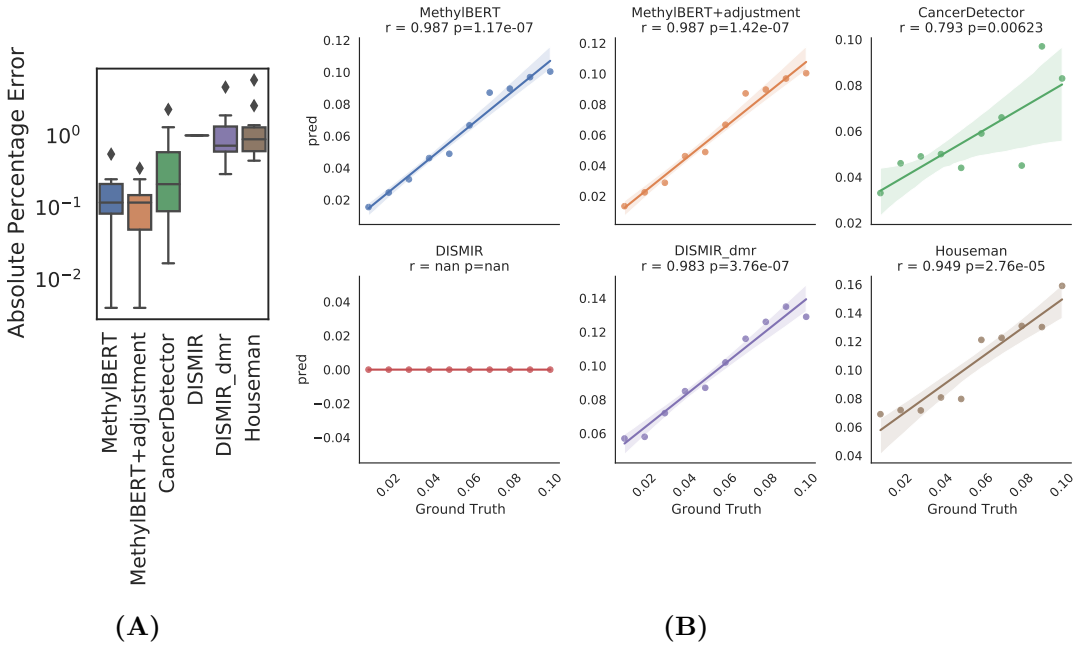


Figure 5.16: Tumour purity estimation results for pseudo-bulk with a very low tumour purity. **(A)** Absolute percentage error of different methods. **(B)** Spearman's correlation between ground truth and estimated tumour purity values and p-value.

5.5 The importance of pre-training

Pre-training is known to not only better generalisation but also show regularisation-like behaviour in large deep neural networks [Erhan et al., 2009]. Particularly, BERT has a unique design of bidirectional pre-training which resolves a conventional problem in NLP models such as task-specific network architectures [Devlin et al., 2018]. For example, Embeddings from Language Models, also known as ELMo, are generally integrated into a task-specific network architecture despite the bidirectional feature extraction [Peters et al., 2018]. On the other hand, BERT can be applied to varying tasks via fine-tuning.

The pre-trained BERT model has also contributed to the advanced analyses of DNA sequences. For example, [Ji et al., 2021] successfully applied the BERT model to genomics-related tasks, such as promoter or enhancer region prediction, by using a pre-trained BERT model with DNA sequences.

From the perspective of NLP, the kind of information gained during BERT pre-training has been meticulously analysed [Clark et al., 2019, Jawahar et al., 2019]. Nevertheless, BERT pre-training on genomic sequences is still poorly understood. For this reason, we analyse the DNA context learnt by the BERT model via pre-training and the efficacy of pre-training in the fine-tuning task.

Primarily, we found that the BERT model can learn the common associations between DNA 3-mer tokens during the pre-training (Figure 5.17). In the UMAP plot of embeddings

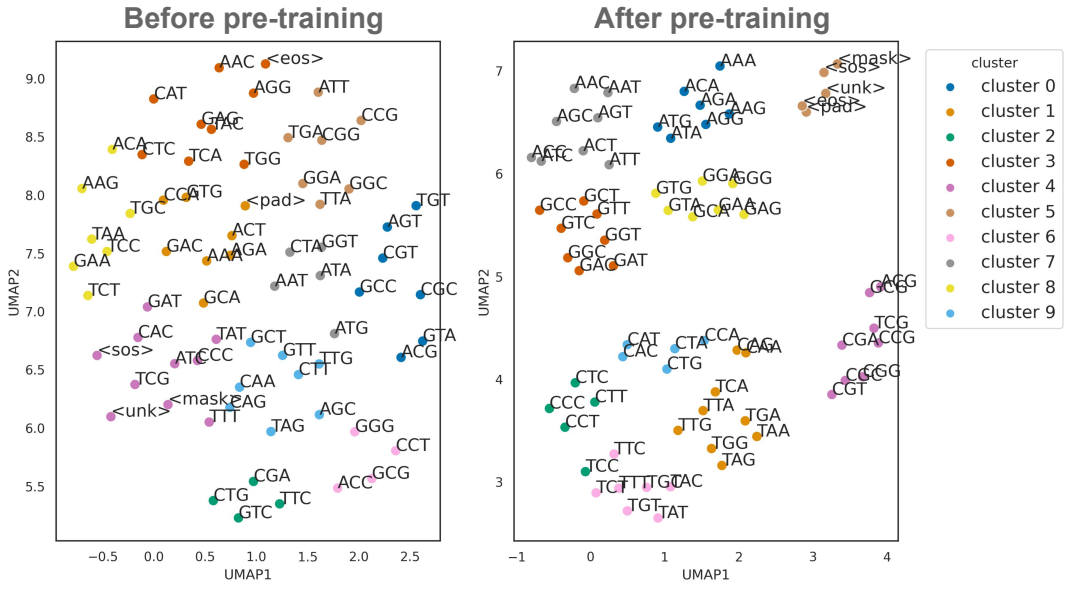


Figure 5.17: UMAP plot of 3-mer token embeddings extracted from the MethylBERT model before pre-training (left) and after pre-training (right). K-means clusters are shown by different colours.

after the pre-training, the tokens that include "CG" context are regarded as a separate cluster (cluster 4) from others, although the information about CpG sites or CpG methylation was not given during the pre-training. We presume the BERT model could already learn the unique characteristics of CpGs from the repetitive occurrence of "CG" on the genome, especially in CpG islands (CGIs). Equally, special tokens are only distinguishable after pre-training (cluster 5). Furthermore, the pre-trained BERT model can relate 3-mer tokens in terms of DNA nucleotide pairs (A-T and C-G). Along the UMAP2 axis, the tokens are separated into two groups: one starts with A/G (UMAP2 > 5) and the other starts with C/T. The majority of 3-mer tokens are also clustered according to the last nucleotides within each group. For example, cluster 0 is comprised of 3-mer tokens starting with A and ending with A or G, which are nucleotide pairs. In our hypothesis, the identification of the nucleotide pairs achieved by the pre-trained BERT model could be feasible due to Chargaff's second parity rule that two paired nucleotides take an approximately equal proportion in a single DNA strand [Rudner et al., 1968].

The final goal of BERT pre-training is to improve the generalisation of a model to an unseen domain without requiring a massive amount of task-specific data sets. Therefore, we investigate how much fine-tuning performance improvement is achieved by pre-training in MethylBERT. For the fine-tuning read classification task, read-level methylomes from DLBCL and normal B cell samples used in Section 5.3 were chosen. Although the DMRs were chosen by the areaStat score in Section 5.4, the hypermethylation in promoter regions can misguide the pre-training efficacy analysis (Figure 5.18). Therefore, we chose 50 tumour-hypermethylated DMRs and 50 tumour-hypomethylated DMRs for the analysis.

At the beginning of fine-tuning, both MethylBERT models without and with pre-training have decreasing loss values, but only the model with pre-training reaches a fine-tuning accuracy value above 0.9 for training and validation data sets (Figure 5.19A). On the other hand, MethylBERT without pre-training cannot increase the accuracy approximately after 100 steps and the accuracy value drastically decreases afterwards. According to the AUC value calculated for the validation set with the best-optimised parameters, MethylBERT with pre-training can classify reads into cell types far more accurately (Figure 5.19B). Read classification is not as accurate when MethylBERT is not pre-trained.

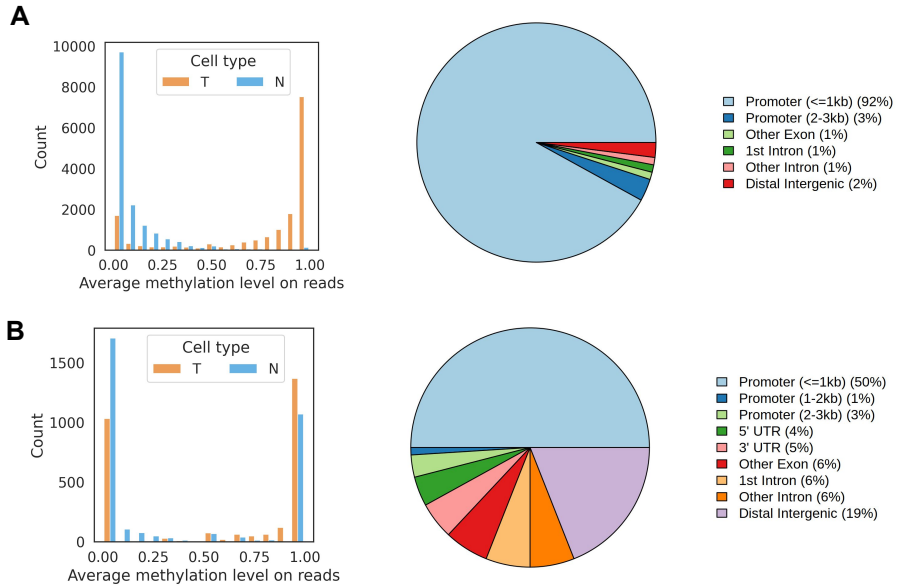


Figure 5.18: Distribution of methylation levels and genomic annotations in selected DMRs when **(A)** DMRs are selected based on the areaStat score and **(B)** DMRs are selected to have the balanced proportion of tumour hyper- and hypomethylation patterns. The histogram shows the distribution of the average methylation levels in each cell type, whereas the pie chart represents the proportions of genomic annotations within selected DMRs.

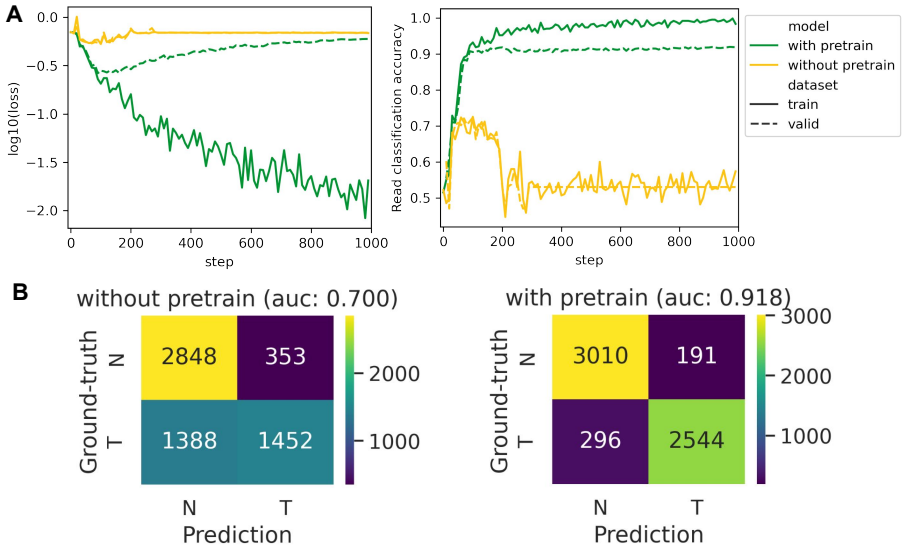


Figure 5.19: Fine-tuning performance of MethylBERT model without and with pre-training. **(A)** Loss and read classification accuracy changes over fine-tuning steps. **(B)** Confusion matrix calculated for read classification by MethylBERT without and with pre-training

