Department of Computer Science

Probabilistic modeling of DNA methylation sequencing data

Viivi Halla-aho

Aalto University

DOCTORAL THESES

Aalto University publication series
DOCTORAL THESES 124/2022

Probabilistic modeling of DNA methylation sequencing data

Viivi Halla-aho

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall AS2 of the school on 8 November 2022 at 12.00.

Aalto University
School of Science
Department of Computer Science
Computational Systems Biology Group

Supervising professor
Professor Harri Lähdesmäki, Aalto University, Finland

Thesis advisor
Professor Harri Lähdesmäki, Aalto University, Finland

Preliminary examiners
Assistant Professor Jing Tang, University of Helsinki, Finland
Professor Matti Nykter, Tampere University, Finland

Opponent
Professor Jan Komorowski, Uppsala University, Sweden

Aalto University publication series
DOCTORAL THESES 124/2022

ISBN 978-952-64-0927-6 (printed)
ISBN 978-952-64-0928-3 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-0928-3

Unigrafia Oy
Helsinki 2022

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

Author
Viivi Halla-aho

Name of the doctoral thesis
Probabilistic modeling of DNA methylation sequencing data

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL THESES 124/2022

Field of research Information and Computer Science

Manuscript submitted 21 March 2022 **Date of the defence** 8 November 2022

Permission for public defence granted (date) 31 August 2022 **Language** English

☐ **Monograph** ☒ **Article thesis** ☐ **Essay thesis**

Abstract

DNA methylation is an epigenetic modification in which methyl groups bind to the DNA molecule. It regulates gene expression and enables the normal function of the cells. On the contrary, aberrant DNA methylation patterns have been associated with diseases such as cancer. Uncovering the mechanisms of gene regulation and utilizing DNA methylation biomarkers in e.g. cancer screening require advanced analysis methods for high-throughput sequencing data.

The aim of this thesis is to improve analysis of DNA methylation data with a probabilistic modeling approach. First, two methods for differential DNA methylation analysis designed for bisulfite sequencing data are proposed. In both methods, the spatial correlation of the methylation states is utilized in a binomial generalized linear mixed model to improve the accuracy of detecting differential methylation. The first method assumes that the DNA methylation across all cytosines in a genomic window have the same correlation characteristics and performs testing for differential methylation by computing one Bayes factor for each genomic window. In the other approach a sparsifying prior is used in the correlation structure to allow individual cytosines to deviate from the general correlation pattern. In the third publication, an analysis workflow for reduced representation bisulfite sequencing data is proposed. The workflow was applied to a cord blood data set, and differential DNA methylation analysis was performed to detect possible pregnancy or delivery-related changes in cord blood DNA methylation. In the fourth publication, methods for cell-free DNA-based cancer classification were developed and compared. To demonstrate the feasibility of liquid biopsies in clinical use, lower sequencing depth was simulated by subsampling the used cell-free methylated DNA immunoprecipitation sequencing data set. Then different generalized linear model classifiers and feature extraction and selection methods were applied and the resulting classification performance was evaluated.

The results presented in this thesis show that probabilistic modeling and Bayesian methods perform well and can improve the accuracy of analysis of DNA methylation sequencing data. Taking spatial correlation into account increased the accuracy of differential DNA methylation analysis. Allowing deviations from the correlation pattern made the analysis more flexible. Most of the differentially methylated cytosines and regions found from the cord-blood data set were sex-associated, and only a few were associated with the other clinical covariates. Additionally, the cord-blood data analysis revealed the problem of inflated p-values and a permutation-based method for solving the issue was proposed. Finally, methods that improved cell-free DNA methylation-based cancer classification included a logistic regression classifier and iterative supervised principal component analysis and Fisher's exact test for feature selection.

Keywords DNA methylation, probabilistic modeling, generalized linear models, bisulfite sequencing, cfMeDIP-seq

ISBN (printed) 978-952-64-0927-6 **ISBN (pdf)** 978-952-64-0928-3

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2022

Pages 146 **urn** http://urn.fi/URN:ISBN:978-952-64-0928-3

Tekijä
Viivi Halla-aho

Väitöskirjan nimi
DNA-metylaatiosekvensointidatan probabilistinen mallintaminen

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietotekniikan laitos

Sarja Aalto University publication series DOCTORAL THESES 124/2022

Tutkimusala Tietojenkäsittelytiede

Käsikirjoituksen pvm 21.03.2022 **Väitöspäivä** 08.11.2022

Väittelyluvan myöntämispäivä 31.08.2022 **Kieli** Englanti

☐ **Monografia** ☒ **Artikkeliväitöskirja** ☐ **Esseeväitöskirja**

Tiivistelmä

DNA-metylaatio on epigeneettinen muutos, jossa metyyliryhmiä kiinnittyy DNA-molekyyliin. Se säätelee geenien ilmentymistä ja mahdollistaa solujen normaalin toiminnan. Poikkeamat DNA-metylaatiotiloissa on toisaalta voitu yhdistää sairauksiin kuten syöpiin. Geenien säätelymekanismien ymmärtäminen ja DNA-metylaatiobiomarkkerien hyödyntäminen esimerkiksi syöpäseulonnoissa vaativat edistyneitä menetelmiä sekvensointidatan analysointiin.

Tämän väitöskirjan tavoite on parantaa DNA-metylaatiodatan analysointia probabilistisella mallinnustavalla. Väitöskirjan kaksi ensimmäistä julkaisua esittelevät kumpikin bisulfiittisekvensointidatalle tarkoitetun työkalun differentiaalista DNA-metylaatioanalyysia varten. Molemmissa hyödynnetään spatiaalista korrelaatiota yleistetyssä lineaarisessa sekamallissa differentiaalisen metylaation havaitsemistarkkuuden parantamiseksi. Ensimmäinen menetelmä olettaa kuhunkin genomi-ikkunaan kuuluvien sytosiinien metylaatiotilojen olevan keskenään korreloituneita. Differentiaalisen metylaation testaus tehdään laskemalla yksi Bayes-tekijä kutakin ikkunaa kohti. Toisessa työkalussa korrelaatiomatriisin määrittelyssä käytetään harvuutta tukevaa prioria, mikä sallii yksittäisten sytosiinien poikkeamisen yleisestä korrelaatiorakenteesta. Kolmannessa julkaisussa esitellään työnkulku RRBS-datan analysointia varten. Sitä käytettiin napaveriaineistoon, josta etsittiin raskauteen tai synnytykseen liittyviä DNA-metylaatiomuutoksia differentiaalisella metylaatioanalyysillä. Neljännessä julkaisussa vertailtiin menetelmiä soluvapaaseen DNA:han perustuvaan luokitteluun. Nestebiopsioiden soveltuvuutta kliiniseen käyttöön havainnollistettiin simuloimalla alempaa sekvensointisyvyyttä alinäytteistämällä käytetty soluvapaa MeDIP-seq-aineisto. Erilaisia yleistettyihin lineaarisiin malleihin perustuvia luokittelijoita ja menetelmiä piirteiden valintaan sovellettiin alinäytteistettyyn aineistoon ja menetelmien luokittelukyky mitattiin.

Väitöskirjassa esitetyt tulokset osoittavat, että probabilistiset mallinnusmenetelmät ja bayesiläiset menetelmät toimivat hyvin ja voivat parantaa DNA-metylaatiosekvensointidatan analyysien tarkkuutta. Spatiaalisen korrelaation ottaminen huomioon paransi differentiaalisen DNA-metylaatioanalyysin tarkkuutta. Sallimalla sytosiinien poikkeamisen korrelaatiorakenteesta analyysistä tuli joustavampi. Napaveriaineistosta löydetyistä differentiaalisesti metyloituneista sytosiineista suurin osa liittyi sukupuoleen ja vain muutama liittyi muihin kliinisiin muuttujiin. Lisäksi napaveriaineiston analyysi paljasti p-arvojen inflaatioon liittyvän ongelman, jonka ratkaisemiseksi esitettiin empiirinen menetelmä. Soluvapaaseen DNA-metylaatioon perustuvaa luokittelua paransivat yksinkertainen bayesiläinen logistinen regressioluokittelija sekä piirteiden valinnassa iteratiivinen ohjattu pääkomponenttianalyysi ja Fisherin tarkka testi.

Avainsanat DNA-metylaatio, probabilistinen mallintaminen, yleistetyt lineaariset mallit, bisulfiittisekvensointi, cfMeDIP-seq

ISBN (painettu) 978-952-64-0927-6 **ISBN (pdf)** 978-952-64-0928-3

ISSN (painettu) 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2022

Sivumäärä 146 **urn** http://urn.fi/URN:ISBN:978-952-64-0928-3

Preface

The research for this thesis was conducted at Computational Systems Biology research group, at the Department of Computer Science, Aalto University. This work has been financially supported by Academy of Finland and Emil Aaltonen Foundation, which I am very grateful for.

First and foremost I would like to express my gratitude to Prof. Harri Lähdesmäki for the possibility to work under his supervision. I am very thankful for your guidance and help during these years. I would also like to thank pre-examiners Prof. Jing Tang and Prof. Matti Nykter for their careful inspection of my thesis.

I would like to thank everyone involved in the DIPP cord blood study, especially Prof. Riitta Lahesmaa, Dr. Riikka Lund and Dr. Essi Laajala, for the fruitful collaboration. I learned a lot about DNA methylation, processing of sequencing data and DNA methylation analysis from Essi. Thank you for the pleasant and inspiring teamwork.

I have had the honor of having great colleagues in the Computational Systems Biology research group during these years. The peer support I have received from you has been very important. Also, the discussions over lunch and going to the parties organized by the department together made my workdays a lot more fun. I would especially like to thank Emmi Jokinen, Dr. Sini Rautio, Henrik Mannerström, Maria Osmala and Dr. Markus Heinonen. I would also like to thank Dr. Juhi Somani and Maria Osmala with whom I had the pleasure of working as a teaching assistant in the High-throughput bioinformatics course. In addition, I would like to thank Maia Malonzo for helping me getting started with processing of bisulfite sequencing data.

Last but not least, I would like to thank my friends and family for their support. Especially my partner, Antti, for always listening to my worries and helping with all the math questions I came up with.

Helsinki, September 3, 2022,

Viivi Halla-aho

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Viivi Halla-aho and Harri Lähdesmäki. LuxUS: DNA methylation analysis using generalized linear mixed model with spatial correlation. *Bioinformatics*, Vol. 36, No. 17, pp. 4535-4543, September 2020.

II Viivi Halla-aho and Harri Lähdesmäki. LuxHS: DNA methylation analysis with spatially varying correlation structure. *Lecture Notes in Bioinformatics (Proceedings of the 8th International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO 2020)*, pp. 505-516, May 2020.

III Essi Laajala, Viivi Halla-aho, Toni Grönroos, Ubaid Ullah Kalim, Mari Vähä-Mäkilä, Mirja Nurmio, Henna Kallionpää, Niina Lietzén, Juha Mykkänen, Omid Rasool, Jorma Toppari, Matej Orešič, Mikael Knip, Riikka Lund, Riitta Lahesmaa and Harri Lähdesmäki. Permutation-based significance analysis reduces the type 1 error rate in bisulfite sequencing data analysis of human umbilical cord blood samples. Accepted for publication in *Epigenetics*, 20 pages, March 2022.

IV Viivi Halla-aho and Harri Lähdesmäki. Probabilistic modeling methods for cell-free DNA methylation based cancer classification. Accepted for publication in *BMC Bioinformatics*, 24 pages, March 2022.

Author's Contribution

Publication I: "LuxUS: DNA methylation analysis using generalized linear mixed model with spatial correlation"

Viivi Halla-aho is the main author of the article. Viivi Halla-aho and Harri Lähdesmäki developed the method. Viivi Halla-aho implemented the method, ran the experiments and wrote the first version of the manuscript. Harri Lähdesmäki came up with the research question and contributed to the interpretation of the results and revision of the article.

Publication II: "LuxHS: DNA methylation analysis with spatially varying correlation structure"

Viivi Halla-aho is the main author of the article. Viivi Halla-aho and Harri Lähdesmäki developed the method. Viivi Halla-aho implemented the method, ran the experiments and wrote the first version of the manuscript. Harri Lähdesmäki came up with the research question and contributed to the interpretation of the results and revision of the article.

Publication III: "Permutation-based significance analysis reduces the type 1 error rate in bisulfite sequencing data analysis of human umbilical cord blood samples"

Essi Laajala is the main author of the article. Essi Laajala developed the analysis workflow, analyzed the data, interpreted the results, prepared the figures and tables, and wrote the manuscript. Viivi Halla-aho participated in the development of the analysis workflow and the interpretation of the results. Toni Grönroos was responsible for technical validation by targeted pyrosequencing. Ubaid Ullah Kalim participated in the interpretation of

the results. Mirja Nurmio, Mari Vähä-Mäkilä, and Juha Mykkänen provided the clinical information compiled by Essi Laajala, Henna Kallionpää, and Niina Lietzén. Omid Rasool supervised the laboratory experiments. Jorma Toppari provided the clinical information and supervised Mirja Nurmio and Mari Vähä-Mäkilä. Mikael Knip and Matej Orešič initiated the study together with Riitta Lahesmaa. Riikka Lund was responsible for the bisulfite sequencing and participated in the interpretation of the results. Riitta Lahesmaa supervised the study and participated in the interpretation of the results. Harri Lähdesmäki supervised Essi Laajala and Viivi Halla-aho, participated in the interpretation of the results, and revised the manuscript. All authors contributed to the final version of the manuscript.

Publication IV: "Probabilistic modeling methods for cell-free DNA methylation based cancer classification"

Viivi Halla-aho is the main author of the article. Viivi Halla-aho and Harri Lähdesmäki developed the methods. Viivi Halla-aho implemented and ran the experiments and wrote the first version of the manuscript. Harri Lähdesmäki came up with the research question, requested the data set for our use and contributed to the interpretation of the results and revision of the article.

Abbreviations

ADVI automatic differentation variational inference

AML acute myeloid leukemia

AUROC area under receiver operating characteristics

BF Bayes factor

bp base pair

BS-seq bisulfite sequencing

cfDNA cell-free DNA

cfMeDIP-seq cell-free methylation immunoprecipitation sequencing

CpG cytosine-phosphate-guanine, CG dinucleotide

ctDNA circulating tumor DNA

DMC differentially methylated cytosine

DMR differentially methylated region

DNA deoxyribonucleic acid

ELBO evidence lower bound

FDR false discovery rate

GLMM generalized linear mixed model

HMC Hamiltonian Monte Carlo

ISPCA iterative supervised principal component analysis

MCMC Markov chain Monte Carlo

NGS next-generation sequencing

PCA principal component analysis

PCR polymerase chain reaction

ROC receiver operating characteristics curve

RRBS-seq reduced representation bisulfite sequencing

SNP single nucleotide polymorphism

TET ten-eleven translocation

VI variational inference

5caC 5-carboxylcytosine

5fC 5-formylcytosine

5hmC 5-hydroxymethylcytosine

5mC 5-methylcytosine

1. Introduction

DNA sequencing technology has evolved greatly in the recent decades, and these more precise and less expensive methods have enabled generation of large amounts of sequencing data. Along with uncovering the genome, i.e. the complete set of genetic material of a cell [113], we can detect epigenetic modifications of the DNA. Epigenetic modifications are inheritable and reversible changes in the DNA, which do not modify the DNA sequences [113]. This includes phenomena such as histone modifications and DNA methylation, of which the latter one is the key topic of this thesis. DNA methylation is crucial for normal function of the cells as it is involved in many cellular processes, but aberrant changes can also be linked to different diseases [8].

In recent years, there has been growing interest in cell-free DNA based liquid biopsy methods for screening and diagnosis of cancer and other diseases [29]. Early detection can significantly decrease mortality rate of certain cancer types, such as colorectal cancer [43]. Cell-free DNA (cfDNA) is DNA that is not associated with any cells [42]. It can be measured from bodily fluids such as blood plasma [101], urine [87] and cerebrospinal fluid [20], from which the samples can be drawn in minimally non-invasive manner, in contrast to tissue sample collection which involves invasive operations. In case of presence of a tumor in the body, part of the cfDNA can be of tumor origin and this part is called circulating tumor DNA (ctDNA). Different types of genetic and epigenetic characteristics can be found from cfDNA: mutations [16, 95], nucleosome positions [105, 116], DNA fragmentation patterns [18, 120, 74], copy number changes [14, 18, 120] and DNA methylation [14, 58, 101]. These characteristics can potentially be used for screening, diagnosis and monitoring of cancer [99, 103] and other diseases [29]. More research is still needed to understand the biology behind cfDNA before liquid biopsies can be used clinically for these purposes [119]. However, cfDNA-based method for prenatal screening is already in use in Finland [50].

The evolution of sequencing technologies has inspired the development of tools for processing and analysis of sequencing data. The volume of

data coming from a high-throughput sequencing experiment is high and requires efficient computational methods for the data processing to be feasible. There are also different types of sequencing protocols, and each has specific features of its own which have to be taken into account. The preprocessing of raw sequencing reads for further analysis requires multiple steps, and for each step there is a variety of tools available [115]. For example, bisulfite sequencing protocol can be used to measure DNA methylation. Bisulfite sequencing data is often used for differential DNA methylation analysis, i.e. finding genomic locations which have statistically significantly different methylation states with respect to some covariate of interest. For this purpose, numerous analysis tools have been proposed [39, 21, 88, 2]. Also the rising interest in cfDNA-based cancer screening has led to development of cancer classification tools which utilize statistical methods and machine learning [14, 58, 101]. All in all, precise and efficient analysis methods are essential for gaining deeper understanding of the molecular biology of DNA methylation from sequencing data and for applying the knowledge e.g. in the cancer classification task. The aim of the doctoral research presented in this thesis is to propose methods for differential DNA methylation analysis and cfDNA-based cancer classification.

1.1 Research questions

The four publications included in this thesis aim to answer the research questions presented in Table 1.1. The publications Publication I and Publication II both propose differential DNA methylation analysis tools, which enable taking spatial correlation of the neighboring cytosines' methylation states into account and show that this increases the accuracy of finding differentially methylated regions. In particular, the aim was to expand previously published tool, LuxGLM [2], into this direction. LuxGLM has many useful features, such as allowing both binary and continuous covariates and taking experimental parameters into account, which makes it an appealing basis for model development. In Publication II, the spatial correlation structure also allows cytosines to not follow the general methylation pattern of the genomic window of interest. In Publication III, the aim is to perform differential DNA methylation analysis for finding DNA methylation changes in umbilical cord blood with respect to maternal and pregnancy-related covariates. This includes building a pipeline for preprocessing and analyzing bisulfite sequencing data. While Publication I, Publication II and Publication III handle bisulfite sequencing data, Publication IV is about analysis of cell-free methylation immunoprecipitation sequencing data. As stated in Table 1.1, the aim of this work is to improve the cfDNA-based cancer type classification using probabilistic modeling

and different feature selection and feature extraction methods. This work is motivated by promising results from previously published method [101], to which the new results are compared against.

Table 1.1. Research questions related to each publication in the thesis.

Publication	Research question
Publication I	How to include spatial correlation in probabilistic model for differential DNA methylation analysis and does it increase accuracy of the method?
Publication II	How to include spatial correlation into model for differential DNA methylation analysis, while allowing cytosines to not follow the general correlation pattern?
Publication III	Is there maternal or pregnancy-related differential methylation in umbilical cord blood and how to set up a pipeline for the analysis?
Publication IV	Can cfMeDIP-seq based cancer type classification be improved with probabilistic modeling and feature selection methods, especially when the sequencing depth is low?

1.2 Outline of the dissertation

This thesis is divided into two parts: an overview and an appendix consisting of a set of four articles. The overview part consists of five chapters. The second chapter gives a brief overview of the biological background of the research presented in this dissertation, namely DNA methylation, its role in diseases and sequencing methods that can be used to measure DNA methylation. The steps of sequencing data processing are also introduced. In the third chapter, the probabilistic modeling framework and the type of models applied in this thesis, generalized linear mixed models, are explained. The fourth chapter summarizes the aims, methods and results of each of the four articles. The dissertation concludes in the fifth chapter, where the article contributions are discussed and conclusions on the thesis are given.

2. DNA methylation

This chapter begins by describing an important epigenetic modification, DNA methylation, and its role in diseases, such as cancer. The focus in this thesis is on human biology. After that, the principles behind two next-generation sequencing protocols, bisulfite sequencing and cell-free methylated DNA immunoprecipitation sequencing, and what kind of data is retrieved from such experiments is explained. Finally, the essential steps of processing bisulfite sequencing data are described.

2.1 DNA methylation and its role in disease

In DNA methylation, a methyl group is attached to a cytosine in a covalent modification, forming a 5-methylcytosine (5mC) [45]. On the contrary, removal of a methyl group is called demethylation. There are two types of methylation: maintenance methylation, where methyl groups are added to newly made DNA to make sure the DNA pattern is inherited correctly, and *de novo* methylation, where the methylation pattern is changed by addition of a methyl group [15]. DNA methylation states of neighboring cytosines have been found out to be correlated, and while the correlation is strong between cytosines with distance under 1000 bp, the correlation diminishes when the distance grows larger than 2000 bp [24]. 5mC cytosines can be converted into 5-hydroxymethylcytosine (5hmC), and even further into 5-formylcytosine (5fC) and finally into 5-carboxylcytosine (5caC) by ten-eleven translocation (TET) proteins [98]. These modifications of 5mC are part of the demethylation process [51].

DNA methylation occurs generally in cytosines with CpG context, i.e. in cytosines which are part of a cytosine-phosphate-guanine dinucleotide [130]. On the other hand, there are some cell type exceptions such as neurons [73] and embryonic stem cells [69], in which non-CpG methylation is relatively common. DNA methylation has different functions in different genomic contexts [55]. In cytosines located in the transcription start site of a gene, a methyl group blocks most transcription factors from binding [15]. Tran-

scription factors are proteins, which bind to regulatory regions of the gene and regulate gene expression [15]. Blocking transcription factors from binding inhibits transcription. Over half of the genes contain a short CpG rich region, which are called CpG islands [55]. Whether a CpG island is methylated or not depends on the type of the gene: in housekeeping genes, which are expressed in all tissue types, the CpG islands are not methylated, while for the tissue specific genes, only the genes particular to the tissue type in question are left unmethylated [15]. This is one of the ways how epigenetic modifications enable cell differentiation to different cell types, even if all cells contain the same genetic material.