The learning trend of MethylBERT with and without pre-training is presented more clearly in Figure 5.20. For the pre-trained MethylBERT model, the P(cell type = Tumour|read) distribution of tumour and normal reads already start separating from each other at training step 50 and compute higher values of P(cell type|read) over further training steps. When MethylBERT is not pre-trained, the P(cell type = Tumour|read) distribution of normal and tumour reads are slightly different from each other at early steps, but both distributions converge at the probability value 0.5 after 250 steps. The correlation between P(cell type = Tumour|read) and the average methylation level of individual reads shows that the read classification of both models is biased towards the methylation level at step 50 (Figure 5.21). In other words, P(cell type = Tumour|read) is usually higher when the read is fully methylated. Afterwards, the pre-trained MethylBERT model can overcome the bias and classify reads more accurately, showing a lower correlation between the probability and the methylation level. The lower correlation and the high performance after step 50 imply that the model can classify reads into correct cell types based on cell type-specific methylation patterns in respective DMRs. However, the MethylBERT model without pre-training keeps classifying reads based on the average methylation level until

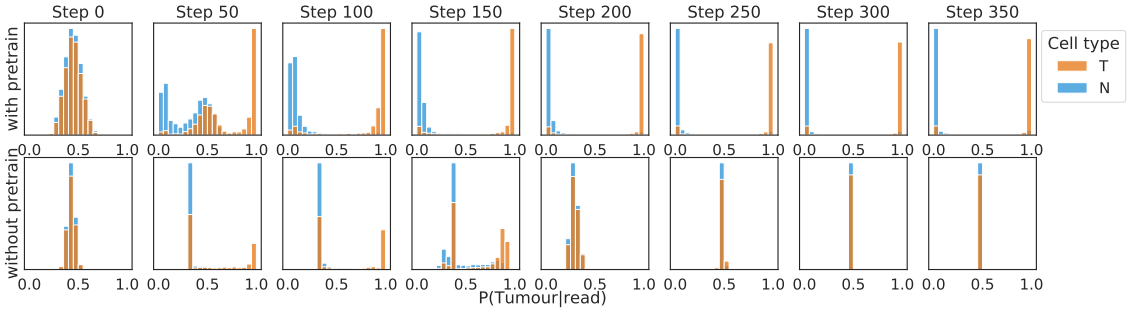


Figure 5.20: Distribution of P(cell type = Tumour|read) during MethylBERT fine-tuning. The top and bottom rows represent MethylBERT with and without pre-training, respectively. The distributions were calculated every 50 steps between step 0 and step 350. 'T' and 'N' refer to tumour and normal cell types.

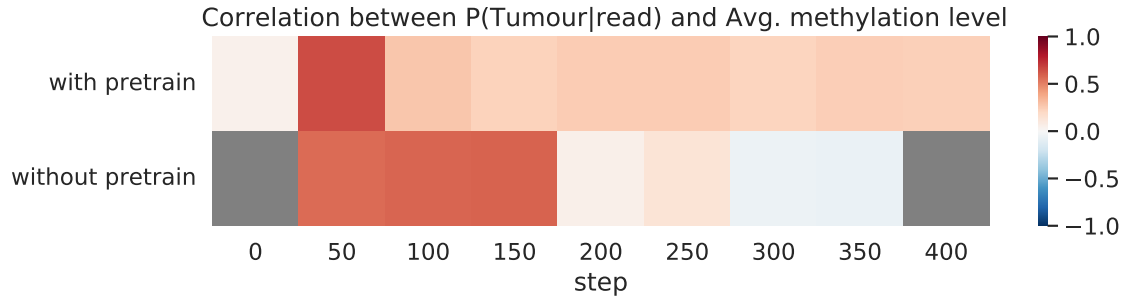


Figure 5.21: Correlation between P(cell type = Tumour|read) and average methylation level calculated over all reads. Correlation with a p-value below 0.05 is coloured by grey. The correlation was calculated every 50 steps between step 0 and step 400.

step 150, and then the classification performance decreases. Consequently, we argue that pre-training plays a key role in the read classification task by preventing the model training from being biased towards average methylation level and making the model detect correct tumour-specific methylation patterns.

The biggest challenge of pre-training is the long training time caused by the large data set. If the same pre-trained MethylBERT model is applicable to different species, it improves usability by broadening the range of analysis without additional pre-training. Figure 5.22 shows the comparison of the fine-tuning results between the two MethylBERT models pre-trained with the human genome (hg19) and mouse genome (mm10). In terms of read classification AUC score for the validation set, the two models achieve a very similar value showing a difference smaller than 0.001. When comparing the distribution of $P(\text{cell type} = \text{Tumour}|\text{read})$ calculated by each model, normal cell-derived reads did not have a significant difference yielding the p-value of paired t-test statistics equal to 1.0. Nevertheless, the distribution of tumour reads still significantly differs between the two models showing the p-value below 1.0×10^{-5}.

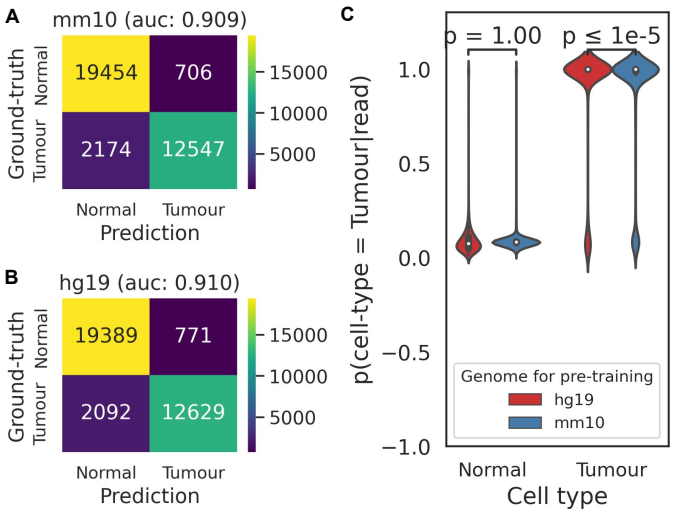


Figure 5.22: Read classification results after pre-training with human genome (hg19) and mouse genome (mm10). **(A)** Confusion matrix of read classification with mm10 pre-trained MethylBERT model. **(B)** Confusion matrix of read classification with hg19 pre-trained MethylBERT model. **(C)** $P(\text{cell type} = \text{Tumour}|\text{read})$ distribution of normal B cell and tumour-derived B cell computed by both pre-trained MethylBERT models.

5.6 The effect of tumour purity estimation adjustment

We also assess the performance of MethylBERT with the adjustment algorithm described in Section 4.3.3. In Figures 5.8 and 5.16, it can be seen that the tumour purity estimation performance slightly increases with the adjustment. On the other hand, CancerDetector yields a higher error value for the bulk samples with tumour cell fraction above 50%

although the read classification performance should be the same regardless of the tumour purity. These imply that the tumour purity estimation relies not only on the read classification results, but also on handling the variability of region-wise tumour purity. Therefore, we compare the adjustment method to handle the variability of region-wise tumour proportion in a bulk between MethylBERT and CancerDetector.

As explained in Section 5.2, we consider the 'region filtering' step in the CancerDetector algorithm conceptually equal to MethylBERT's tumour purity adjustment algorithm. Therefore, in this section, we call the region filtering step 'CancerDetector tumour purity adjustment' for convenience. The CancerDetector tumour purity adjustment algorithm is described in Algorithm 3. CancerDetector employs the EM algorithm to remove DMRs whose estimated purity is outside of the standard deviation of region-wise tumour purity. The removal is conducted iteratively until the estimated tumour purity value converges. Then, the converged value is regarded as the final estimation.

For a fair comparison of the two adjustment methods, the evaluation is done using the same read-wise probabilities P(cell type|read) for pseudo-bulk samples computed by MethylBERT. We note that the performance of CancerDetector in this section is thus different from the CancerDetector result shown in Figure 5.8, which was yielded using P(cell type|read) calculated by CancerDetector.

When the adjustment is applied, MethylBERT achieves a lower median absolute error

Algorithm 3 CancerDetector tumour purity estimation adjustment algorithm

Input: $\mathbf{P}^t = \{p_1^t ... p_R^t\}$, $\mathbf{P}^n = \{p_1^n ... p_R^n\}$ Calculated P(read|cell type) for all R reads
$\boldsymbol{\mathcal{M}} = \{m_1 ... m_K\}$ K DMRs
λ Factor for std^a of region-wise estimation
θ Threshold for EM algorithm iteration

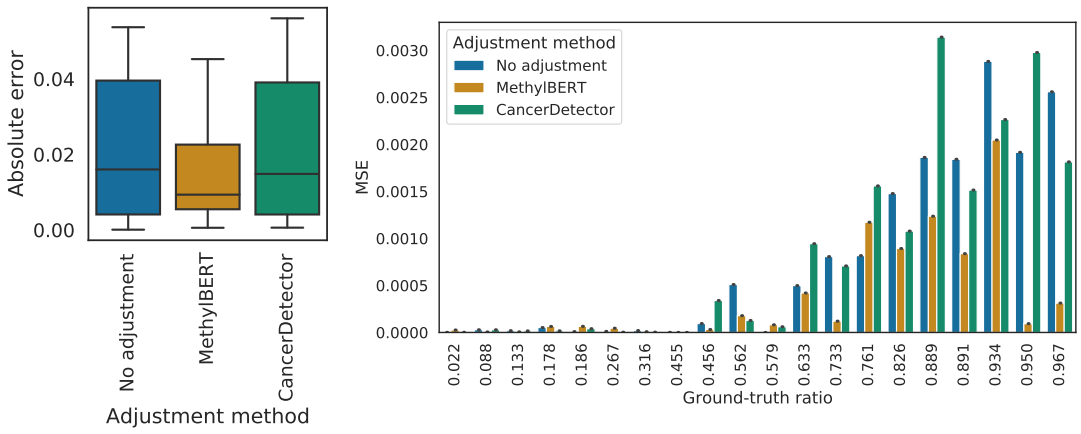
Output: δ_{new} Estimated tumour purity

$\delta_{new} \leftarrow \arg\max_{\delta} \prod_{r_i=1}^{R} \{\delta p_{r_i}^t + (1-\delta) p_{r_i}^n\}$
while $|\delta_{new} - \delta_{prev}| > \theta$ **do**
  for $m_k \in \boldsymbol{\mathcal{M}}$ **do** ▷ Estimate region-wise tumour purity
    $\delta_k \leftarrow \arg\max_{\delta} \prod_{r_i \in m_k} \{\delta p_{r_i}^t + (1-\delta) p_{r_i}^n\}$
  for $m_k \in \boldsymbol{\mathcal{M}}$ **do** ▷ Exclude regions whose estimation is not in the std
    if $\delta_k > \delta_{new} + \lambda std(\{\delta_1, ..., \delta_K\})$ **then**
      Remove m_k from $\boldsymbol{\mathcal{M}}$
  $\delta_{prev} \leftarrow \delta_{new}$
  $\delta_{new} \leftarrow \arg\max_{\delta} \prod_{r_i=1}^{R} \{\delta p_{r_i}^t + (1-\delta) p_{r_i}^n\}$

[a] The standard deviation

than CancerDetector (Figure 5.23A). The performance gain of MethylBERT with the adjustment mainly comes from the pseudo-bulk samples where the tumour proportion is higher than 50% (Figure 5.23B). Especially when the tumour proportion is above 95%, MethylBERT with the adjustment performs better than the others.

Figure 5.24 shows why MethylBERT adjustment works better than CancerDetector adjustment when the tumour purity is very high. CancerDetector was originally designed for ctDNA analysis, so the adjustment algorithm excludes regions whose estimated purity is higher than the sum of the standard deviation of local purities and the estimated global purity. This way of removing regions causes an underestimation of final tumour purity when tumour-derived reads are not rare. On the other hand, MethylBERT adjustment does not filter out any regions, and rather it transforms the region-wise estimated tumour purities to a less biased distribution centred around the global tumour purity. This alleviates the underestimation problem occurring in the CancerDetector adjustment and improves the estimation regardless of the ground-truth tumour purity.



(A) **(B)**

Figure 5.23: Performance evaluation of MethylBERT adjustment method. **(A)** Absolute error calculated by each method for 20 pseudo-bulk samples. **(B)** Mean squared error calculated by each method ordered by ground-truth tumour purity.