The mechanisms in which DNA methylation has an important role include genomic imprinting and silencing [15]. In genomic imprinting, a gene which is inactivated by methylation in one of the haploid gametes keeps its methylation status even after forming of a diploid zygote and after methylation pattern reprogramming, which happens in the early embryo [15]. In other words, the activity of the gene copy is determined by its parent of origin. Silencing means, that a genomic region consisting of one gene or a larger region is inactivated in a nonspecific manner by addition of DNA methylation and covalent changes to the histone proteins, which makes DNA condensate and become non-accessible [15]. X chromosome inactivation is also one example of silencing [15]. X chromosome inactivation happens in female cells, where there are two X chromosomes present. One of the X chromosomes is silenced by methylation, while the genes from the other X chromosome can be expressed.

Above some of the functions of DNA methylation in healthy cells were briefly described. Modifications to the normal DNA methylation patterns can lead to aberrant behavior of the affected cell. It is known that environment can cause changes in DNA methylation [27]. For example, maternal diet can affect the methylome of the offspring [57], and exposure to chemicals such as tobacco smoke [11] can alter the methylome and increase the risk of diseases [27]. The changes in DNA methylation can also be a result of a disease rather than the cause [47, 118]. For example, certain DNA methylation patterns have been associated to type 1 diabetes [30], cardiovascular disease [47] and rheumatoid arthritis [76]. Be it the cause or result of a disease, DNA methylation patterns can be used as biomarkers for different diseases [47].

Tumor cells usually differ from the normal cells by their lack of response to control mechanisms, leading to abnormal cell proliferation [60]. Malignant tumor cells can also invade surrounding tissues and spread to other locations of the body, i.e. metastasize [90]. The mechanisms of how DNA methylation is related to cancer development are not yet well understood, but evidence has been found of aberrant DNA methylation affecting the regulation of cancer-related genes, global hypomethylation promoting chromosome instability and 5mC cytosines being more prone

to mutagenesis [60]. Also, mutations in genes encoding TET proteins [49] and in genes regulating methylation catalyzing enzymes [102] can lead into aberrant methylation patterns and cancer as a consequence. As of 2018, there were over 14000 publications describing cancer-related DNA methylation biomarkers [61]. However, only few of the found biomarkers have yet been developed into clinical tests [61]. For example, DNA and cfDNA methylation signatures for colorectal cancer [54, 6], sarcomas [62] and intracranial tumors [84] have been discovered.

2.2 High-throughput sequencing DNA methylation data

Bisulfite sequencing (BS-seq) is a next-generation sequencing method which gives information of the methylation state of the cytosines in the sequenced DNA with base pair resolution. The sequencing assay can cover the whole genome (WGBS-seq) [32], or be targeted to CpG rich areas only [36]. The latter approach is called reduced representation bisulfite sequencing (RRBS-seq). In bisulfite sequencing, bisulfite sodium treatment is applied on the denatured DNA and it turns the unmethylated cytosines into uracils [66]. In the case of RRBS-seq, bisulfite conversion step is preceded by a step where DNA is fragmented with a restriction enzyme, which picks CpG-rich regions from the DNA [8]. The reduced material covers around 1% of the whole genome. Next, the bisulfite converted DNA is amplified with polymerase chain reaction (PCR), which produces DNA where the uracil-converted unmethylated cytosines appear as thymines [66]. This is followed by high-throughput sequencing [8]. The steps before sequencing are demonstrated in Fig. 2.1A.

From the resulting sequencing data it is possible to distinguish between methylated and unmethylated cytosines by comparing the reads to the original sequence [66]. Processing the data leads to total read count and methylated read count for each cytosine, which can be used for DNA methylation analysis. The proportion of methylated reads versus total number of reads for a cytosine is called methylation level [106]. The DNA methylation state of each cytosine is binary in a single cell, but as the input of BS-seq experiment is a population of cells [106], the DNA methylation state over cell population can be anything between 0 and 1. For example, in a case of whole blood sample, the cell population could consist of different blood cell types, which could each have their own distinctive DNA methylation patterns [96, 53].

Cell-free methylated DNA immunoprecipitation sequencing (cfMeDIP-seq) is a modification of methylated DNA immunoprecipitation sequencing (MeDIP-seq) method which is tailored to be used with cell-free DNA [101]. MeDIP-seq is a method where DNA is first fragmented and denatured, and the fragments with methylated cytosines are picked using methylation-

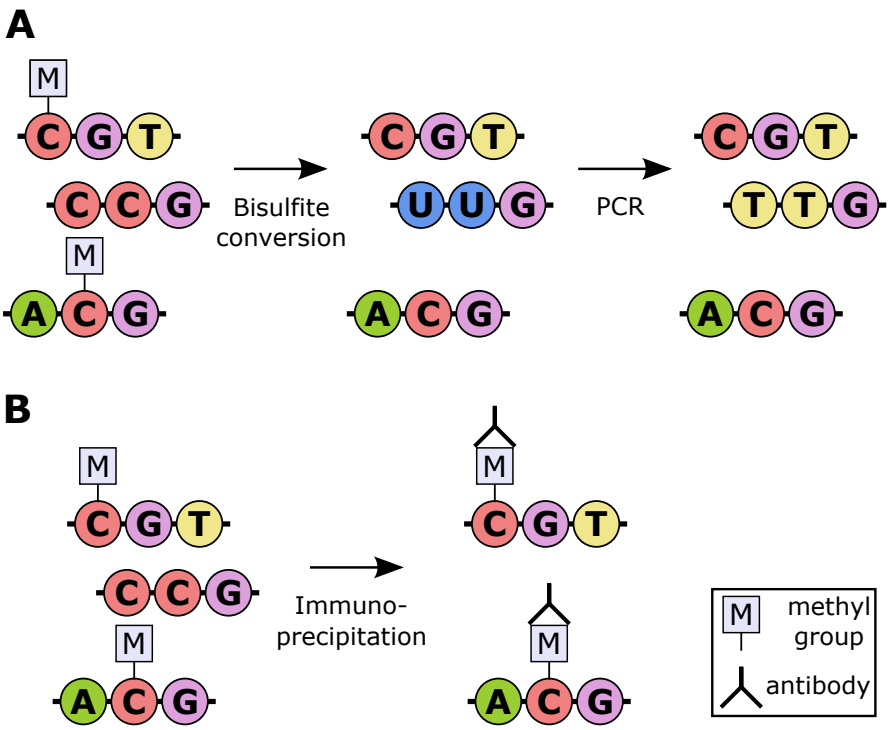

Figure 2.1. Key steps of (**A**) bisulfite sequencing and (**B**) MeDIP-seq protocols before sequencing the reads.

specific immunoprecipitation [121]. See Fig. 2.1B for illustration of the immunoprecipitation step. Then the DNA is sequenced with a high-throughput method [23]. As the selected fragments are sequenced and mapped to reference genome, we get a signal telling how many fragments with methylated cytosines cover each genomic region. The sequencing data can be turned into e.g. read counts per genomic windows of constant length [101]. The native average fragment length of cfDNA is 167 bp [105, 100]. This length originates from chromatosome structure, where 147 bp of DNA is wrapped around histone protein core, forming a nucleosome, while a linker histone protein is bound to the nucleosome. 20 bp of DNA is bound to the linker histone, thus the total length is 167 bp.

There are some limitations in both methods. In BS-seq, the bisulfite conversion step can be incomplete, which means that not all unmethylated cytosines are converted to uracils [66]. This affects the reliability of the resulting sequencing data. The bisulfite conversion step also degrades at least 84–96% of the input DNA [35]. This is a problem especially if the amount of input material is limited from the beginning, which is the case for example in cell-free DNA sequencing [100]. As 5hmC modification is resistant to bisulfite conversion, BS-seq cannot distinguish between 5mC and 5hmC [48]. On the contrary, MeDIP-seq only detects 5mC, as the antibody

used in the experiment is specific to the modification [112]. MeDIP-seq and cfMeDIP-seq methods do not require bisulfite conversion, and so extensive input DNA degradation is avoided. However, the resolution of methylation state information retrieved from MeDIP-seq and cfMeDIP-seq experiments is lower than in BS-seq, rather, the resolution is restricted to the DNA fragment length [112]. From BS-seq, the number of methylated and unmethylated reads is retrieved for each cytosine separately [112]. Other limitations of cfMeDIP-seq include that changes in signal caused by copy number variation cannot be distinguished from presence of DNA methylation and that signal is only retrieved from methylated reads [100]. The latter limitation means, that it is not possible to tell from cfMeDIP-seq data whether the absence of signal is caused by the genomic region in question not being covered for some reason from there simply being no DNA methylation in that region. To solve this problem in the case of MeDIP-seq, it has been proposed that it could be combined with methylation-sensitive restricting enzyme sequencing, which detects unmethylated CpGs [40]. In addition to experimental limitations, there lies uncertainty in the input DNA material too: as mentioned above, if a heterogenic sample is used as input to the BS-seq experiment, the cell type distribution can affect the methylation level estimates [106] and in cfMeDIP-seq the DNA fragments can originate from both normal and tumor tissues [100]. In summary, both approaches have their advances and limitations, which should be taken into account when choosing an appropriate method among the different DNA methylation measurement protocols, and also during downstream analysis of the retrieved sequencing data.

2.3 Preprocessing bisulfite sequencing data

Bisulfite data preprocessing starts with raw sequencing read files and results as count data matrix, which can be used for further analysis, such as differential DNA methylation analysis. Simplified preprocessing workflow is presented in Fig. 2.2. Below each of the steps is described and examples of possible tools made for each step are given. Pipeline tools which can perform all or part of the steps consecutively are also available. Examples of such tools are SAAP-RRBS [110], MethyQA [109] and gemBS [82], which use tools for separate steps as building blocks to form a pipeline.

First step in processing raw bisulfite sequencing data is quality control. For example, general tools for raw sequencing files, FastQC [5], HTQC [127] and Kraken [19], can be utilized while taking into consideration the special features of bisulfite sequencing data. FastQC takes raw sequencing data file as an input, and prints out a report with visualizations of sequence and per base quality scores, sequence lengths, duplicate

and overrepresented sequences and adapter content, along with other key statistics which inform about the quality of the data. If the sequencing reads are contaminated with adapter sequences, the reads must be trimmed or removed altogether [79]. Adapters are specific sequences which are attached to the ends of the DNA fragments in the library preparation step of sequencing [59]. Trimming can be done for example with Cutadapt tool [79]. HTQC [127] and Kraken [19] tool kits also include programs for read filtering. Quality control and read trimming steps can be repeated as needed to ensure appropriate data quality.