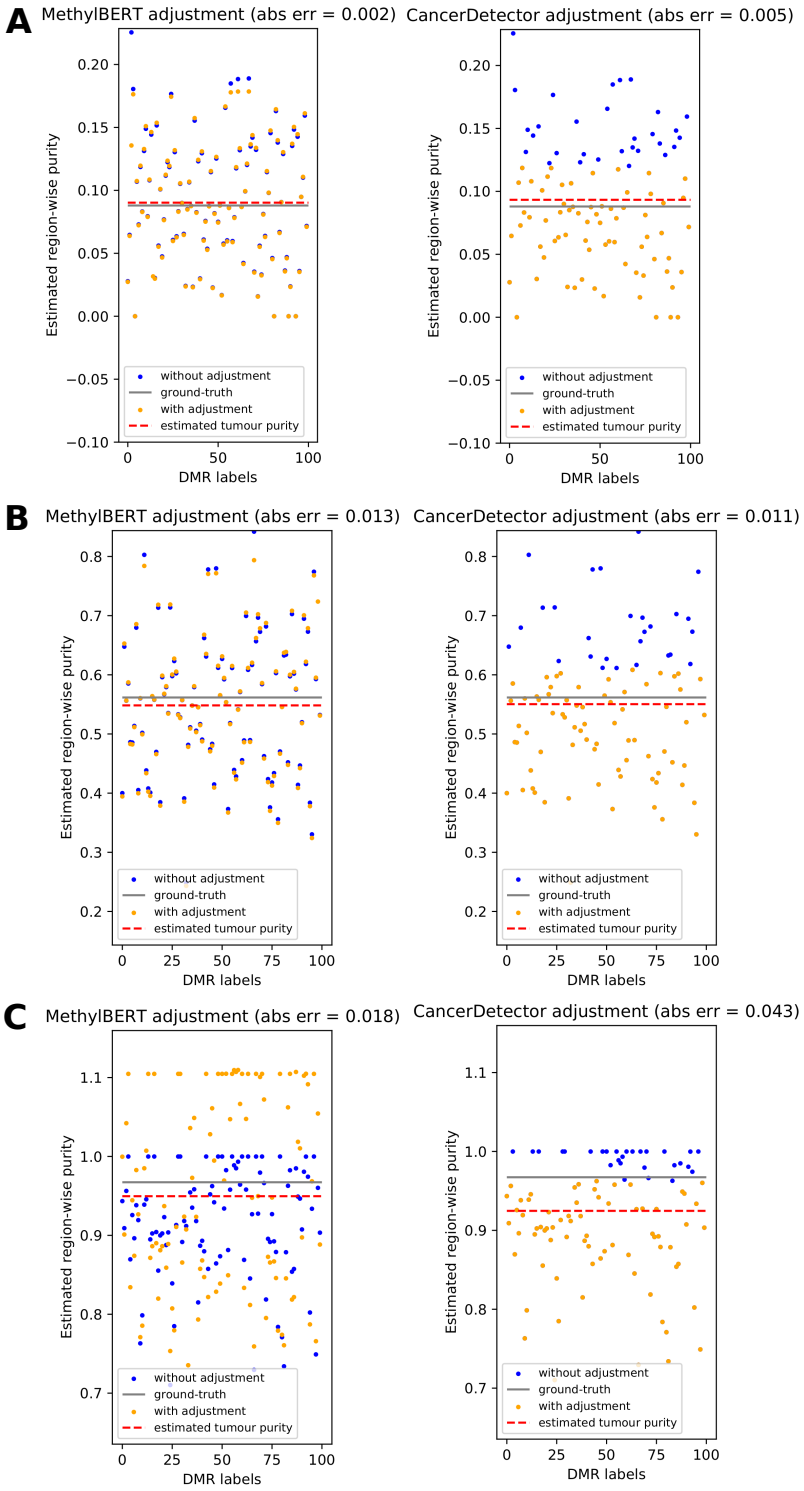


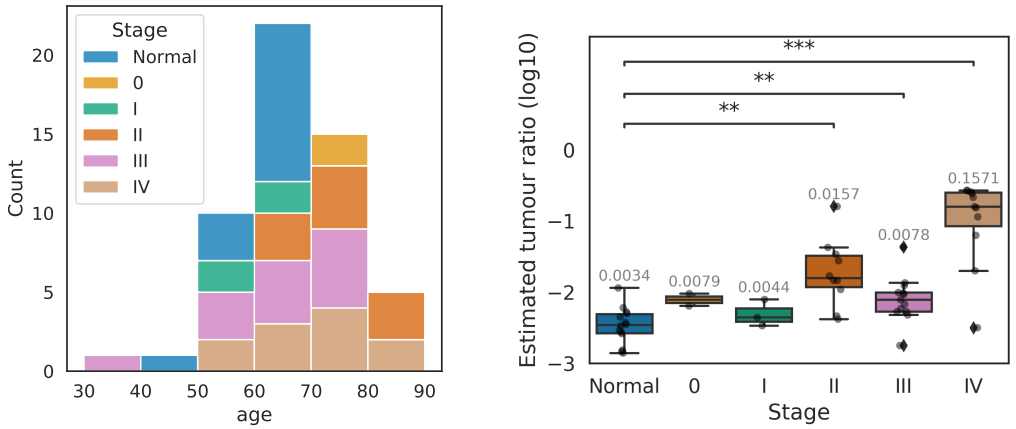
Figure 5.24: Comparison between region-wise estimation before and after adjustment. The left column shows the result of the MethylBERT adjustment, and the right column presents the result of the CancerDetector adjustment. The example has three different cases of pseudo-bulks whose tumour content is 0.088, 0.579, and 0.967, respectively.

5.7 Tumour diagnosis based on ctDNA in blood plasma samples

In Section 2.1.2, we explained ctDNA and evaluated the performance of MethylBERT using pseudo-bulk samples resembling ctDNA. ctDNA analysis enables non-invasive early cancer diagnosis because the samples can be collected via liquid biopsy. Nonetheless, the very low percentage of tumour-derived DNA fragments found in liquid biopsies requires highly sensitive cell-type deconvolution methods which can detect rare cell-type signals. In this case, sequencing-based cell-type deconvolution methods are considered more suitable than array-based methods because the methylation beta-value given by array-based profiling can obscure the rare cell-type signals. The two benchmarked sequencing-based cell-type deconvolution methods, CancerDetector and DISMIR, were also developed for this purpose. Hence, we evaluate the usability of MethylBERT for non-invasive cancer diagnosis using blood plasma samples from cancer patients.

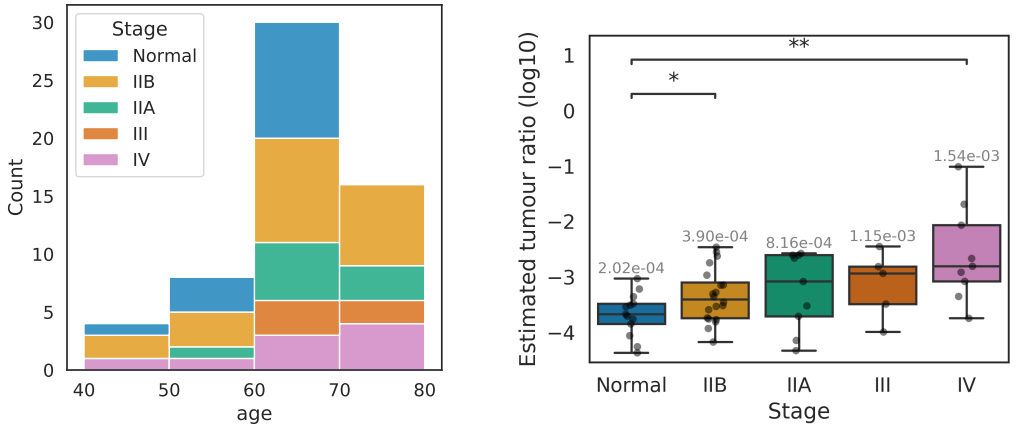
We evaluated tumour purity estimation with blood plasma samples collected from 14 healthy donors and 40 colorectal cancer (CRC) patients (GEO accession number GSE149438). The patient cohort is comprised of five different stages of CRC from stage 0 to stage IV. The MethylBERT model was trained with the other non-cancer plasma samples from the same data set and scBS-seq data from CRC tissue samples (GEO accession number GSE97693). Figure 5.25 shows the analysis result. Based on the estimated tumour purity, some of the early-stage CRC patients (II-III) are distinguished from the healthy donors with p-value ≤ 0.01. Furthermore, the median value of estimated tumour purity is higher for all stages of CRC patient groups than the healthy donor group. The estimated median value is lower than 0.01 in all early stages of CRC patients (0-III), thus a sensitive model for tumour purity estimation like MethylBERT is undoubtedly required to enable non-invasive early tumour diagnosis.

44 pancreatic ductal adenocarcinoma (PDAC) patient samples were also analysed compared to the 14 healthy donors (Figure 5.26). The patient samples are categorised by four stages (IIA, IIB, III, and IV). The MethylBERT model was trained with the other healthy donor blood plasma samples from GSE149438 and PDAC tissue WGBS data downloaded with the GEO accession number GSE63123. Likewise, the median estimated tumour purity is higher in all stages of PDAC patients than in healthy donors. In particular, stage IIB patients are distinguished from the healthy donors with the p-value < 0.05. This result is particularly meaningful because PDAC is widely recognised as one of the trickiest cancer types to be identified during the early stages. The original paper providing this data set also confirmed this by showing their cell-type deconvolution results where PDAC patients were the most difficult cases to be identified based on DNAm profiling compared to other cancer patients such as colorectal, liver, and gastric cancers [Kandimalla et al., 2021].



(A) **(B)**

Figure 5.25: ctDNA analysis by MethylBERT using blood plasma samples from healthy donors and CRC patients. **(A)** Distribution of age and cancer stages in the cohort. **(B)** Estimated tumour purity plotted by cancer stages. The p-value is calculated via a two-sided Mann-Whitney-Wilxcoxon test. '**' and '***' mean p-value$\leq$ 0.01 and $\leq$ 0.001, respectively. The median value is written at the top of each box.



(A) **(B)**

Figure 5.26: ctDNA analysis by MethylBERT using blood plasma samples from healthy donors and PDAC patients. **(A)** Distribution of age and cancer stages in the cohort. **(B)** Estimated tumour purity plotted by cancer stages. The p-value is calculated via a two-sided Mann-Whitney-Wilxcoxon test. '*' and '**' mean p-value$\leq$ 0.05 and $\leq$ 0.01, respectively. The median value is written at the top of each box.

5.8 Discussion

Given the limitations of current sequencing-based cell-type deconvolution methods evaluated in Chapter 3, we proposed a new Transformer-based model for cell-type deconvolution using read-level methylomes specifically targeting tumour, MethylBERT, in Chapter 4 and performed an evaluation in Chapter 5. Transformer encoders in the BERT model not only enable bidirectional training over input sequences, but also apply an attention mechanism to the entire sequence. The calculation of an attention matrix over the entire sequence compensates for the lack of global context learning in CNNs, as well as the order-dependent modelling in RNNs. Even though several Transformer-based methods have been successfully used for modelling sequencing-based DNAm data [Yu et al., 2021, De Waele et al., 2022], application for cell-type deconvolution has not been reported to date.

MethylBERT consists of three steps. The first step of MethylBERT is pre-training using reference genome data split into 3-mer tokens. In the second step, the model is fine-tuned to classify reads into cell types based on DNA sequences and read-level methylation patterns. The classification is done by assigning the cell type with the highest posterior probability given the reads. The calculated posterior probability is used in the third step of MethylBERT, which is maximum likelihood estimation of the tumour purity within a given bulk sample.

Accurate estimation of the posterior probability is imperative to ensure high-performance tumour purity estimation. Therefore, we evaluated the read-level methylation classification performance of MethylBERT and existing methods. Since using read-level tumour-derived methylomes obtained from biological samples for evaluation is highly complicated due to other confounding factors or possible contamination, the major evaluation was done using read-level methylation patterns simulated from a beta-binomial distribution. CancerDetector and DISMIR, which were originally developed for ctDNA analysis, as well as Houseman's method which showed the best performance in the evaluation in the evaluation in Chapter 3 were included in the comparison. The simulation was conducted by controlling methylation pattern complexity, read length, and read coverage. MethylBERT mostly outperforms other methods, regardless of simulation setups.