The next step is read mapping, where the sequencing reads are mapped into the reference genome. Due to the bisulfite conversion, which turns unmethylated 'C's into 'T's, there are multiple issues in BS-seq data that must be taken into consideration in read mapping: the reads are not complementary to the reference genome, lowered read complexity due to majority of non-CpG cytosines being transformed to 'T' leads more easily to misalignments and a 'T' in a read might align to a 'C' in the reference genome but not the other way around [124, 111]. The read mapping and the consequent methylation calling is even more complicated in the case of paired-end sequencing [111]. Bismark, a tool for read mapping and methylation calling, does mapping by converting both the reads and the reference genome into C-to-T and G-to-A versions and then finding the best matches [65]. Other mapping tools using the same approach include BS-Seeker2 [37] and GEM version 3 (GEM3) [78, 82]. Mapping BS-seq reads can also be done with hash table seeding algorithm, where all possible methylation status combinations are considered [39]. This is used for example in BSMAP tool [124]. BSMAP is used as a building block in SAAP-RRBS pipeline [110], while gemBS pipeline uses GEM3 [82].

From the aligned reads the Bismark tool can infer the total read count and methylated read count for each cytosine [65]. This can be called methylation calling or extraction. E.g. MethyQA [109] and gemBS [82] can also perform methylation extraction. Biases in the methylation level related to the position in the read, also known as M-bias, can happen in the both ends of the read [72]. This bias can be detected by comparing methylation levels in the middle of the read and in the read ends and then removed by trimming the read ends if needed [72]. The quality control tool BSeQC for mapped reads performs automatic trimming of nucleotides with potential technical biases which emerge as M-bias [72]. The BS-seq data pipeline and analysis tool BSmooth [39] and above-mentioned Bismark tool [65] provide information on the M-bias which can be used for making filtering decisions.

After read alignment, it is also important to investigate whether there are cytosine locations with single nucleotide polymorphisms (SNPs), meaning genetic variation occurring in a single nucleotide, which could bias the methylation calling [33]. Tools for BS-seq-specific SNP finding include

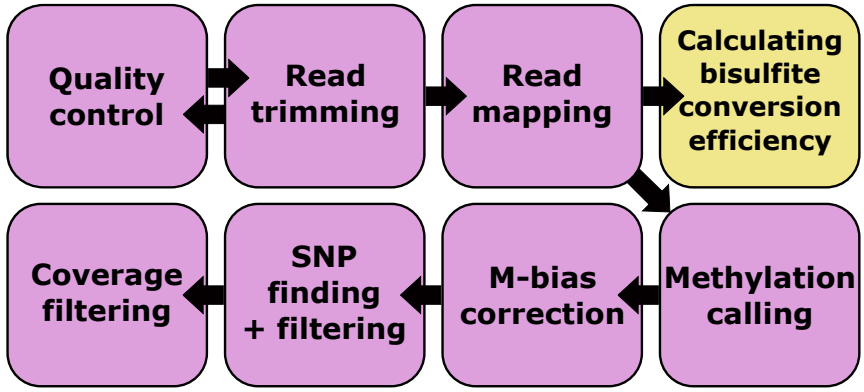

Figure 2.2. The preprocessing steps for BS-seq data. The quality control and read trimming steps can be repeated until desired data quality is achieved. The bisulfite conversion efficiency calculation step applies only to data from experiments with appropriate spike-in DNA.

BS-SNPer [33], Bis-SNP [75] and MethylExtract [7]. BS-SNPer does SNP calling in two steps [33]. First, it finds a list of putative mutations with high enough frequencies, and stores the position, reference base, number of supporting reads and their sequencing qualities. In the second step, this information is used for calculating the posterior probability for each genotype using Bayes' theorem. The genotype with the highest probability is then picked as the SNP. Bis-SNP is based on Bayesian inference model for SNP finding and genotyping, Genome Analysis Toolkit (GATK) [75], while MethylExtract is based on VarScan method [7]. SNP finding is performed for each sample, and if a SNP is found, the corresponding cytosine must be filtered out before sample data is used for further analysis.

Often it is of interest to evaluate how successful the bisulfite conversion step of the sample preparation has been, as it tells about the quality of the experiment [106]. Samples with relatively low bisulfite conversion rate might have to be removed from further analysis, as incomplete bisulfite conversion makes the data unreliable. However, some differential DNA methylation analysis tools, such as LuxGLM [2], take the bisulfite conversion rate into account and so it might not be necessary to remove samples with low conversion rates. The conversion rate can be estimated by spiking the sample DNA with unmethylated DNA, such as lambda virus DNA [106]. When doing the mapping step, the spike-in organism reference genome is used together with the actual reference genome. The read count of bisulfite converted cytosines divided by the total cytosine read count aligned to the spike-in reference genome then tells the conversion efficiency [106].

Before further analysis, cytosines with a very low read count are typically removed sample-wise, as the methylation levels of cytosines with low coverages cannot be estimated reliably [111]. However, if the data is

modeled with a distribution that takes the coverage information into account, such as binomial distribution, there is no need to filter such observations. This is discussed in Publication III. It is also recommended to filter out cytosines with very high read counts, as high coverage could be result of PCR duplication bias [3].

3. Probabilistic modeling

In this chapter, an overview of the used probabilistic modeling and statistical methods is given. First, the principles behind probabilistic modeling are described. Next, the generalized linear mixed models are introduced, focusing on models with binomial likelihood and logistic regression, which were utilized in differential DNA methylation analysis and cancer type classification tasks, respectively. In addition, a brief overview of the used feature selection, extraction and model regularization methods is given. The methods used for posterior inference, MCMC sampling and variational inference, are described in the third section. Finally, hypothesis testing and related methods are introduced in the fourth section.

3.1 Probabilistic modeling and Bayesian inference

In probabilistic modeling data and parameters are described using probability distributions, which quantify the uncertainty about the underlying process. Bayesian methods are a natural choice for probabilistic modeling. The steps of Bayesian data analysis include setting up the full probability model, calculating the posterior distribution and evaluating the fit of the resulting model [34].

In the first step of Bayesian data analysis, all observed and unobserved quantities are given a joint probability distribution [34]. This includes choosing prior probabilities for the parameters and possible hyperparameters. The joint probability for data y and parameter θ is

$$p(\theta, y) = p(\theta)p(y|\theta), \tag{3.1}$$

where $p(\theta)$ is the parameter prior and $p(y|\theta)$ is data distribution. The prior for parameter θ should give non-zero probability for all possible values for θ, and it can express the prior knowledge or uncertainty about the parameter, if such knowledge is available. When there is no prior information available, or if the effect of the prior is desired to be minimal, a flat, noninformative prior can be chosen. Weakly informative priors

contain slightly more information to keep the posterior within reasonable bounds.

In the second step, the Bayes' rule

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}, \tag{3.2}$$

is applied to calculate the posterior distribution [34]. Bayes' rule defines how the posterior distribution for parameter θ depends on the prior and observed data y. When calculating the posterior $p(\theta|y)$ the data distribution $p(y|\theta)$ can be regarded as a function of parameter θ, and so the term $p(y|\theta)$ can be called likelihood. In practice, the denominator term of Eq. 3.2, $p(y)$, is often omitted to retrieve unnormalized posterior density

$$p(\theta|y) \propto p(\theta)p(y|\theta), \tag{3.3}$$

as $p(y)$ does not depend on parameter θ. How the posterior can be evaluated in practice is discussed later in Section 3.3.

The third step of evaluating the fit of the model can include a variety of tasks, for example sensitivity analysis, evaluation of how well the model fits to the data and of whether the conclusions implied by the model are reasonable [34]. External data can be used for validating model predictions. Another technique for checking the model fit is to generate data under the learned model and compare it to the actual observed data. This is called posterior predictive checking. The observed and generated data can be compared e.g. with statistical tests. If needed, the three steps of the analysis can be repeated and modifications to the model can be done in the first step.

3.2 Generalized linear mixed models

In this section, the basic principles of generalized linear mixed modeling are described. The data retrieved from sequencing experiments can be of non-Gaussian type, such as binomial count data. Therefore, generalized linear models, which allow non-Gaussian likelihoods for the data, are often used in sequencing data analyses. In addition to the fixed linear effects the linear model part can be expanded by introducing random effects to the model, leading to mixed models. In some applications, the number of features in the model can be very high, even though not all features are expected to be important in predicting the outcome. This calls for methods for feature selection, feature extraction and regularization of the model.

3.2.1 Generalized linear models

Generalized linear models are an expansion of general linear models. The general linear model assumptions include normality, homoscedasticity and

linearity [77]. However, in the case of non-Gaussian response variable, these assumptions cannot be met. Examples of such response variable types include count, binary and proportion of counts data [77]. A popular approach for modeling non-Gaussian data is generalized linear model framework. Generalized linear models can be defined with three components: distribution of the response variable, linear predictor and link function [1]. Response variables are stored in vector $\mathbf{y}$ of length N, which is the number of observations [1]. Response variable is given a distribution from the exponential family, and the choice depends on the properties of the data type. For example, count data could be given Poisson distribution and proportion of counts data could be assumed to have binomial distribution [77]. The linear predictor can be expressed as

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \tag{3.4}$$

where $\mathbf{X}$ is a design matrix of size $N \times D$ and $\boldsymbol{\beta}$ is a vector of coefficients of length D [1]. D is the number of explanatory variables and in the presence of an intercept term in the model, D is increased by one. Explanatory variables can also be called covariates or features. The design matrix contains covariate information and often the first column is set to ones, corresponding to an intercept term. The coefficients in $\boldsymbol{\beta}$ describe the covariate effects on the response variable $\mathbf{y}$ and the vector can also contain the possible intercept term. The linear predictor is connected to the expectation of y_i through a bijective link function $g(\cdot)$

$$g(\mu_i) = \eta_i, \tag{3.5}$$

where $\mu_i = E(y_i)$ [1]. The choice of the link function depends on the distribution of the response variable. For example, logit link function is commonly used with binomial distribution while for Poisson distribution logarithm link function is a popular choice [77]. These link functions are also the canonical link functions of the corresponding distributions, meaning that the link functions transform the means to the distributions' canonical location parameters [77].

Next, two specific generalized models are described in more detailed manner. These are GLMs with binomial likelihood and logistic regression for binary response variable. The models with binomial likelihood were used in Publication I and Publication II to model the count data retrieved from BS-seq experiments and to perform differential DNA methylation analysis. In a binomial likelihood the data describes the number of successes y out of the total number of trials n with success probability p: $y \sim \mathrm{Bin}(n, p)$ [77]. The probability mass function of the binomial distribution with number of successes z, total number of trials n and success probability p is

$$f(z, n, p) = \binom{n}{z} p^z (1-p)^{n-z}. \tag{3.6}$$

The link function should be chosen so that the values of $\boldsymbol{\mu}$ are in the range $(0, 1)$, as the parameter is the success probability of the binomial distribution, notated as p in Eq. 3.6. As mentioned above, the canonical link function for binomial GLM is logit function [77].

In Publication I and Publication II, rather than using μ_i straightforwardly as p, a sigmoid function is used to transform the linear predictor into methylation proportion. The methylation proportion can then be used for calculating the success probability parameter of the binomial distribution.

Logistic regression can be used for modeling a binary response variable. The generation of each of the binary observations y_i can be considered as Bernoulli trial, with the probability of observation i being 1 is $P(y_i = 1) = \mu_i$ [1]. The probability mass function for Bernoulli distribution is $f(z, p) = p^z(1-p)^{1-z}$, where $z \in \{0, 1\}$ [77]. The link function connecting μ_i to the linear predictor in logistic regression is logit function

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta}, \tag{3.7}$$

where vector $\mathbf{x}_i$ contains the values from row of $\mathbf{X}$ corresponding to observation y_i [1].

Generalized linear models have been widely utilized for sequencing data analysis, such as differential DNA methylation analysis which uses bisulfite sequencing data. The generalized linear model approaches often enable including confounding covariates, such as age or sex, into the model through the design matrix. Some tools proposed for this purpose model the BS-seq counts as binomial [2, 108], while some use the methylation proportion instead of counts and apply logistic regression [3]. Another popular choice for modeling bisulfite sequencing data is beta-binomial regression [21, 88]. In beta-binomial regression the count data is binomially distributed $y_i \sim \mathrm{Bin}(n_i, p_i)$, and the success probability parameter p_i is assumed to have beta distribution $p_i \sim \mathrm{Beta}(\gamma\pi_i, \gamma(1-\pi_i))$ [17]. γ is a common dispersion parameter and π_i is calculated with logit link function. Beta-binomial model is often used for handling binomial data with overdispersion, i.e. the variation of the data is higher than expected for a binomial model [17].

GLMs have also been utilized in cfDNA-based cancer classification literature. Logistic regression has been used for diagnostic prediction of hepatocellular carcinoma patients from healthy controls [126]. The model used cfDNA methylation markers as features. Logistic regression was also used in Publication IV for building a binary classifier for distinguishing different cancer types.

3.2.2 Mixed models

The generalized linear model described in previous section has only fixed effects. Next the model is expanded further by addition of random effects. Models containing both fixed and random effects are called mixed models [77]. Mixed models can be used to describe dependence between and within underlying groups in the data [77]. It is assumed that there is one or more latent variables per each group, and they are assumed to be random. These are referred to as random effects, and they can be either intercept or slope type of terms [107]. The linear predictor in a generalized linear mixed model can be expressed as

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \tag{3.8}$$

where $\mathbf{Z}$ is design matrix for random effects and vector $\mathbf{u}$ contains the random effects with a desired distribution [107]. $\mathbf{u}$ is typically given a Gaussian distribution $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_u)$. Mixed models allow diverse correlation structures, and thereby can be used for different types of grouped data, such as repeated observations and longitudinal or spatial data [77]. The desired correlation structure of the random effect can be conveyed through $\boldsymbol{\Sigma}_u$.

Mixed effects are included in the differential DNA methylation analysis tool PQLseq [108], where heritability is inserted into the model as a mixed effect. In Publication I and Publication II, a random effect was added to the model to account for replicate effects. Also, in Publication I spatial correlation between neighboring cytosines' methylation states was included to the model through a random effect.

3.2.3 Feature selection, feature extraction and regularization

If the number of potential covariates explaining the variation of the response variable is high, it is often reasonable and even necessary to attempt to reduce the number of covariates. Simpler models with fewer covariates require less computational resources, are more robust especially when data set is small and are easier to interpret and visualize [4]. The methods for reducing the number of covariates before model fitting can be divided into feature selection and feature extraction methods [4]. Feature selection means that the most important features are picked and the rest of the covariates are discarded, while in feature extraction new features are computationally constructed from the original features. Another approach is to aim for model sparsity, i.e. some of the covariate coefficients being zero or close to zero [13]. This can be enabled by using sparsifying priors and regularizing penalty terms when learning the model parameters in Bayesian and frequentist approaches, respectively [92].

A simple feature selection approach is to perform statistical tests to see

whether the covariate is important for predicting the value of response variable. Suitable test for this purpose depends on the application, but could be for example Fisher's exact test or moderated t-test. The test is performed for each of the covariates and the ones with smallest p-values can then be picked to be used in the model. This technique was applied in Publication IV, following a previously published classifier pipeline [101]. Both Fisher's exact test and moderated t-test test were experimented with. The moderated t-test has also been applied in finding methylation markers for hepatocellular carcinoma to be used as features in a diagnostic model [126]. The statistical tests will be described in more detail in Section 3.4.

Principle component analysis (PCA) and its supervised variants, such as iterative supervised PCA (ISPCA) [94], can be used on a high-dimensional set of features to generate principal components which can be used as covariates instead of the original features, i.e. PCA can be used for feature extraction. Let us define the number of observations as N, the number of features as D and let $\mathbf{X}$ be a matrix of size $N \times D$ containing the feature values for each observation [86]. Data matrix $\mathbf{X}$ is centered by substracting the mean value for each feature. The sample covariance matrix of data $\mathbf{X}$ is $\mathbf{\Sigma} = \frac{1}{N-1}\mathbf{X}^T\mathbf{X}$. In principal component analysis the eigenvalues $\lambda_1, ..., \lambda_L$ of $\mathbf{\Sigma}$ and corresponding orthonormal eigenvectors $\mathbf{w}_i$, $i = 1, ..., L$ are calculated. L is the number of non-zero eigenvalues, for which applies $1 \leq L \leq \min(N-1, D)$ [52]. The eigenvectors, or their subset with the highest eigenvalues, can be collected into transformation matrix $\mathbf{W}$, where vectors $\mathbf{w}_i$ are the columns. The i'th principal axis is then defined as the direction of eigenvector corresponding to i'th largest eigenvalue, and the principal components $\mathbf{Z}$ are the projection of $\mathbf{X}$ on the principal axes, $\mathbf{Z} = \mathbf{W}^T\mathbf{X}$. The number of components needed can be chosen so that majority of the variance is explained by them.

The ISPCA method utilizes the knowledge of the values of the binary or multiclass response variable $\mathbf{y}$ and finds principal components which maximize the separation of the classes [94]. The algorithm consists of four steps, which are iterated K times, each iteration returning a supervised principal component. First, an univariate score $S(\mathbf{x}_j, \mathbf{y})$ telling how well a feature separates the binary classes from each other is calculated for each feature $\mathbf{x}_j$, $j = 1, ..., D$. Then the features with score higher than a threshold γ are collected into data matrix $\mathbf{X}_\gamma$ and first principal component is calculated using this matrix. $\mathbf{w}_\gamma$ is the eigenvector of the retrieved principal component. Threshold γ is chosen so that the score $S(\mathbf{z}_\gamma, \mathbf{y})$ for the projection $\mathbf{z}_\gamma = \mathbf{w}_\gamma^T X_\gamma$ is maximized. After finding the best $\mathbf{z}_\gamma$, the variation explained by it is substracted from $\mathbf{X}$ and a modified data matrix $\mathbf{X}'$ is retrieved. Finally, $\mathbf{X}'$ is set as $\mathbf{X}$ and another iteration can begin. The number of iterations K can be decided using a permutation test, which is performed after each iteration. The test gives low p-values if there is still relevant variation left in the current data matrix $\mathbf{X}$. After K iterations,

$\mathbf{X}$ can be used for producing unsupervised principal components. Above the algorithm was presented for setting with binary classes, but it can be extended to multiclass problems.

Both PCA and ISPCA approaches were experimented with in Publication IV for feature extraction, when the number of potential features for a cancer type classifier was very high. The principal components were used in logistic regression model for classifying different cancer types from each other. PCA has also been applied in similar manner for example in building a classifier for separating acute myeloid leukemia and acute lymphoblastic leukemia samples from each other based on their methylation patterns [83]. Instead of using all CpG sites as features in the classifier, principal components derived from the methylation pattern data were used. This approach was compared with feature selection methods, such as finding differentially methylated CpG sites using t-test.

When the potential features have been chosen and it is time to fit the model, model sparsity can be enabled with the choice of the prior for coefficients $\boldsymbol{\beta}$. There are many different sparsity enhancing priors, but they share the same key idea: the prior attempts to shrink the coefficients of the unimportant covariates to or near to zero, and give the important covariates a non-zero coefficient [93]. Also, it is assumed that the number of important covariates is considerably lower than the total number of covariates [93]. Two important categories of sparsifying priors are spike-and-slab priors and continuous shrinkage priors [93]. Next, horseshoe prior [13] and regularized horseshoe prior [93], which belong to the latter class, are described in detail. Following the notation in [13], the horseshoe prior for generalized linear model coefficients $\boldsymbol{\beta}$ from 3.4 can be defined as

$$\beta_j \sim \mathcal{N}(0, \lambda_j^2 \tau^2), \tag{3.9}$$

where $j = 1, ..., D$ and D is the number of covariates in the model. λ_j is the local shrinkage parameter for coefficient β_j and τ is the global shrinkage parameter, shared by all indices j. The global shrinkage parameter shrinks all parameters towards zero, but the local parameter lets some parameters to have higher values [92]. The local shrinkage parameters λ_j and the global shrinkage parameter τ are given half-Cauchy hyperpriors

$$\lambda_j \sim C^+(0, 1), \tag{3.10}$$

and

$$\tau \sim C^+(0, 1). \tag{3.11}$$

The hyperprior choice for the global shrinkage has been discussed in [92]. The regularized horseshoe prior was proposed to enable specification of level of sparsity and level of regularization for the larger coefficients [93]. The regularized horseshoe prior for coefficients $\boldsymbol{\beta}$ is defined as

$$\beta_j \sim \mathcal{N}(0, \tau^2 \tilde{\lambda}_j^2), \tag{3.12}$$

where the local shrinkage parameter $\tilde{\lambda}_j$ is defined as

$$\tilde{\lambda}_j = \frac{c^2 \lambda_j^2}{c^2 + \tau^2 \lambda_j^2}, \tag{3.13}$$

and $\lambda_j \sim C^+(0, 1)$. c^2 is the slab width and can be given a fixed value or for example an inverse-Gamma prior. Through c it is possible to regularize the coefficients in case they are weakly identified and to define the expected maximum effect size. The global shrinkage parameter is given a half-Cauchy prior

$$\tau \sim C^+(0, \tau_0^2), \tag{3.14}$$

where

$$\tau_0 = \frac{p_0}{D - p_0} \frac{\sigma}{\sqrt{N}}. \tag{3.15}$$

In this formula, p_0 is the prior guess of the number of non-zero coefficients, N is the number of observations, D is the number of covariates and σ is noise level, for which a pseudo-variance value can be used.

In frequentist approaches for general and generalized linear models, penalty parameters in the optimization target function can be used for regularization. Examples of this are ℓ_1 penalty lasso [114], ℓ_2 penalty ridge regression [4] and their combination, elastic-net penalty [131]. For example, the GLMnet tool [31] for fitting generalized linear models uses the elastic-net penalties. GLMnet has been applied in cfDNA 5mC-based [101] and cfDNA 5hmC-based cancer classification [71].

3.3 Posterior inference

This section covers the second step of Bayesian data analysis, calculating the posterior distribution. The difficulty in Bayesian analysis is that the posterior distribution, which is of high interest, is often of intractable nature. When the posterior is not analytically solvable easily or at all, Markov chain Monte Carlo (MCMC) methods and approximate methods such as variational inference can be used.