The MethylBERT tumour purity estimation was also evaluated and compared with three previous methods. As an evaluation data set, we used pseudo-bulk samples created using DLBCL and normal B cell WGBS data. Again, MethylBERT shows the lowest absolute error values. Houseman's method performs better than MethylBERT only when the tumour proportion is higher than 0.5, but cannot perform tumour purity estimation for a low percentage of tumour-derived reads. On the contrary, the sequencing-based methods CancerDetector and DISMIR tend to perform much better only when the ground-truth tumour purity is low because they were originally designed to handle ctDNA analysis where a low tumour signal is the biggest challenge.

From the MethylBERT cell-type deconvolution results, we discovered that estimation gets less accurate with increasing ground-truth tumour purity. This can be caused by partially methylated reads which occur more frequently in tumour cells. The Fisher information increases with a higher tumour purity for MethylBERT, as well as for MethylBERT with the adjustment. Nevertheless, the adjustment in MethylBERT mitigates the problem of biased local tumour purity distribution and improves the deconvolution accuracy in tumour-dominant pseudo-bulk samples.

Furthermore, we interrogated the information the BERT model learned from DNA sequences during pre-training, which has not been done in previous work. The analysis of pre-training results shows that MethylBERT can learn the basic context of DNA sequences, such as CpG sites or DNA nucleotide pairs without any prior knowledge. Accordingly, accurate read classification is achieved only after pre-training. If the model is not pre-trained, it can classify reads into cell types only based on average methylation level, although cell type-specific methylation patterns differ among genomic regions. These results show that pre-training is essential in MethylBERT, despite the absence of information about DNA methylation in the training data.

ctDNA analysis, as mentioned in previous chapters, has recently received a lot of attention as a non-invasive tumour detection method. DNA methylation modifications particularly occur at the very early stage of the tumour. Thus, DNAm profiles are considered a better biomarker to conduct ctDNA analyses than mutational-based profiles. Yet, the very low quantity of tumour-derived DNA fragments in blood plasma collected from cancer patients demands a highly sensitive cell-type deconvolution method for the diagnosis of cancers using liquid biopsy. For this reason, in previous work, CancerDetector and DISMIR were proposed to use read-level methylomes to detect rare tumour signals in blood plasma and have shown promising performance. Using 10 DLBCL pseudo-bulk samples whose tumour purity is below 1%, we confirmed that MethylBERT performs better than these two methods as a sensitive cell-type deconvolution method for ctDNA analysis. Moreover, MethylBERT successfully distinguishes early CRC and PDAC patients from healthy donors based on tumour purity estimation for blood plasma samples. Taken together, we claim that MethylBERT can be used as a non-invasive early tumour diagnosis method in the future.

Chapter 6

Conclusion and future work

6.1 Conclusion

This thesis focuses on methods for cell-type deconvolution using read-level DNA methylomes. The thesis consists of three main parts:

- First, we have comprehensively analysed and benchmarked previously published sequencing-based cell-type deconvolution methods. This work is the first benchmarking study specifically targeting sequencing-based cell-type deconvolution methods and contributes important input to the field of bioinformatics. We provide systematic and unbiased evaluations using the same data sets and metrics, as well as a thorough examination of their algorithmic designs. Furthermore, we established a new paradigm in evaluating sequencing-based cell-type deconvolution methods. We subdivided the evaluation into two major steps: informative region selection and cell-type composition estimation. Our benchmarking results confirm that the informative region selection algorithm has a significant impact on cell-type composition estimation performance. In particular, for the bi-component pseudo-bulk samples, the cell-type composition estimation accuracy is higher when the selected regions are more similar to the differentially methylated regions (DMRs). Therefore, in the case of reference-based methods requiring pure cell-type samples, we recommend using pre-calculated DMRs based on the reference data rather than estimating informative regions based on the variance of the methylation level in given bulk samples. Consequently, our benchmarking study showed that the currently available methods do not significantly outperform array-based deconvolution methods, despite the much more abundant and detailed information in sequencing-based DNAm data.

- Second, taking the limitations of current methods into account, we have developed a new cell-type deconvolution method based on Transformers for read-level DNAm pat-

terns derived from tumours, named MethylBERT. MethylBERT is the first application of Transformers to cell-type deconvolution for DNAm data. In the MethylBERT network, the BERT model encodes read-level methylation patterns and corresponding DNA sequences together to identify tumour-specific signals in single sequencing reads. For pre-training, the network performs the masked language model task on 3-mer DNA sequences. Afterwards, MethylBERT is fine-tuned to classify cell types based on read-level methylomes. Based on the posterior class probability calculated by the MethylBERT network, it finally estimates the proportion of tumour cells within a bulk sample. We also suggest using Fisher information to determine the precision of the tumour purity estimation. MethylBERT supports the adjustment of purity estimation by minimising the skewness of estimated local proportions. This scheme is designed to avoid over- or underestimation of tumour purity, especially when bulks have a very high or low percentage of tumour-derived reads.

- Third, we evaluated MethylBERT and performed a comparison with previous methods. An analysis of pre-training results emphasises the importance of the pre-training step in MethylBERT, although it does not exploit any information about DNA methylation. The masked language model training using 3-mer DNA sequences enables the model to learn the basic context of DNA sequences, such as CpG sites and DNA nucleotide pairs. We showed that the DNA sequence context learnt during the pre-training eventually renders the model fine-tuning unbiased towards the average methylation level and aids the successful detection of the region-wise tumour-specific signals. MethylBERT shows superior performance compared to other methods in both read classification for simulated read-level DNAm patterns, and tumour purity estimation for diffuse large B-cell lymphoma (DLBCL) pseudo-bulk data. The successful deconvolution of pseudo-bulk samples with a very low percentage of tumour-derived reads shows the potential of MethylBERT as a non-invasive early tumour diagnosis method. This is also confirmed by the tumour purity estimation results for the blood plasma samples obtained from colorectal cancer (CRC) and pancreatic ductal adenocarcinoma (PDAC) patients. MethylBERT is capable of distinguishing early-stage cancer patients from healthy donors in both cancer types. To sum up, the evaluation results show that MethylBERT is not bound to a specific type of tumour bulk and is able to accurately estimate tumour purity for bulks regardless of the level of ground-truth tumour purity. This advances the state-of-the-art in tumour purity estimation compared to previous methods that are restricted to a specific type of tumour bulks (e.g., solid tumour tissue or liquid biopsies).

6.2 Future work

In the benchmarking study described in Chapter 3, we evaluated sequencing-based cell-type deconvolution methods for unspecified types of biological samples and tumour bulks. The benchmarking study can be extended to cell-type deconvolution methods for other cell types (e.g., immune cell types). Immune cell-type deconvolution is used for quantifying immune infiltration in the tumour microenvironment. The immune infiltration level contributes to grouping tumour subtypes in terms of clinical relevance as well as drug discovery in immunotherapy [Singh et al., 2021, Tang et al., 2021]. [Sturm et al., 2019] already performed a review study on transcriptome-based immune cell-type deconvolution methods. Yet, such a review study has not been performed for methylation data, even though DNAm-based lymphocyte[1] infiltration estimation has improved diagnosis of various cancer types like breast cancer and glioneuronal tumour [Jeschke et al., 2017]. [Singh et al., 2021] also developed a method specifically for immune cell-type deconvolution using DNA methylation data. Therefore, we suggest a further evaluation of sequencing-based cell-type deconvolution methods for DNA methylomes from tumour microenvironment samples for future work.

Regarding the MethylBERT result analysis, additional analyses can be done by examining calculated attention matrices in the model. Through simulated read-level methylation pattern analysis, we found that MethylBERT accurately performs read classification, even when the neighbouring CpGs do not have the same methylation patterns. This implies that the model successfully distinguishes important CpG sites clearly presenting cell type-specific signals. Therefore, we plan to conduct further analysis to see whether the calculated attention matrices correspond to the informative genomic loci to identify tumour signals. In addition, since blood plasma methylation data is publicly available for varying types of cancer, MethylBERT could be applied to diagnose more cancer types in the early stage using the blood plasma data.

Third-generation sequencing, which is introduced in Section 2.1.3, recently has shed light upon long-range genomic makeup and features. Especially in cancer epigenomics, third-generation sequencing has received attention as a bisulfite-free methylation sequencing method by distinguishing methylation status based on ionic current [Sakamoto et al., 2020]. Third-generation sequencing has already been attempted to profile circulating tumour DNA methylation using long-read sequencing [Katsman et al., 2022]. Nonetheless, MethylBERT application to long-read sequencing is not feasible due to the high number of parameters demanding large memory consumption. Modified versions of the BERT model have been proposed to reduce the computational complexity in the fields of machine learning and computational biology [Zaheer et al., 2020, Beltagy et al., 2020, Avsec et al., 2021]. Therefore, we intend to improve MethylBERT referring to the previous works so that the model has fewer parameters. This can broaden the applicability of MethylBERT

[1] A lymphocyte is a type of immune cell whose subtypes include B cells and T cells.

to long-read sequencing data.

Deconvolution for multiple cell types is another potential direction to develop MethylBERT further. Currently, MethylBERT can only distinguish binary subpopulations, like tumour and normal. Identifying more than two cell types in a bulk sample can stratify subpopulations and enable a more thorough classification of bulk samples. For instance, in the tumour microenvironment, it is necessary to distinguish tumour, normal (stromal) and multiple immune cell types to avoid the confusion between normal and immune cell types. Hence, as another future work, we propose extending the MethylBERT model to handle complex cell populations and subpopulations to support a more comprehensive analysis of bulk samples.

Supplementary

Proof of the softmax function satisfying the probability axioms

In this section, we prove that the standard softmax function σ in Equation (4.3) satisfies the probability axiom 3 when σ_i is used as a probability function of an event that label i is observed, E_i. For the classification with C labels, the function σ is $\sigma : \mathbb{R}^C \to (0,1]^C$ and the event space $F = \{E_1, ..., E_C\}$.

Theorem (probability axiom 3). For any countable sequence of mutually exclusive event sets $\{E_i\}$ chosen from the event space F, $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$.

Proof. Let n be the number of events in a chosen event set $\{E_i\}$, then

$$P(\bigcup_{i=1}^{\infty} E_i) = P(\bigcup_{i=1}^{n} E_i).$$

We note that all events in F are mutually exclusive by our assumption that input is classified into only one of C labels (see Section 4.2.2). Therefore, any event in F can be selected to create $\{E_i\}$.

With this assumption, we prove *Theorem* by mathematical induction. When n is 1, $P(\bigcup_{i=1}^{\infty} E_i) = P(\bigcup_{i=1}^{1} E_i) = P(E_1)$ is obviously true. For the case $n = 2$,

$$\begin{aligned} P(\bigcup_{i=1}^{\infty} E_i) &= P(\bigcup_{i=1}^{2} E_i) \\ &= P(E_1) + P(E_2) - P(E_1 \cap E_2) \end{aligned}$$

by Equation (4.4):

$$= P(E_1) + P(E_2).$$

Assume $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$ is true for a natural number k. Then, for $n = k + 1$,

$$\begin{aligned} P(\bigcup_{i=1}^{\infty} E_i) &= P(\bigcup_{i=1}^{k+1} E_i) \\ &= P((\bigcup_{i=1}^{k} E_i) \cup E_{k+1}) \\ &= P(\bigcup_{i=1}^{k} E_i) + P(E_{k+1}) - P((\bigcup_{i=1}^{k} E_i) \cap E_{k+1}) \end{aligned}$$

by Equation (4.4):

$$= P(\bigcup_{i=1}^{k} E_i) + P(E_{k+1})$$

by the assumption $n = k$ is true:

$$= \sum_{i=1}^{k+1} P(E_i).$$

Thus, $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$ holds for any natural number n. $\square$

Bibliography

[Pic, 2019] (2019). Picard toolkit. `https://broadinstitute.github.io/picard/`.

[Abbaszadegan et al., 2008] Abbaszadegan, M. R., Moaven, O., Sima, H. R., Ghafarzadegan, K., A'rabi, A., Forghani, M. N., Raziee, H. R., Mashhadinejad, A., Jafarzadeh, M., Esmaili-Shandiz, E. et al. (2008). p16 promoter hypermethylation: a useful serum marker for early detection of gastric cancer. World Journal of Gastroenterology: WJG *14*, 2055.