3.3.1 Markov chain Monte Carlo

In Markov chain Monte Carlo methods, the posterior distribution of interest is sampled from by forming a sequence of random numbers [34]. Each new value drawn depends only on the last value of the sequence, thus forming a Markov chain. The values are drawn using rules which ensure convergence to the true posterior distribution, if infinitely many samples are drawn. Gibbs sampler, Metropolis and Metropolis-Hastings are examples of simple MCMC algorithms. To sample parameter θ with Metropolis algorithm, a starting point θ_0 is first drawn from a starting distribution

$p_0(\theta)$ [34]. After this, for each time point $t = 1, 2, ...$ the following three steps are repeated. A proposal θ^* is sampled from a symmetric jumping distribution $J_t(\theta^*|\theta^{t-1})$. Then a ratio of densities

$$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}, \tag{3.16}$$

is calculated. Finally, θ^t is set to

$$\theta^t = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{t-1} & \text{otherwise.} \end{cases} \tag{3.17}$$

Hamiltonian Monte Carlo (HMC) is a Metropolis algorithm, which explores the posterior space efficiently [85]. In HMC, each parameter θ to be sampled is assigned a momentum variable ϕ [34]. A key part of an HMC iteration is updating the parameters and momentum variables simultaneously, taking L leapfrog steps. Each leapfrog step involves updating the parameters and momentum variables, both in relation to the other and imitating the physical dynamics. The updates are scaled with step size factor ϵ. The proposals θ^* and ϕ^* are the values of parameter and momentum vectors after the leapfrog steps. In the accept-reject step the momentum variable is taken into account in the density ratio:

$$r = \frac{p(\theta^*|y)p(\phi^*)}{p(\theta^{t-1}|y)p(\phi^{t-1})}, \tag{3.18}$$

where $p(\phi)$ is the momentum distribution with covariance M. In an extension of HMC algorithm called the no-U-turn sampler (NUTS) [46], the number of leapfrog steps L is adapted in each iteration [34]. This allows efficient exploration of the posterior distribution even in more difficult sampling problems. The step size ϵ and the momentum variables' covariance matrix M, also known as mass matrix, can also be tuned in NUTS [34].

After running an MCMC algorithm, it is essential to check the retrieved samples for convergence [34]. Typically, MCMC algorithms are run to produce a couple of independent chains of samples, from which the samples can be combined after removing the samples from the warm-up period of the sampling [34]. However, it must be checked that all of the separate chains are stationary and that the chains are mixing together well. Convergence checks are performed for ensuring that the sampling iterations have proceeded long enough and the samples represent the target distribution, not the starting approximations [34]. Two commonly used convergence statistics are potential scale reduction $\hat{R}$ and the effective sample size n_{eff} [34]. The convergence checks must be performed on each parameter separately.

Implementing and tuning an MCMC sampling algorithm can be a strenuous task and requires expert knowledge, especially in case of complex sampling algorithms and distributions which are hard to sample from.

Probabilistic programming language Stan [12] is one of the solutions developed for making MCMC sampling easier. Stan is supplied a Bayesian model, input data and sampling parameters. Using these inputs, it runs MCMC sampling and finally returns chains of posterior samples. Stan also returns summary statistics of the samples and convergence diagnostics. The functionalities of Stan include MCMC sampling with regular Hamiltonian Monte Carlo and no U-turn sampler (NUTS) algorithms, the latter of which is the default option. In Publication I, Publication II and Publication IV Stan was utilized for model fitting.

3.3.2 Variational inference

In variational inference (VI) the posterior distribution $p(\theta|y)$ is approximated with distribution $q(\theta)$, which is often chosen from a family of simple distributions that are easy to work with [34]. The approximation is found by minimizing the Kullback-Leibler (KL) divergence, a measure of difference between $q(\theta)$ and $p(\theta|y)$

$$KL(q||p) = -E_q\left(\log\left(\frac{p(\theta|y)}{q(\theta)}\right)\right) = -\int \log\left(\frac{p(\theta|y)}{q(\theta)}\right) q(\theta) d\theta. \tag{3.19}$$

Utilizing the rule for conditional probabilities $p(\theta|y) = \frac{p(\theta,y)}{p(y)}$ the divergence can be written as

$$KL(q||p) = -\int \log\left(\frac{p(\theta,y)}{q(\theta)}\right) q(\theta) d\theta + \int \log(p(y)) q(\theta) d\theta \tag{3.20}$$

$$= -E_q\left(\log\left(\frac{p(\theta,y)}{q(\theta)}\right)\right) + \log p(y). \tag{3.21}$$

The term $\log p(y)$ does not depend on $q(\theta)$, and thus minimizing the first term on the right hand side equals to minimizing the Kullback-Leibler divergence. This term is called evidence lower bound (ELBO). In VI applications the posterior $p(\theta|y)$ and consequently the KL divergence is often intractable, but evidence lower bound can be minimized instead.

Stan software provides a possibility to perform variational inference with automatic differentiation variational inference (ADVI) algorithm [67]. Using ADVI in Stan requires the user to only provide a model and a data set, the rest has been automated. The parameters are transformed into real-coordinate space and a fully-factorized Gaussian is chosen as the variational distribution family. The evidence lower bound is then optimized using stochastic gradient descent. Approximations of the ELBO and its gradients are retrieved using Monte Carlo integration.

Compared to MCMC methods, variational inference is often faster. However, due to its approximate nature, VI methods do not guarantee convergence to the true posterior like MCMC methods do [34]. In Publication I and Publication II HMC (NUTS) and ADVI were utilized for the same problem and the resulting accuracies were compared.

3.4 Hypothesis testing

The aim of building and fitting a generalized linear (mixed) model often is to find out whether the chosen covariates have a statistically significant effect on the value of the response variable [107]. This question can be formulated into a hypothesis test. The null hypothesis is that the coefficient of the covariate of interest is expected to be zero if there is no effect

$$H_0 : \beta_j = 0, \tag{3.22}$$

where j is the index of the covariate of interest [44]. The alternative hypothesis can then be set as

$$H_1 : \beta_j \neq 0. \tag{3.23}$$

The test defined by hypotheses in Eq. 3.22 and Eq. 3.23 can also be interpreted as a comparison of two nested models: full model with covariate j and a reduced model without covariate j.

After setting the null and alternative hypotheses, the corresponding test statistic can be calculated. In Bayesian inference, testing is often done using Bayes factors, defined as

$$BF_{01} = \frac{p(y|H_0)}{p(y|H_1)}, \tag{3.24}$$

for the hypotheses presented in Eq. 3.22-3.23. The Bayes factor can be estimated with Savage-Dickey density ratio. The Savage-Dickey estimate for BF_{01} presented above for parameters $\theta = (\beta_j, \psi)$, where ψ denotes the set of nuisance parameters is

$$BF_{01} = \frac{p(\beta_j = 0|y, H_1)}{p(\beta_j = 0|H_1)} E\left[\frac{p(\psi|H_0)}{p(\psi|\beta_j = 0, H_1)}\right]. \tag{3.25}$$

Eq. 3.25 simplifies to $BF_{01} = p(\beta_j = 0|x)/p(\beta_j = 0)$ if $p(\psi|H_0) = p(\psi|\beta_j = 0, H_1)$ [117]. The simplified estimator is easy to evaluate. MCMC samples can be used to make a kernel density estimate of the posterior and evaluate it at $\beta_j = 0$ to get $p(\beta_j = 0|x)$, while the denominator $p(\beta_j = 0)$ can be evaluated straightforwardly using the prior. This approach is used in Publication I and Publication II.

In frequentist analysis, test statistics such as log-likelihood ratio test, Wald test, F-test and moderated t-test can be used for testing linear model coefficients. In likelihood ratio test the likelihoods for the reduced L_0 and full L_1 models, corresponding to the likelihoods under null hypothesis and alternative hypothesis respectively, are compared [107]. Likelihood ratio is defined as $\Lambda = L_0/L_1$, but in general a distribution for Λ cannot be determined. However, the test statistic

$$-2 \log \Lambda = 2\ell_1 - 2\ell_0, \tag{3.26}$$

which can be shown to asymptotically approach χ^2_ν distribution [123], can be used instead. ℓ denotes log-likelihood and ν is the difference in number of parameters in the reduced and full model. Log-likelihood ratio test was used in RADMeth tool [21] for testing differential methylation for individual cytosines.

Wald test can be used for testing null hypothesis $H_0 : \beta_j = 0$ for individual coefficients [44]. The test statistic is calculated as

$$W = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)}, \tag{3.27}$$

where $\hat{\beta}_j$ is the coefficient estimate and $\mathrm{SE}(\hat{\beta}_j)$ is the standard error of the estimate. W follows Gaussian distribution $\mathcal{N}(0, 1)$. Wald test is used for testing individual cytosines for differential methylation in MACAU [70] and PQLseq tools [108].

F-test can be used for comparing nested models [44]. In the reduced model under null hypothesis one or more coefficients are set to zero $H_0 : \beta_{j_1} = ... = \beta_{j_k} = 0$. The alternative hypothesis corresponds to the full model. The test statistic is

$$F = \frac{\mathrm{RSS}_0 - \mathrm{RSS}/k}{\mathrm{RSS}/(n-q)}, \tag{3.28}$$

which follows F-distribution with degrees of freedom $(k, n-q)$. RSS_0 and RSS are the sums of squared residuals for the reduced and full models respectively. k is the difference in the number of parameters between the full and reduced models, while n is the number of data points and q is the number of parameters in the full model. F-test was applied in preanalysis step of LuxUS tool in Publication I for finding genomic windows for further analysis.

Moderated t-test was originally proposed for differential gene expression analysis [104, 91], but has been also applied for finding differentially methylated regions [101]. The log-expression level of G genes is modelled with a linear model. The statistical test for differential expression is defined with null hypothesis $\beta_{gj} = 0$, where β_{gj} is the coefficient of the covariate of interest for gene g. The moderated t-statistic for gene g, $g = 1, ..., G$

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}} \tag{3.29}$$

follows a t-distribution with degrees of freedom $d_g + d_0$ under null hypothesis. $\hat{\beta}_{gj}$ is the estimate for β_{gj} and v_{gj} is its unscaled variance. In the moderated t-test, the gene-wise residual variances are estimated using Empirical Bayes method, and the resulting variance estimate $\tilde{s}_g$ is a compromise of the individual gene-wise variance and global variance. d_g is the residual degrees of freedom for the ordinary t-statistic and d_0 corresponds

to the extra information brought in by pooling information from other genes.

The tests described above are for testing the significance of one or more covariates in a linear model. Fisher's exact test can be applied for testing the independence of two factors with two levels [10]. The null hypothesis is that the two factors are independent of each other. A contingency table containing the numbers of observations in each factor level is formed as presented in Table 3.1. The test statistic is $T = f$, which follows hypergeometric distribution under null hypothesis. The moderated t-test and Fisher's exact test have been applied in feature selection as described earlier in Section 3.2.3.