[Acharjee et al., 2023] Acharjee, S., Chauhan, S., Pal, R. and Tomar, R. S. (2023). Mechanisms of DNA methylation and histone modifications. Progress in Molecular Biology and Translational Science *197*, 51–92.

[Affinito et al., 2020] Affinito, O., Palumbo, D., Fierro, A., Cuomo, M., De Riso, G., Monticelli, A., Miele, G., Chiariotti, L. and Cocozza, S. (2020). Nucleotide distance influences co-methylation between nearby CpG sites. Genomics *112*, 144–150.

[Akalin et al., 2012] Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A. and Mason, C. E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biology *13*, 1–9.

[Almamun et al., 2017] Almamun, M., Kholod, O., Stuckel, A. J., Levinson, B. T., Johnson, N. T., Arthur, G. L., Davis, J. W. and Taylor, K. H. (2017). Inferring a role for methylation of intergenic DNA in the regulation of genes aberrantly expressed in precursor B-cell acute lymphoblastic leukemia. Leukemia & Lymphoma *58*, 2156–2164.

[Ambrosi et al., 2017] Ambrosi, C., Manzo, M. and Baubec, T. (2017). Dynamics and context-dependent roles of DNA methylation. Journal of Molecular Biology *429*, 1459–1475.

[Anastasiadi et al., 2018] Anastasiadi, D., Esteve-Codina, A. and Piferrer, F. (2018). Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. Epigenetics & Chromatin *11*, 1–17.

[Angermueller et al., 2017] Angermueller, C., Lee, H. J., Reik, W. and Stegle, O. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learn-

ing. Genome Biology *18*, 1–13.

[Arneson et al., 2020] Arneson, D., Yang, X. and Wang, K. (2020). MethylResolver—a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents. Communications Biology *3*, 422.

[Atlasi and Stunnenberg, 2017] Atlasi, Y. and Stunnenberg, H. G. (2017). The interplay of epigenetic marks during stem cell differentiation and development. Nature Reviews Genetics *18*, 643–658.

[Avsec et al., 2021] Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P. and Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods *18*, 1196–1203.

[Azad et al., 2013] Azad, N., Zahnow, C. A., Rudin, C. M. and Baylin, S. B. (2013). The future of epigenetic therapy in solid tumours—lessons from the past. Nature Reviews Clinical Oncology *10*, 256–266.

[Bansal and Pinney, 2017] Bansal, A. and Pinney, S. E. (2017). DNA methylation and its role in the pathogenesis of diabetes. Pediatric Diabetes *18*, 167–177.

[Barrett et al., 2017] Barrett, J. E., Feber, A., Herrero, J., Tanic, M., Wilson, G. A., Swanton, C. and Beck, S. (2017). Quantification of tumour evolution and heterogeneity via Bayesian epiallele detection. BMC Bioinformatics *18*, 1–10.

[Basu and Tiwari, 2021] Basu, A. and Tiwari, V. K. (2021). Epigenetic reprogramming of cell identity: lessons from development for regenerative medicine. Clinical Epigenetics *13*, 1–11.

[Baum, 1968] Baum, L. E. (1968). Growth functions for trasformations on manifolds. Pacific Journal of Mathematics *27*, 211–227.

[Baum and Eagon, 1967] Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. Bulletin of the American Mathematical Society *73*, 360–363.

[Baum and Petrie, 1966] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. The Annals of Mathematical Statistics *37*, 1554–1563.

[Baum et al., 1970] Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics *41*, 164–171.

[Baum et al., 1972] Baum, L. E. et al. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities *3*, 1–8.

[Behjati and Tarpey, 2013] Behjati, S. and Tarpey, P. S. (2013). What is next generation sequencing? Archives of Disease in Childhood - Education and Practice *98*, 236–238.

[Beltagy et al., 2020] Beltagy, I., Peters, M. E. and Cohan, A. (2020). Longformer: The Long-Document Transformer. arXiv preprint *arXiv:2004.05150*.

[Bishop, 1995] Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford university press.

[Blagitko-Dorfs et al., 2019] Blagitko-Dorfs, N., Schlosser, P., Greve, G., Pfeifer, D., Meier, R., Baude, A., Brocks, D., Plass, C. and Lübbert, M. (2019). Combination treatment of acute myeloid leukemia cells with DNMT and HDAC inhibitors: predominant synergistic gene downregulation associated with gene body demethylation. Leukemia *33*, 945–956.

[Boks et al., 2009] Boks, M. P., Derks, E. M., Weisenberger, D. J., Strengman, E., Janson, E., Sommer, I. E., Kahn, R. S. and Ophoff, R. A. (2009). The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. PloS One *4*, e6767.

[Bowman and Shenton, 2007] Bowman, K. O. and Shenton, L. (2007). The beta distribution, moment method, Karl Pearson and RA Fisher. Far East Journal of Theoretical Statistics *23*, 133.

[Bridle, 1989] Bridle, J. (1989). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. Advances in Neural Information Processing Systems *2*.

[Brierley et al., 2016] Brierley, J., Gospodarowicz, M. and O'Sullivan, B. (2016). The principles of cancer staging. Cancer *10*.

[Capper et al., 2018] Capper, D., Jones, D. T., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D. E. et al. (2018). DNA methylation-based classification of central nervous system tumours. Nature *555*, 469–474.

[Carlson and Maintainer, 2015] Carlson, M. and Maintainer, B. P. (2015). TxDb. Hsapiens. UCSC. hg19. knownGene. R package *3.2.2*.

[Catoni et al., 2018] Catoni, M., Tsang, J. M., Greco, A. P. and Zabet, N. R. (2018). DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. Nucleic Acids Research *46*, e114–e114.

[Chakravarthy et al., 2018] Chakravarthy, A., Furness, A., Joshi, K., Ghorani, E., Ford, K., Ward, M. J., King, E. V., Lechner, M., Marafioti, T., Quezada, S. A. et al. (2018). Pan-cancer deconvolution of tumour composition using DNA methylation. Nature Communications *9*, 3220.

[Cheng et al., 2019] Cheng, Y., He, C., Wang, M., Ma, X., Mo, F., Yang, S., Han, J. and Wei, X. (2019). Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. Signal Transduction and Targeted Therapy *4*, 62.

[Cho et al., 2014] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint *arXiv:1406.1078*.

[Chu et al., 2022] Chu, T., Wang, Z., Pe'er, D. and Danko, C. G. (2022). Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. Nature Cancer *3*, 505–517.

[Chuanben et al., 2018] Chuanben, C., Zhaodong, F., Chaoxiong, H., Jianming, D. and Lisha, C. (2018). Prognostic value of tumor burden in nasopharyngeal carcinoma. Cancer Management and Research *10*, 3169–3175.

[Clark et al., 2019] Clark, K., Khandelwal, U., Levy, O. and Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. arXiv preprint *arXiv:1906.04341*.

[Cormen et al., 2022] Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C. (2022). Introduction to algorithms. MIT press.

[Crowder, 1978] Crowder, M. J. (1978). Beta-binomial ANOVA for proportions. pp. 34–37, JSTOR.

[Cuadrat et al., 2021] Cuadrat, R. R., Kratzer, A., Arnal, H. G., Wreczycka, K., Blume, A., Ebenal, V., Mauno, T., Hartung, J., Seppelt, C., Meteva, D. et al. (2021). Cardiovascular disease biomarkers derived from circulating cell-free DNA methylation. pp. 2021–11, Cold Spring Harbor Laboratory Press.

[Danecek et al., 2021] Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M. et al. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*, giab008.

[Das and Singal, 2004] Das, P. M. and Singal, R. (2004). DNA methylation and cancer. Journal of Clinical Oncology *22*, 4632–4642.

[De Mattos-Arruda et al., 2013] De Mattos-Arruda, L., Cortes, J., Santarpia, L., Vivancos, A., Tabernero, J., Reis-Filho, J. S. and Seoane, J. (2013). Circulating tumour cells and cell-free DNA as tools for managing breast cancer. Nature Reviews Clinical Oncology *10*, 377–389.

[de Ridder et al., 2005] de Ridder, D., Van Der Linden, C., Schonewille, T., Dik, W., Reinders, M., Van Dongen, J. and Staal, F. (2005). Purity for clarity: the need for purification of tumor cells in DNA microarray studies. Leukemia *19*, 618–627.

[De Waele et al., 2022] De Waele, G., Clauwaert, J., Menschaert, G. and Waegeman, W. (2022). CpG Transformer for imputation of single-cell methylomes. Bioinformatics *38*,

597–603.

[Decamps et al., 2020] Decamps, C., Privé, F., Bacher, R., Jost, D., Waguet, A., Houseman, E. A., Lurie, E., Lutsik, P., Milosavljevic, A. et al. (2020). Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation deconvolution software. BMC Bioinformatics *21*, 1–15.

[Deng et al., 2021] Deng, Y., Song, Z., Huang, L., Guo, Z., Tong, B., Sun, M., Zhao, J., Zhang, H., Zhang, Z. and Li, G. (2021). Tumor purity as a prognosis and immunotherapy relevant feature in cervical cancer. Aging *13*, 24768.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint *arXiv:1810.04805*.

[D'haeseleer, 2006] D'haeseleer, P. (2006). What are DNA sequence motifs? Nature Biotechnology *24*, 423–425.

[Do et al., 2020] Do, C., Dumont, E. L., Salas, M., Castano, A., Mujahed, H., Maldonado, L., Singh, A., DaSilva-Arnold, S. C., Bhagat, G., Lehman, S. et al. (2020). Allele-specific DNA methylation is increased in cancers and its dense mapping in normal plus neoplastic cells increases the yield of disease-associated regulatory SNPs. Genome Biology *21*, 1–39.

[Doane and Seward, 2011] Doane, D. P. and Seward, L. E. (2011). Measuring skewness: a forgotten statistic? Journal of Statistics Education *19*.

[Dolzhenko and Smith, 2014] Dolzhenko, E. and Smith, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinformatics *15*, 1–8.

[Dong et al., 2018] Dong, L., Xu, S. and Xu, B. (2018). Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 5884–5888, IEEE.

[Dorri et al., 2016] Dorri, F., Mendelowitz, L. and Corrada Bravo, H. (2016). methylFlow: cell-specific methylation pattern reconstruction from high-throughput bisulfite-converted DNA sequencing. Bioinformatics *32*, 1618–1624.

[Duffy and Crown, 2022] Duffy, M. J. and Crown, J. (2022). Circulating tumor DNA as a biomarker for monitoring patients with solid cancers: Comparison with standard protein biomarkers. Clinical Chemistry *68*, 1381–1390.

[Egyud et al., 2019] Egyud, M., Tejani, M., Pennathur, A., Luketich, J., Sridhar, P., Yamada, E., Ståhlberg, A., Filges, S., Krzyzanowski, P., Jackson, J. et al. (2019). Detection of circulating tumor DNA in plasma: a potential biomarker for esophageal adenocarcinoma. The Annals of Thoracic Surgery *108*, 343–349.

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

[Ehrlich, 2009] Ehrlich, M. (2009). DNA hypomethylation in cancer cells. Epigenomics *1*, 239–259.

[Erhan et al., 2009] Erhan, D., Manzagol, P.-A., Bengio, Y., Bengio, S. and Vincent, P. (2009). The difficulty of training deep architectures and the effect of unsupervised pre-training. In Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics vol. 5, of Proceedings of Machine Learning Research pp. 153–160, PMLR.

[Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd vol. 96, pp. 226–231,.

[Ewing and Green, 1998] Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research *8*, 186–194.

[Farrell, 2011] Farrell, A. (2011). Encyclopedia of fish physiology: from genome to environment. Academic press.

[Favorov et al., 2012] Favorov, A., Mularoni, L., Cope, L. M., Medvedeva, Y., Mironov, A. A., Makeev, V. J. and Wheelan, S. J. (2012). Exploring massive, genome scale datasets with the GenometriCorr package. PLoS Computational Biology *8*, e1002529.

[Felix and Cecil, 2019] Felix, J. and Cecil, C. A. (2019). Population DNA methylation studies in the Developmental Origins of Health and Disease (DOHaD) framework. Journal of Developmental Origins of Health and Disease *10*, 306–313.

[Feng et al., 2014] Feng, H., Conneely, K. N. and Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Research *42*, e69–e69.

[Ferreira et al., 2019] Ferreira, L. J., Donoghue, M. T., Barros, P., Saibo, N. J., Santos, A. P. and Oliveira, M. M. (2019). Uncovering differentially methylated regions (DMRs) in a salt-tolerant rice variety under stress: one step towards new regulatory regions for enhanced salt tolerance. Epigenomes *3*, 4.