Table 3.1. Contingency table for testing the independence of two factors by Fisher's exact test [10]. Contingency table is a way to present frequencies for the different factor levels. n is the total number of observations and f is the frequency in the cell on the first row and column. $f_{1\bullet}$ and $f_{\bullet 1}$ are the marginal frequencies of the first row and first column of the table, respectively.

Factor 1	**Factor 2**		
	Level 1	Level 2	
Level 1	f	$f_{1\bullet} - f$	$f_{1\bullet}$
Level 2	$f_{\bullet 1} - f$	$n - f_{1\bullet} - f_{\bullet 1} + f$	$n - f_{1\bullet}$
	$f_{\bullet 1}$	$n - f_{\bullet 1}$	n

As mentioned in Chapter 2, DNA methylation states of neighboring cytosines are often correlated. However, differential methylation tests for bisulfite sequencing data are often performed cytosine-wise. To utilize the autocorrelation between cytosines, the cytosine-wise p-values can be combined. A method for this is weighted Z-test which is also known as Stouffer-Liptak test [21]. The weighed Z-test starts by transforming a sequence of p-values $p_1, ..., p_n$ into Z-scores with $z_i = \Phi^{-1}(1 - p_i)$ [129]. Then the scores are combined by using

$$p_z = 1 - \Phi\left(\frac{\sum_i z_i}{\sqrt{n + \sum_{i<j} \mathrm{cor}(z_i, z_j)}}\right), \tag{3.30}$$

where $\mathrm{cor}(z_i, z_j)$ is correlation coefficient. Φ and Φ^{-1} are the cumulative distribution function and inverse cumulative distribution function for the standard normal distribution. Weighted z-test is used in RADMeth tool [21] and Publication III. comb-p method implements the weighted Z-test which can be applied in any across-genome analysis with auto-correlated observations [89]. Other solutions for utilizing the spatial correlation of neighboring cytosines include data smoothing before statistical testing [39, 41] and performing tests for differentially methylated regions or cytosine clusters rather than single differentially methylated cytosines [81, 122, 63]. The latter approach was used in Publication I.

In bioinformatics applications, the number of tests performed is often very high. For example, when testing differential DNA methylation for the cytosines in the whole genome, the number of tests can be in the magnitude of millions. In such a multiple testing setting, false discovery rate (FDR) must be controlled. False discovery rate is the proportion of tests where the null hypothesis is falsely rejected out of all tests for which null hypothesis was rejected [9]. The act of falsely rejecting null hypothesis is also called type 1 error, while falsely accepting null hypothesis is called type 2 error [97]. One commonly used method for FDR control is Benjamini-Hochberg correction [9]. If m tests with hypotheses $H_1, ..., H_m$ were performed and corresponding p-values $p_1, ..., p_m$ were retrieved the Benjamini-Hochberg correction would be performed in following manner: First, the p-values are sorted $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$, while the hypothesis corresponding to $p_{(i)}$ is denoted as $H_{(i)}$. Then the largest index i for which applies

$$p_{(i)} \leq \frac{i}{m} q^* \tag{3.31}$$

is found. This index is called k. q^* is the desired false discovery rate. Finally, all $H_{(i)}$ where $i = 1, ..., k$ are rejected.

Another technique for controlling FDR is to estimate the test statistic null distribution by using permutation method [125]. In this approach, the null hypothesis testing is repeated using permuted data. Then the p-value threshold can be set using the estimated null distribution to reach the desired false discovery rate. This approach was used in Publication III.

4. Summary of the publications

In this chapter, methods and results of each of the four publications are summarized.

4.1 Publication I: LuxUS: DNA methylation analysis using generalized linear mixed model with spatial correlation

Publication I and Publication II propose new differential DNA methylation analysis tools, LuxUS and LuxHS, which utilize spatial correlation between the neighboring cytosines' methylation states to increase the accuracy of the analysis. The generalized linear model framework from LuxGLM tool [2] was used as a starting point for the new models. In Publication I, the LuxUS tool is proposed. The method is based on a GLMM with binomial likelihood. Let us consider an experiment with N_R samples, for which bisulfite sequencing has been performed. From sequencing and downstream processing the total read count N_{BS} and $N_{BS,C}$, count of reads where a 'C' was observed, are obtained for each cytosine. N_P is the number of fixed effect covariates. In LuxUS, the analysis is performed for a genomic window at a time. The window contains N_C cytosines and their genomic coordinates are stored in vector $\mathbf{c}$. The genomic windows can be either predefined, or the preanalysis method coming with LuxUS can be used. The preanalysis step allows filtering individual cytosines or genomic windows. For example, minimum total read count, maximum genomic length of the window, maximum number of cytosines in the window and a p-value threshold for a simple statistical F-test comparing the case and control sample methylation states can be set by the user.

The read counts $N_{\mathrm{BS,C},i}$ are considered binomially distributed with number of trials being $N_{\mathrm{BS},i}$ and the success probability, the probability of observing a 'C', being $p_{\mathrm{BS,C},i}$, $i = 1, ..., N_R \cdot N_C$. The binomial probability mass function was presented in Eq. 3.6. The success probability, i.e. the probability of observing a 'C', parameter is defined in the same way as in LuxGLM by taking into account the possibilities of there being an error in

the bisulfite conversion or sequencing steps of the experiment

$$\begin{aligned} p_{\mathrm{BS,C},i} = \theta_i((1 - \mathrm{BS}_{\mathrm{Eff},i})(1 - \mathrm{seq}_{\mathrm{Err},i}) \\ + \mathrm{BS}_{\mathrm{Eff},i} \cdot \mathrm{seq}_{\mathrm{Err},i}) \\ + (1 - \theta_i)((1 - \mathrm{BS}^*_{\mathrm{Eff},i})(1 - \mathrm{seq}_{\mathrm{Err},i}) \\ + \mathrm{BS}^*_{\mathrm{Eff},i} \cdot \mathrm{seq}_{\mathrm{Err},i}), \end{aligned}$$

where θ_i is methylation proportion, $\mathrm{BS}_{\mathrm{Eff},i}$ is bisulfite conversion efficiency, $\mathrm{seq}_{\mathrm{Err},i}$ is sequencing error probability and $\mathrm{BS}^*_{\mathrm{Eff},i}$ is incorrect bisulfite conversion rate. The latter three are experimental parameters, which may be evaluated using spike-in sequencing material. If it is not possible to evaluate the experimental parameters, they can be set to 1, 1 and 0 respectively. These values correspond to perfect bisulfite conversion and sequencing.

The methylation proportion vector $\boldsymbol{\theta}$ of length $N_R \cdot N_C$ is calculated with sigmoid link function

$$\boldsymbol{\theta} = \frac{1}{1 + \exp(-\boldsymbol{\eta})}, \tag{4.1}$$

where $\boldsymbol{\eta}$ is the linear predictor with mixed effects. Following the GLMM notation in Eq. 3.6, $\boldsymbol{\eta}$ is defined as

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_R\mathbf{u}_R + \mathbf{Z}_C\mathbf{u}_C + \mathbf{e}, \tag{4.2}$$

where $\mathbf{X}$ is a matrix of size $(N_C \cdot N_R) \times N_P$, $\boldsymbol{\beta}$ is a vector of length N_P, terms $\mathbf{Z}_R\mathbf{u}_R$ and $\mathbf{Z}_C\mathbf{u}_C$ correspond to sample and cytosine random effects respectively. Fixed effect coefficients are given a Gaussian prior $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2_\beta\mathbf{I})$, where σ^2_β should be set so that it allows high enough variation for $\boldsymbol{\beta}$. The last term is a noise term with distribution $\mathbf{e} \sim \mathcal{N}(0, \sigma^2_E\mathbf{I})$. The sample random effect $\mathbf{u}_R$ of length N_R is given a Gaussian prior $\mathbf{u}_R \sim \mathcal{N}(0, \sigma^2_R\mathbf{I})$. The spatial correlation is brought to the model through the cytosine random effect $\mathbf{u}_C$ of length N_C. The prior for the cytosine random effect is $\mathbf{u}_C \sim \mathcal{N}(0, \boldsymbol{\Sigma}_C)$, where the covariance terms depend on the distance between the cytosines

$$\mathrm{cov}(u_{c,i}, u_{c,j}) = \sigma^2_C \cdot \exp\big(\frac{-|c_i - c_j|}{\ell^2}\big). \tag{4.3}$$

ℓ is a lengthscale parameter with Gamma prior. The variance parameters σ^2_E, σ^2_R and σ^2_C are each given a Gamma or inverse-Gamma prior.

The model is fitted using probabilistic programming language Stan. For each genomic window, a Bayes factor is calculated to determine how much evidence there is that the binomial GLMM model coefficient for the covariate of interest is non-zero. In practice, the Savage-Dickey estimate of the Bayes factor is used as described in Section 3.4, using the samples retrieved from Stan.

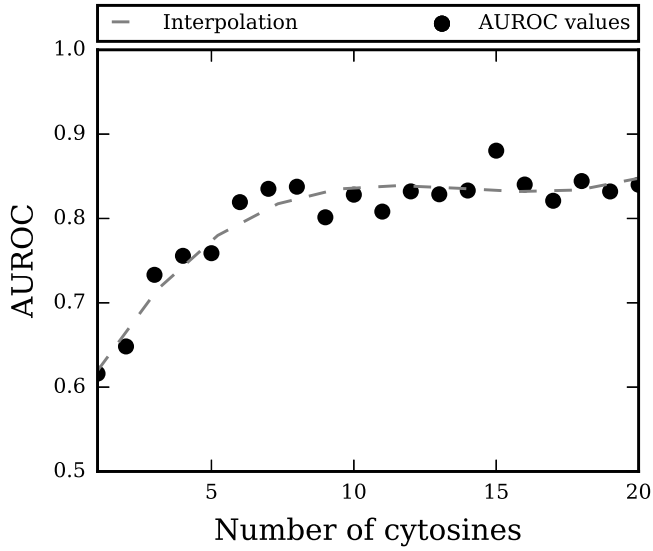

Figure 4.1. AUROC values as a function of number of cytosines in genomic window. Each value (black dots) has been evaluated over 400 simulated genomic windows, consisting of corresponding number of cytosines. The interpolation (dashed line) has been done with third degree polynomial. In this simulation setting, the AUROC values grow along with the number of cytosines per window until 5–10 cytosines per window, after which analyzing more cytosines together seems not to further increase the accuracy. Figure from Publication I. Reprinted with permission.

The LuxUS tool was tested using simulated and real BS-seq count data. For the simulation settings, a number of genomic windows with and without differential methylation were generated for model evaluation purposes. The performance of the tool was measured using receiver operating characteristics curve (ROC), which is a common method for classifier performance evaluation [26]. ROC graph shows true positive rate

$$\mathrm{TPR} = \frac{\text{correctly classified positives}}{\text{total positives}}, \tag{4.4}$$

against false positive rate

$$\mathrm{FPR} = \frac{\text{incorrectly classified negatives}}{\text{total negatives}}. \tag{4.5}$$

The area under ROC curve (AUROC) was used as a scalar measure of accuracy [26]. The AUROC values range from 0 to 1, with 1 meaning that the classifier has separated two classes perfectly. In practice, the AUROC values lie between 0.5 and 1, as a random guessing classifier should already result in AUROC value of 0.5. Also, the predictions from classifier with AUROC value below 0.5 can be flipped, resulting in classifier with AUROC value higher than 0.5. When evaluating differential DNA methylation analysis methods, the differentially methylated cytosines (DMCs) were

Table 4.1. AUROC values for LuxUS with HMC and ADVI, LuxUS for separate cytosines, RADMeth, M^{3}D, DSS, metilene and bsseq for simulated data set with two confounding covariates, highest value for each setting is bolded. N_{BS} and N_R denote the number of sequencing reads overlapping a cytosine and the number of samples, respectively. Table from Publication I. Reprinted with permission.