[Fiala and Diamandis, 2018] Fiala, C. and Diamandis, E. P. (2018). Utility of circulating tumor DNA in cancer diagnostics with emphasis on early detection. BMC Medicine *16*, 1–10.

[Fong et al., 2021] Fong, J., Gardner, J. R., Andrews, J. M., Cashen, A. F., Payton, J. E., Weinberger, K. Q. and Edwards, J. R. (2021). Determining subpopulation methylation profiles from bisulfite sequencing data of heterogeneous samples using DXM. Nucleic Acids Research *49*, e93–e93.

[Fujita et al., 2022] Fujita, K., Okada, K. and Katahira, K. (2022). The Fisher information matrix: A tutorial for calculation for decision making models. PsyArXiv.

[Goh et al., 2017] Goh, G. B., Hodas, N. O. and Vishnu, A. (2017). Deep learning for computational chemistry. Journal of Computational Chemistry *38*, 1291–1307.

[Goldberg et al., 2007] Goldberg, A. D., Allis, C. D. and Bernstein, E. (2007). Epigenetics: a landscape takes shape. Cell *128*, 635–638.

[Goldman et al., 2019] Goldman, S. L., MacKay, M., Afshinnekoo, E., Melnick, A. M., Wu, S. and Mason, C. E. (2019). The impact of heterogeneity on single-cell sequencing. Frontiers in Genetics *10*, 8.

[Gosselt et al., 2021] Gosselt, H. R., Griffioen, P. H., van Zelst, B. D., Oosterom, N., de Jonge, R. and Heil, S. G. (2021). Global DNA (hydroxy) methylation is stable over time under several storage conditions and temperatures. Epigenetics *16*, 45–53.

[Graves et al., 2013] Graves, A., Mohamed, A.-r. and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing pp. 6645–6649, IEEE.

[Greenberg and Bourc'his, 2019] Greenberg, M. V. and Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. Nature Reviews Molecular Cell Biology *20*, 590–607.

[Guo et al., 2019] Guo, M., Peng, Y., Gao, A., Du, C. and Herman, J. G. (2019). Epigenetic heterogeneity in cancer. Biomarker Research *7*, 1–19.

[Gwak and Rho, 2022] Gwak, H.-J. and Rho, M. (2022). ViBE: a hierarchical BERT model to identify eukaryotic viruses using metagenome sequencing data. Briefings in Bioinformatics *23*, bbac204.

[Hall et al., 2008] Hall, G. L., Shaw, R. J., Field, E. A., Rogers, S. N., Sutton, D. N., Woolgar, J. A., Lowe, D., Liloglou, T., Field, J. K. and Risk, J. M. (2008). p16 Promoter methylation is a potential predictor of malignant transformation in oral epithelial dysplasia. Cancer Epidemiology Biomarkers & Prevention *17*, 2174–2179.

[Han and He, 2016] Han, Y. and He, X. (2016). Integrating epigenomics into the understanding of biomedical insight. Bioinformatics and Biology Insights *10*, BBI–S38427.

[Hengartner, 2000] Hengartner, M. O. (2000). The biochemistry of apoptosis. Nature *407*, 770–776.

[Ho et al., 2021] Ho, S.-Y., Liu, P.-H., Hsu, C.-Y., Ko, C.-C., Huang, Y.-H., Su, C.-W., Lee, R.-C., Tsai, P.-H., Hou, M.-C. and Huo, T.-I. (2021). Tumor burden score as a new prognostic marker for patients with hepatocellular carcinoma undergoing transarterial chemoembolization. Journal of Gastroenterology and Hepatology *36*, 3196–3203.

[Houseman et al., 2012] Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K. and Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics *13*, 1–16.

[Houseman et al., 2016] Houseman, E. A., Kile, M. L., Christiani, D. C., Ince, T. A., Kelsey, K. T. and Marsit, C. J. (2016). Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC Bioinformatics *17*, 1–15.

[Hui et al., 2018] Hui, T., Cao, Q., Wegrzyn-Woltosz, J., O'Neill, K., Hammond, C. A., Knapp, D. J., Laks, E., Moksa, M., Aparicio, S., Eaves, C. J. et al. (2018). High-resolution single-cell DNA methylation measurements reveal epigenetically distinct hematopoietic stem cell subpopulations. Stem Cell Reports *11*, 578–592.

[Humeau-Heurtier, 2018] Humeau-Heurtier, A. (2018). Evaluation of systems' irregularity and complexity: Sample entropy, its derivatives, and their applications across scales and disciplines. Entropy *20*, 794.

[Inouye et al., 1991] Inouye, T., Shinosaki, K., Sakamoto, H., Toi, S., Ukai, S., Iyama, A., Katsuda, Y. and Hirano, M. (1991). Quantification of EEG irregularity by use of the entropy of the power spectrum. Electroencephalography and Clinical Neurophysiology *79*, 204–210.

[Jang et al., 2020] Jang, B. G., Kim, H. S., Bae, J. M., Kim, W. H., Kim, H. U. and Kang, G. H. (2020). SMOC2, an intestinal stem cell marker, is an independent prognostic marker associated with better survival in colorectal cancers. Scientific Reports *10*, 14591.

[Jawahar et al., 2019] Jawahar, G., Sagot, B. and Seddah, D. (2019). What does BERT learn about the structure of language? In ACL 2019-57th Annual Meeting of the Association for Computational Linguistics.

[Jeong et al., 2022] Jeong, Y., de Andrade e Sousa, L. B., Thalmeier, D., Toth, R., Ganslmeier, M., Breuer, K., Plass, C. and Lutsik, P. (2022). Systematic evaluation of cell-type deconvolution pipelines for sequencing-based bulk DNA methylomes. Briefings in Bioinformatics *23*, bbac248.

[Jeong et al., 2023a] Jeong, Y., Rohr, K. and Lutsik, P. (2023a). MethylBERT: A Transformer-based model for read-level DNA methylation pattern identification and tumour deconvolution. Cold Spring Harbor Laboratory.

[Jeong et al., 2023b] Jeong, Y., Ronen, J., Kopp, W., Lutsik, P. and Akalin, A. (2023b). Decoding single-cell multiomics: scMaui-A deep learning framework for uncovering cellular heterogeneity in presence of batch effects and missing data. pp. 2023–01, Cold Spring Harbor Laboratory.

[Jeschke et al., 2017] Jeschke, J., Bizet, M., Desmedt, C., Calonne, E., Dedeurwaerder, S., Garaud, S., Koch, A., Larsimont, D., Salgado, R., Van den Eynden, G. et al. (2017).

DNA methylation–based immune response signature improves patient diagnosis in multiple cancers. The Journal of Clinical Investigation *127*, 3090–3102.

[Ji et al., 2021] Ji, Y., Zhou, Z., Liu, H. and Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics *37*, 2112–2120.

[Joanes and Gill, 1998] Joanes, D. N. and Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. Journal of the Royal Statistical Society: Series D (The Statistician) *47*, 183–189.

[Joo et al., 2018] Joo, J. E., Dowty, J. G., Milne, R. L., Wong, E. M., Dugué, P.-A., English, D., Hopper, J. L., Goldgar, D. E., Giles, G. G., Southey, M. C. et al. (2018). Heritable DNA methylation marks associated with susceptibility to breast cancer. Nature Communications *9*, 867.

[Jumper et al., 2021] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

[Kamilaris and Prenafeta-Boldú, 2018] Kamilaris, A. and Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture *147*, 70–90.

[Kandimalla et al., 2021] Kandimalla, R., Xu, J., Link, A., Matsuyama, T., Yamamura, K., Parker, M. I., Uetake, H., Balaguer, F., Borazanci, E., Tsai, S. et al. (2021). EpiPanGI Dx: a cell-free DNA methylation fingerprint for the early detection of gastrointestinal cancers. Clinical Cancer Research *27*, 6135–6144.

[Kapourani and Sanguinetti, 2019] Kapourani, C.-A. and Sanguinetti, G. (2019). Melissa: Bayesian clustering and imputation of single-cell methylomes. Genome Biology *20*, 61.

[Katsman et al., 2022] Katsman, E., Orlanski, S., Martignano, F., Fox-Fisher, I., Shemer, R., Dor, Y., Zick, A., Eden, A., Petrini, I., Conticello, S. G. et al. (2022). Detecting cell-of-origin and cancer-specific methylation features of cell-free DNA from Nanopore sequencing. Genome Biology *23*, 1–25.

[Kim and Kim, 2016] Kim, S. and Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. International Journal of Forecasting *32*, 669–679.

[Kim et al., 2021] Kim, S. I., Cassella, C. R. and Byrne, K. T. (2021). Tumor burden and immunotherapy: impact on immune infiltration and therapeutic outcomes. Frontiers in Immunology *11*, 629722.

[Klco et al., 2013] Klco, J. M., Spencer, D. H., Lamprecht, T. L., Sarkaria, S. M., Wylie, T., Magrini, V., Hundal, J., Walker, J., Varghese, N., Erdmann-Gilmore, P. et al.

(2013). Genomic impact of transient low-dose decitabine treatment on primary AML cells. Blood, The Journal of the American Society of Hematology *121*, 1633–1643.

[Kolmogorov, 1933] Kolmogorov, A. N. (1933). Foundations of the theory of Probability, 2nd English edition.

[Kretzmer et al., 2015] Kretzmer, H., Bernhart, S. H., Wang, W., Haake, A., Weniger, M. A., Bergmann, A. K., Betts, M. J., Carrillo-de Santa-Pau, E., Doose, G., Gutwein, J. et al. (2015). DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. Nature Genetics *47*, 1316–1325.

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

[Kulis and Esteller, 2010] Kulis, M. and Esteller, M. (2010). DNA methylation and cancer. Advances in Genetics *70*, 27–56.

[Kyriakopoulos et al., 2019] Kyriakopoulos, C., Giehr, P., Lück, A., Walter, J. and Wolf, V. (2019). A hybrid HMM approach for the dynamics of DNA methylation. In Hybrid Systems Biology: 6th International Workshop, HSB 2019, Prague, Czech Republic, April 6-7, 2019, Revised Selected Papers 6 pp. 117–131, Springer.

[Lawrence et al., 2013] Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T. and Carey, V. J. (2013). Software for computing and annotating genomic ranges. PLoS Computational Biology *9*, e1003118.

[Lee et al., 2019] Lee, D., Lee, S. and Kim, S. (2019). PRISM: methylation pattern-based, reference-free inference of subclonal makeup. Bioinformatics *35*, i520–i529.

[Leung et al., 2020] Leung, J. Y., Chia, K., Ong, D. S. T. and Taneja, R. (2020). Interweaving tumor heterogeneity into the cancer epigenetic/metabolic axis. Antioxidants & Redox Signaling *33*, 946–965.

[Li et al., 2021] Li, J., Wei, L., Zhang, X., Zhang, W., Wang, H., Zhong, B., Xie, Z., Lv, H. and Wang, X. (2021). DISMIR: Deep learning-based noninvasive cancer detection by integrating DNA sequence and methylation information of individual cell-free DNA reads. Briefings in Bioinformatics *22*, bbab250.

[Li et al., 2016] Li, S., Garrett-Bakelman, F. E., Chung, S. S., Sanders, M. A., Hricik, T., Rapaport, F., Patel, J., Dillon, R., Vijay, P., Brown, A. L. et al. (2016). Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. Nature Medicine *22*, 792–799.

[Li et al., 2018] Li, W., Li, Q., Kang, S., Same, M., Zhou, Y., Sun, C., Liu, C.-C., Matsuoka, L., Sher, L., Wong, W. H. et al. (2018). CancerDetector: ultrasensitive and

non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. Nucleic Acids Research *46*, e89–e89.

[Lietz et al., 2022] Lietz, C. E., Newman, E. T., Kelly, A. D., Xiang, D. H., Zhang, Z., Luscko, C. A., Lozano-Calderon, S. A., Ebb, D. H., Raskin, K. A., Cote, G. M. et al. (2022). Genome-wide DNA methylation patterns reveal clinically relevant predictive and prognostic subtypes in human osteosarcoma. Communications Biology *5*, 213.

[Lin et al., 2022] Lin, J., Nogueira, R. and Yates, A. (2022). Pretrained transformers for text ranking: Bert and beyond. Springer Nature.

[Lipton et al., 2015] Lipton, Z. C., Berkowitz, J. and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. arXiv preprint *arXiv:1506.00019*.

[Liu et al., 2021] Liu, Q., Chen, J., Wang, Y., Li, S., Jia, C., Song, J. and Li, F. (2021). DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. Briefings in Bioinformatics *22*, bbaa124.

[Liu et al., 2019] Liu, Q., Fang, L., Yu, G., Wang, D., Xiao, C.-L. and Wang, K. (2019). Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nature Communications *10*, 2449.

[Liu et al., 2013] Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M. et al. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nature Biotechnology *31*, 142–147.

[Liu et al., 2019] Liu, Y., Gu, Y., Su, M., Liu, H., Zhang, S. and Zhang, Y. (2019). An analysis about heterogeneity among cancers based on the DNA methylation patterns. BMC Cancer *19*, 1–15.

[Loman et al., 2015] Loman, N. J., Quick, J. and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. Nature Methods *12*, 733–735.

[Loshchilov and Hutter, 2017] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint *arXiv:1711.05101*.

[Loyfer et al., 2023] Loyfer, N., Magenheim, J., Peretz, A., Cann, G., Bredno, J., Klochendler, A., Fox-Fisher, I., Shabi-Porat, S., Hecht, M., Pelet, T. et al. (2023). A DNA methylation atlas of normal human cell types. Nature *613*, 355–364.

[Luo et al., 2017] Luo, C., Keown, C. L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J. R., Sandoval, J. P. et al. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science *357*, 600–604.

[Lutsik et al., 2017] Lutsik, P., Slawski, M., Gasparoni, G., Vedeneev, N., Hein, M. and Walter, J. (2017). MeDeCom: discovery and quantification of latent components of

heterogeneous methylomes. Genome Biology *18*, 1–20.

[Lv et al., 2020] Lv, H., Dao, F.-Y., Zhang, D., Guan, Z.-X., Yang, H., Su, W., Liu, M.-L., Ding, H., Chen, W. and Lin, H. (2020). iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. Iscience *23*.

[Martin, 2011] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12.

[Martins et al., 2021] Martins, I., Ribeiro, I. P., Jorge, J., Gonçalves, A. C., Sarmento-Ribeiro, A. B., Melo, J. B. and Carreira, I. M. (2021). Liquid biopsies: applications for cancer diagnosis and monitoring. Genes *12*, 349.

[Maruyama et al., 2022] Maruyama, O., Li, Y., Narita, H., Toh, H., Au Yeung, W. K. and Sasaki, H. (2022). CMIC: predicting DNA methylation inheritance of CpG islands with embedding vectors of variable-length k-mers. BMC Bioinformatics *23*, 1–20.

[Mayakonda et al., 2020] Mayakonda, A., Schönung, M., Hey, J., Batra, R. N., Feuerstein-Akgoz, C., Köhler, K., Lipka, D. B., Sotillo, R., Plass, C., Lutsik, P. et al. (2020). Methrix: an R/Bioconductor package for systematic aggregation and analysis of bisulfite sequencing data. Bioinformatics *36*, 5524–5525.

[McCabe et al., 2009] McCabe, M. T., Brandes, J. C. and Vertino, P. M. (2009). Cancer DNA methylation: molecular mechanisms and clinical implications. Clinical Cancer Research *15*, 3927–3937.

[Melamed et al., 2018] Melamed, P., Yosefzon, Y., David, C., Tsukerman, A. and Pnueli, L. (2018). Tet enzymes, variants, and differential effects on function. Frontiers in Cell and Developmental Biology *6*, 22.

[Meyer et al., 2021] Meyer, B., Clifton, S., Locke, W., Luu, P.-L., Du, Q., Lam, D., Armstrong, N. J., Kumar, B., Deng, N., Harvey, K. et al. (2021). Identification of DNA methylation biomarkers with potential to predict response to neoadjuvant chemotherapy in triple-negative breast cancer. Clinical Epigenetics *13*, 1–7.

[Min et al., 2017] Min, S., Lee, B. and Yoon, S. (2017). Deep learning in bioinformatics. Briefings in Bioinformatics *18*, 851–869.

[Newman et al., 2019] Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D. et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology *37*, 773–782.

[Ni et al., 2019] Ni, P., Huang, N., Zhang, Z., Wang, D.-P., Liang, F., Miao, Y., Xiao, C.-L., Luo, F. and Wang, J. (2019). DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. Bioinformatics *35*, 4586–4595.

[Nishiyama and Nakanishi, 2021] Nishiyama, A. and Nakanishi, M. (2021). Navigating the DNA methylation landscape of cancer. Trends in Genetics *37*, 1012–1027.

[Orkin and Zon, 2008] Orkin, S. H. and Zon, L. I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. Cell *132*, 631–644.

[Pantel and Alix-Panabières, 2017] Pantel, K. and Alix-Panabières, C. (2017). Circulating tumour cells and cell-free DNA in gastrointestinal cancer. Nature Reviews Gastroenterology & Hepatology *14*, 73–74.

[Park and Wu, 2016] Park, Y. and Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics *32*, 1446–1453.

[Pascanu et al., 2013] Pascanu, R., Mikolov, T. and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In International Conference on Machine Learning pp. 1310–1318, PMLR.

[Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems *32*.

[Peters et al., 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) pp. 2227–2237, Association for Computational Linguistics.

[Petralia et al., 2021] Petralia, F., Krek, A., Calinawan, A. P., Feng, S., Gosline, S., Pugliese, P., Ceccarelli, M. and Wang, P. (2021). BayesDeBulk: a flexible Bayesian algorithm for the deconvolution of bulk tumor data. pp. 2021–06, Cold Spring Harbor Laboratory.

[Phillips et al., 2008] Phillips, T. et al. (2008). The role of methylation in gene expression. Nature Education *1*, 116.

[Plass et al., 2013] Plass, C., Pfister, S. M., Lindroth, A. M., Bogatyrova, O., Claus, R. and Lichter, P. (2013). Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. Nature Reviews Genetics *14*, 765–780.

[Portela and Esteller, 2010] Portela, A. and Esteller, M. (2010). Epigenetic modifications and human disease. Nature Biotechnology *28*, 1057–1068.

[Prince et al., 2007] Prince, M., Sivanandan, R., Kaczorowski, A., Wolf, G., Kaplan, M., Dalerba, P., Weissman, I., Clarke, M. and Ailles, L. (2007). Identification of a subpopulation of cells with cancer stem cell properties in head and neck squamous cell carcinoma. Proceedings of the National Academy of Sciences *104*, 973–978.

[Pujadas and Feinberg, 2012] Pujadas, E. and Feinberg, A. P. (2012). Regulated noise in the epigenetic landscape of development and disease. Cell *148*, 1123–1131.

[Radford et al., 2018] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. et al. (2018). Improving language understanding by generative pre-training. OpenAI.

[Rahmani et al., 2019] Rahmani, E., Schweiger, R., Rhead, B., Criswell, L. A., Barcellos, L. F., Eskin, E., Rosset, S., Sankararaman, S. and Halperin, E. (2019). Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. Nature Communications *10*, 3417.

[Rajpurkar et al., 2018] Rajpurkar, P., Jia, R. and Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. arXiv preprint *arXiv:1806.03822*.

[Rajpurkar et al., 2016] Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint *arXiv:1606.05250*.

[Rand et al., 2017] Rand, A. C., Jain, M., Eizenga, J. M., Musselman-Brown, A., Olsen, H. E., Akeson, M. and Paten, B. (2017). Mapping DNA methylation with high-throughput nanopore sequencing. Nature Methods *14*, 411–413.

[Rasmi et al., 2023] Rasmi, Y., Shokati, A., Hassan, A., Aziz, S. G.-G., Bastani, S., Jalali, L., Moradi, F. and Alipour, S. (2023). The role of DNA methylation in progression of neurological disorders and neurodegenerative diseases as well as the prospect of using DNA methylation inhibitors as therapeutic agents for such disorders. IBRO Neuroscience Reports *14*, 28–37.

[Reid and Fridley, 2020] Reid, B. M. and Fridley, B. L. (2020). DNA methylation in ovarian cancer susceptibility. Cancers *13*, 108.

[Richard and Lippmann, 1991] Richard, M. D. and Lippmann, R. P. (1991). Neural network classifiers estimate Bayesian a posteriori probabilities. Neural Computation *3*, 461–483.

[Rudner et al., 1968] Rudner, R., Karkas, J. D. and Chargaff, E. (1968). Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. Proceedings of the National Academy of Sciences *60*, 921–922.

[Saerens et al., 2002] Saerens, M., Latinne, P. and Decaestecker, C. (2002). Any reasonable cost function can be used for a posteriori probability approximation. IEEE Transactions on Neural Networks *13*, 1204–1210.

[Saghafinia et al., 2018] Saghafinia, S., Mina, M., Riggi, N., Hanahan, D. and Ciriello, G. (2018). Pan-cancer landscape of aberrant DNA methylation across human tumors. Cell Reports *25*, 1066–1080.

[Sakamoto et al., 2020] Sakamoto, Y., Sereewattanawoot, S. and Suzuki, A. (2020). A new era of long-read sequencing for cancer genomics. Journal of Human Genetics *65*, 3–10.

[Salvianti et al., 2021] Salvianti, F., Gelmini, S., Mancini, I., Pazzagli, M., Pillozzi, S., Giommoni, E., Brugia, M., Di Costanzo, F., Galardi, F., De Luca, F. et al. (2021). Circulating tumour cells and cell-free DNA as a prognostic factor in metastatic colorectal cancer: the OMITERC prospective study. British Journal of Cancer *125*, 94–100.

[Sangaletti et al., 2020] Sangaletti, S., Iannelli, F., Zanardi, F., Cancila, V., Portararo, P., Botti, L., Vacca, D., Chiodoni, C., Di Napoli, A., Valenti, C. et al. (2020). Intratumour heterogeneity of diffuse large B-cell lymphoma involves the induction of diversified stroma-tumour interfaces. EBioMedicine *61*.

[Sarzynska-Wawer et al., 2021] Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M. and Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. Psychiatry Research *304*, 114135.

[Scherer et al., 2020] Scherer, M., Nazarov, P. V., Toth, R., Sahay, S., Kaoma, T., Maurer, V., Vedeneev, N., Plass, C., Lengauer, T., Walter, J. et al. (2020). Reference-free deconvolution, visualization and interpretation of complex DNA methylation data using DecompPipeline, MeDeCom and FactorViz. Nature Protocols *15*, 3240–3263.

[Schmid et al., 1999] Schmid, M. H., Bird, P., Dummer, R., Kempf, W. and Burg, G. (1999). Tumor burden index as a prognostic tool for cutaneous T-cell lymphoma: a new concept. Archives of Dermatology *135*, 1204–1208.

[Schmidt et al., 2010] Schmidt, B., Liebenberg, V., Dietrich, D., Schlegel, T., Kneip, C., Seegebarth, A., Flemming, N., Seemann, S., Distler, J., Lewin, J. et al. (2010). SHOX2 DNA methylation is a biomarker for the diagnosis of lung cancer based on bronchial aspirates. BMC Cancer *10*, 1–9.

[Schönung et al., 2021] Schönung, M., Meyer, J., Nöllke, P., Olshen, A. B., Hartmann, M., Murakami, N., Wakamatsu, M., Okuno, Y., Plass, C., Loh, M. L. et al. (2021). International consensus definition of DNA methylation subgroups in juvenile myelomonocytic leukemia. Clinical Cancer Research *27*, 158–168.

[Scott et al., 2020] Scott, C. A., Duryea, J. D., MacKay, H., Baker, M. S., Laritsky, E., Gunasekara, C. J., Coarfa, C. and Waterland, R. A. (2020). Identification of cell type-specific methylation signals in bulk whole genome bisulfite sequencing data. Genome Biology *21*, 1–23.

[Sen et al., 2021] Sen, M., Mooijman, D., Chialastri, A., Boisset, J.-C., Popovic, M., Heindryckx, B., Chuva de Sousa Lopes, S. M., Dey, S. S. and van Oudenaarden, A. (2021). Strand-specific single-cell methylomics reveals distinct modes of DNA demethy-

lation dynamics during early mammalian development. Nature Communications *12*, 1286.

[Sheffield et al., 2017] Sheffield, N. C., Pierron, G., Klughammer, J., Datlinger, P., Schönegger, A., Schuster, M., Hadler, J., Surdez, D., Guillemot, D., Lapouble, E. et al. (2017). DNA methylation heterogeneity defines a disease spectrum in Ewing sarcoma. Nature Medicine *23*, 386–395.

[Shu et al., 2020] Shu, C., Zhang, X., Aouizerat, B. E. and Xu, K. (2020). Comparison of methylation capture sequencing and Infinium MethylationEPIC array in peripheral blood mononuclear cells. Epigenetics & Chromatin *13*, 1–15.

[Sill et al., 2020] Sill, M., Plass, C., Pfister, S. M. and Lipka, D. B. (2020). Molecular tumor classification using DNA methylome analysis. Human Molecular Genetics *29*, R205–R213.

[Singh et al., 2021] Singh, O., Pratt, D. and Aldape, K. (2021). 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*, 1–17.

[Smallwood et al., 2014] Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W. and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nature Methods *11*, 817–820.

[Sollier et al., 2023] Sollier, E., Kuipers, J., Takahashi, K., Beerenwinkel, N. and Jahn, K. (2023). COMPASS: joint copy number and mutation phylogeny reconstruction from amplicon single-cell sequencing data. Nature Communications *14*, 4921.

[Song and Kuan, 2022] Song, J. and Kuan, P.-F. (2022). A systematic assessment of cell type deconvolution algorithms for DNA methylation data. Briefings in Bioinformatics *23*, bbac449.

[Sontag et al., 2006] Sontag, L. B., Lorincz, M. C. and Luebeck, E. G. (2006). Dynamics, stability and inheritance of somatic DNA methylation imprints. Journal of Theoretical Biology *242*, 890–899.

[Stankovic et al., 2019] Stankovic, B., Bjørhovde, H. A. K., Skarshaug, R., Aamodt, H., Frafjord, A., Müller, E., Hammarström, C., Beraki, K., Bækkevold, E. S., Woldbæk, P. R. et al. (2019). Immune cell composition in human non-small cell lung cancer. Frontiers in Immunology *9*, 3101.

[Stanojević et al., 2022] Stanojević, D., Li, Z., Foo, R. and Šikić, M. (2022). Rockfish: A Transformer-based Model for Accurate 5-Methylcytosine Prediction from Nanopore Sequencing. pp. 2022–11, Cold Spring Harbor Laboratory.

[Sturm et al., 2019] Sturm, G., Finotello, F., Petitprez, F., Zhang, J. D., Baumbach, J., Fridman, W. H., List, M. and Aneichyk, T. (2019). Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics *35*, i436–i445.

[Subakan et al., 2021] Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M. and Zhong, J. (2021). Attention is all you need in speech separation. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 21–25, IEEE.

[Sun and Yu, 2016] Sun, S. and Yu, X. (2016). HMM-Fisher: identifying differential methylation using a hidden Markov model and Fisher's exact test. Statistical Applications in Genetics and Molecular Biology *15*, 55–67.

[Suzuki and Bird, 2008] Suzuki, M. M. and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nature Reviews Genetics *9*, 465–476.

[Tang et al., 2021] Tang, T., Huang, X., Zhang, G., Hong, Z., Bai, X. and Liang, T. (2021). Advantages of targeting the tumor immune microenvironment over blocking immune checkpoint in cancer immunotherapy. Signal Transduction and Targeted Therapy *6*, 72.

[Taudt et al., 2018] Taudt, A., Roquis, D., Vidalis, A., Wardenaar, R., Johannes, F. and Colomé-Tatché, M. (2018). METHimpute: imputation-guided construction of complete methylomes from WGBS data. BMC Genomics *19*, 1–14.

[Team and Maintainer, 2020] Team, B. and Maintainer, B. (2020). TxDb. Mmusculus. UCSC. mm10. knownGene: Annotation Package for TxDb Object (s). R package *3.10.0*.

[Titus et al., 2017a] Titus, A. J., Gallimore, R. M., Salas, L. A. and Christensen, B. C. (2017a). Cell-type deconvolution from DNA methylation: a review of recent applications. Human Molecular Genetics *26*, R216–R224.

[Titus et al., 2017b] Titus, A. J., Way, G. P., Johnson, K. C. and Christensen, B. C. (2017b). Deconvolution of DNA methylation identifies differentially methylated gene regions on 1p36 across breast cancer subtypes. Scientific Reports *7*, 11594.

[Vandiver et al., 2015] Vandiver, A. R., Idrizi, A., Rizzardi, L., Feinberg, A. P. and Hansen, K. D. (2015). DNA methylation is stable during replication and cell cycle arrest. Scientific Reports *5*, 17911.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems *30*.

[Veganzones-de Castro et al., 2012] Veganzones-de Castro, S., Rafael-Fernández, S., Vidaurreta-Lázaro, M., de-la Orden, V., Mediero-Valeros, B., Fernández, C. and Maestro-de las Casas, M. L. (2012). p16 gene methylation in colorectal cancer pa-

tients with long-term follow-up. Revista Espanola de Enfermedades Digestivas *104*, 111.

[Videtic Paska and Hudler, 2015] Videtic Paska, A. and Hudler, P. (2015). Aberrant methylation patterns in cancer: a clinical view. Biochemia Medica *25*, 161–176.

[Waas and Kislinger, 2020] Waas, M. and Kislinger, T. (2020). Addressing cellular heterogeneity in cancer through precision proteomics. Journal of Proteome Research *19*, 3607–3619.

[Waddington, 2014] Waddington, C. H. (2014). The strategy of the genes. Routledge.

[Wagner and Rohr, 2022] Wagner, R. and Rohr, K. (2022). Cellcentroidformer: Combining self-attention and convolution for cell detection. In Annual Conference on Medical Image Understanding and Analysis pp. 212–222, Springer.

[Waldmann and Schneider, 2013] Waldmann, T. and Schneider, R. (2013). Targeting histone modifications—epigenetics in cancer. Current Opinion in Cell Biology *25*, 184–189.

[Wang et al., 2018] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint *arXiv:1804.07461*.

[Wang et al., 2022] Wang, F., Yang, F., Huang, L., Song, J., Gasser, R. B., Aebersold, R., Wang, G. and Yao, J. (2022). Deep Domain Adversarial Neural Network for the Deconvolution of Cell Type Mixtures in Tissue Proteome Profiling. pp. 2022–11, Cold Spring Harbor Laboratory.

[Wang et al., 2012] Wang, L., Tang, L., Xie, R., Nie, W., Chen, L. and Guan, X. (2012). p16 promoter hypermethylation is associated with increased breast cancer risk. Molecular Medicine Reports *6*, 904–908.

[Wang et al., 2023] Wang, X., Ahsan, M. U., Zhou, Y. and Wang, K. (2023). Transformer-based DNA methylation detection on ionic signals from Oxford Nanopore sequencing data. Quantitative Biology *11*, 287–296.

[Wang et al., 2019] Wang, X., Park, J., Susztak, K., Zhang, N. R. and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature Communications *10*, 380.

[Wang et al., 2020] Wang, Z., Yin, J., Zhou, W., Bai, J., Xie, Y., Xu, K., Zheng, X., Xiao, J., Zhou, L., Qi, X. et al. (2020). Complex impact of DNA methylation on transcriptional dysregulation across 22 human cancer types. Nucleic Acids Research *48*, 2287–2302.

[Wen et al., 2017] Wen, Y., Wei, Y., Zhang, S., Li, S., Liu, H., Wang, F., Zhao, Y., Zhang, D. and Zhang, Y. (2017). Cell subpopulation deconvolution reveals breast cancer

heterogeneity based on DNA methylation signature. Briefings in Bioinformatics *18*, 426–440.

[Wu et al., 2017] Wu, S. H., Schwartz, R. S., Winter, D. J., Conrad, D. F. and Cartwright, R. A. (2017). Estimating error models for whole genome sequencing using mixtures of Dirichlet-multinomial distributions. Bioinformatics *33*, 2322–2329.

[Yan et al., 2021] Yan, Y.-y., Guo, Q.-r., Wang, F.-h., Adhikari, R., Zhu, Z.-y., Zhang, H.-y., Zhou, W.-m., Yu, H., Li, J.-q. and Zhang, J.-y. (2021). Cell-free DNA: hope and potential application in cancer. Frontiers in Cell and Developmental Biology *9*, 639233.

[Yang et al., 2020a] Yang, G. S., Mi, X., Jackson-Cook, C. K., Starkweather, A. R., Lynch Kelly, D., Archer, K. J., Zou, F. and Lyon, D. E. (2020a). Differential DNA methylation following chemotherapy for breast cancer is associated with lack of memory improvement at one year. Epigenetics *15*, 499–510.

[Yang et al., 2020b] Yang, J., Lang, K., Zhang, G., Fan, X., Chen, Y. and Pian, C. (2020b). SOMM4mC: a second-order Markov model for DNA N4-methylcytosine site prediction in six species. Bioinformatics *36*, 4103–4105.

[Yang et al., 2010] Yang, X., Lay, F., Han, H. and Jones, P. A. (2010). Targeting DNA methylation for epigenetic therapy. Trends in Pharmacological Sciences *31*, 536–546.

[Yang et al., 2022] Yang, Y., Zhang, T., Wang, J., Wang, J., Xu, Y., Zhao, X., Ou, Q., Shao, Y., Wang, X., Wu, Y. et al. (2022). The clinical utility of dynamic ctDNA monitoring in inoperable localized NSCLC patients. Molecular Cancer *21*, 1–6.

[Yin et al., 2019] Yin, L., Luo, Y., Xu, X., Wen, S., Wu, X., Lu, X. and Xie, H. (2019). Virtual methylome dissection facilitated by single-cell analyses. Epigenetics & Chromatin *12*, 1–13.

[Yin et al., 2017] Yin, W., Kann, K., Yu, M. and Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. arXiv preprint *arXiv:1702.01923*.

[Yoon, 2009] Yoon, B.-J. (2009). Hidden Markov models and their applications in biological sequence analysis. Current Genomics *10*, 402–415.

[Yoshihara et al., 2013] Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P. W., Levine, D. A. et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nature Communications *4*, 2612.

[Yu et al., 2021] Yu, Y., He, W., Jin, J., Xiao, G., Cui, L., Zeng, R. and Wei, L. (2021). iDNA-ABT: advanced deep learning model for detecting DNA methylation with adaptive features and transductive information maximization. Bioinformatics *37*, 4603–4610.

[Zaheer et al., 2020] Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L. et al. (2020). Big bird: Trans-

formers for longer sequences. Advances in Neural Information Processing Systems *33*, 17283–17297.

[Zellers et al., 2018] Zellers, R., Bisk, Y., Schwartz, R. and Choi, Y. (2018). SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing pp. 93–104, Association for Computational Linguistics.

[Zeng et al., 2023] Zeng, W., Gautam, A. and Huson, D. H. (2023). MuLan-Methyl—multiple transformer-based language models for accurate DNA methylation prediction. GigaScience *12*, giad054.

[Zhang et al., 2011] Zhang, F. F., Cardarelli, R., Carroll, J., Fulda, K. G., Kaur, M., Gonzalez, K., Vishwanatha, J. K., Santella, R. M. and Morabia, A. (2011). Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. Epigenetics *6*, 623–629.

[Zhang et al., 2017] Zhang, W., Feng, H., Wu, H. and Zheng, X. (2017). Accounting for tumor purity improves cancer subtype classification from DNA methylation data. Bioinformatics *33*, 2651–2657.

[Zhao et al., 2021] Zhao, L., Wu, X., Zheng, J. and Dong, D. (2021). DNA methylome profiling of circulating tumor cells in lung cancer at single base-pair resolution. Oncogene *40*, 1884–1895.

[Zheng et al., 2014] Zheng, X., Zhao, Q., Wu, H.-J., Li, W., Wang, H., Meyer, C. A., Qin, Q. A., Xu, H., Zang, C., Jiang, P. et al. (2014). MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. Genome Biology *15*, 1–13.

[Zhou et al., 2017] Zhou, J., Sears, R. L., Xing, X., Zhang, B., Li, D., Rockweiler, N. B., Jang, H. S., Choudhary, M. N., Lee, H. J., Lowdon, R. F. et al. (2017). Tissue-specific DNA methylation is conserved across human, mouse, and rat, and driven by primary sequence conservation. BMC Genomics *18*, 1–17.