N_{BS}	N_R	LuxUS HMC	LuxUS ADVI	LuxUS sep.	RAD-Meth	M^{3}D	metilene	metilene mode 2	DSS	bsseq
6	12	**0.859**	0.851	0.605	0.728	0.674	0.625	0.717	0.845	0.614
6	24	**0.907**	0.883	0.682	0.800	0.618	0.735	0.840	0.870	0.688
12	12	**0.809**	0.802	0.634	0.702	0.644	0.676	0.712	0.757	0.628
12	24	**0.938**	0.899	0.748	0.821	0.722	0.772	0.861	0.915	0.737
24	12	**0.796**	0.738	0.641	0.717	0.658	0.592	0.684	0.750	0.626
24	24	**0.915**	0.880	0.731	0.827	0.690	0.709	0.836	0.874	0.714

given label 1 and correspondingly the non-DMC loci were set to have label 0.

One of the simulation experiments was aimed to demonstrate the benefit of utilizing the spatial correlation between cytosines. The simulated data consisted of sets of genomic windows, where the number of cytosines in the genomic window was different. The results showed, that analyzing 5–10 cytosines together increased the accuracy measured in AUROC when compared to cytosines being analyzed individually. Increasing the number of cytosines in a window from that seemed not to increase the accuracy further. The results from this experiment are shown in Fig. 4.1.

LuxUS was also compared to other published differential DNA methylation analysis tools which take spatial correlation into account. These tools were M^{3}D [81], DSS [28, 88], metilene [56], bsseq [39] and RADMeth [21]. The comparisons were done with simulated data. For simple simulation setting with no confounding covariates, the best performing tools by AUROC values were LuxUS, metilene and DSS. When two confounding effects, one binary covariate and one continuous covariate, were added to the simulation, LuxUS tool performed the best, followed by DSS. The AUROC values for the simulation setting with confounding covariates are presented in Table 4.1. This demonstrates the benefit of taking confounding effects into account, which is done by LuxUS and DSS. For metilene, M^{3}D and bsseq tools the results were weaker, which is as expected as they do not allow confounding covariates. The continuous covariate had to be binarized for RADMeth, as it only allows binary covariates. Perhaps due to this simplification RADMeth did not reach as good results as LuxUS and DSS. Full description of the results can be found from Publication I.

In Section 4.2 the methods of model fitting in LuxUS and LuxHS are described and the two tools are discussed.

4.2 Publication II: LuxHS: DNA methylation analysis with spatially varying correlation structure

In Publication II, LuxHS tool was proposed. The motivation for another formulation of a binomial model with included spatial correlation was to allow some of the cytosines in the genomic window of interest to not follow the general correlation pattern. Such cytosines will be called deviating cytosines in this section. Even if the DNA methylation is assumed to be a spatially correlated phenomenon, there might be cases in which one or more cytosines do not follow the pattern. This might be because of e.g. transcription factor binding [22]. The LuxUS model proposed in Publication I does not take the possibility of such deviations into account. For this purpose, a different model formulation is needed.

In LuxHS, a GLMM model similar to the one in LuxUS is fitted, but in contrary to LuxUS, Bayes factors are calculated separately for each of the cytosines in the genomic window. The main difference between the models is that in LuxHS there is no cytosine random effect, and that there are separate fixed effect coefficients for each cytosine. So the linear predictor can be expressed as

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_R\mathbf{u}_R + \mathbf{e}, \tag{4.6}$$

where $\mathbf{X}$ is a matrix of size $(N_C \cdot N_R) \times (N_C \cdot N_P)$ and $\boldsymbol{\beta}$ is a vector of length $N_C \cdot N_P$. $\boldsymbol{\beta}$ consists of cytosine-specific coefficient vectors . The replicate effect term and noise term are the same as in Eq. 4.2. The fixed effect coefficients have prior $\boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\beta)$. The spatial correlation is included in the model trough the covariance matrix $\boldsymbol{\Sigma}_\beta$, which is constructed in the following way

$$\boldsymbol{\Sigma}_\beta = \begin{pmatrix} \sigma_\beta^2 & \operatorname{cov}(\beta_{1,1}, \beta_{1,2}) & \cdots & \operatorname{cov}(\beta_{1,1}, \beta_{N_C,N_P}) \\ \operatorname{cov}(\beta_{1,2}, \beta_{1,1}) & \sigma_\beta^2 & \cdots & \operatorname{cov}(\beta_{1,2}, \beta_{N_C,N_P}) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(\beta_{N_C,N_P}, \beta_{1,1}) & \operatorname{cov}(\beta_{N_C,N_P}, \beta_{1,2}) & \cdots & \sigma_\beta^2 \end{pmatrix}. \tag{4.7}$$

The diagonal terms are the coefficient variances and the off-diagonal terms describe the covariance between the different cytosines' fixed effect coefficients. The covariance terms depend on the distance between the cytosines:

$$\operatorname{cov}(\beta_{j,k}, \beta_{j',k'}) = \begin{cases} \sigma_\beta^2 \cdot \exp \frac{-|c_j - c_{j'}|}{\ell^2} \cdot d_j \cdot d_{j'} & \text{if } k = k' \\ 0 & \text{if } k \neq k', \end{cases} \tag{4.8}$$

where j and j' are the cytosine indices while k and k' are the covariate indices. Covariance is set to zero between different covariates, as that would break the assumption of uncorrelated covariates. The distance-dependent terms are multiplied by corresponding cytosine-specific correlation weight

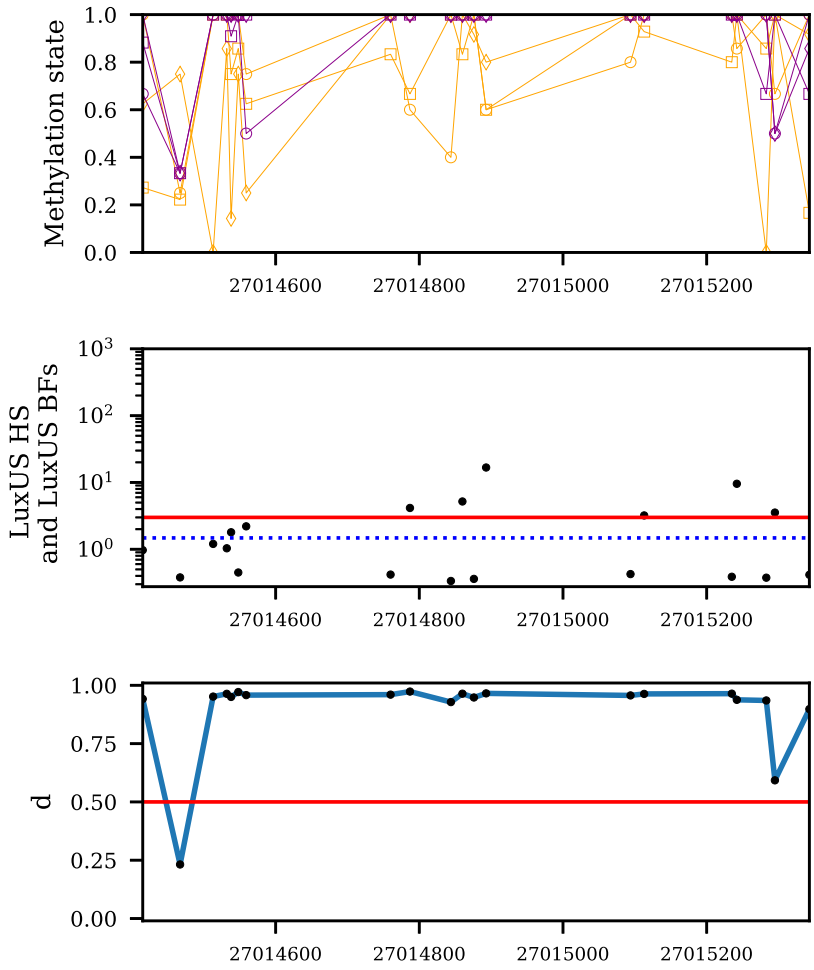

Figure 4.2. Comparison of LuxUS and LuxHS results. The figure shows the genome region chr22:27014415–27015343 from a real BS-seq data set [38]. The top panel shows the data, with cases plotted in purple and controls plotted in orange. The experimental design is matched, and the matching cases and controls have been plotted with the same markers. The middle panel shows the LuxHS (black dots) and LuxUS (dotted blue line) Bayes factor values. The red line shows BF value 3 for comparison. In the bottom panel the black dots show the cytosine weights d for the LuxHS model, with red line showing value 0.5 for comparison. For most of the cytosines d is close to 1, indicating that the cytosines follow the general correlation pattern. However, there are two cytosines with considerably lower d values, and from the data it can be seen that the methylation states for these cytosines seem to diverge from the rest of the cytosines in this window. The middle panel shows that in such a case, it might be favorable to calculate separate BFs for each cytosines in this kind of a case like LuxHS does, instead of assigning the same BF to the whole window like LuxUS. Figure from Publication II. Reprinted with permission.

variables d_j. The correlation weights d_j are retrieved through transformation

$$d_j = 1 - f(\tilde{d}_j), \tag{4.9}$$

Table 4.2. AUROC values for different simulated data settings. The best AUROC value for each setting is bolded. N_D is the number of deviating cytosines in a genomic window, N_R is the number of samples, N_{BS} is the total number of reads and μ_B contains the means of the coefficients β used in the simulations. $\mu_B = [-1.4, 1]$ and $\mu_B = [-1.4, 2.3]$ correspond to methylation state differences of 0.2 and 0.5 between cases and controls, respectively. Table adapted from Publication II. Reprinted with permission.

		$\mu_B = [-1.4, 1]$					$\mu_B = [-1.4, 2.3]$				
N_R	N_{BS}	LuxHS HMC	LuxHS ADVI	LuxUS sep	LuxUS	RADMeth (NaN values)	LuxHS HMC	LuxHS ADVI	LuxUS sep	LuxUS	RADMeth (NaN values)
		$N_D = 1$									
6	24	0.59	0.584	0.569	**0.618**	0.59 (10)	**0.828**	0.812	0.792	0.812	0.791 (10)
12	24	0.664	0.656	0.641	**0.688**	0.651 (30)	**0.906**	0.894	0.894	0.852	0.839 (10)
24	24	0.75	0.747	0.74	**0.767**	0.727 (10)	**0.974**	0.969	0.97	0.891	0.884 (30)
		$N_D = 2$									
6	24	0.557	0.555	0.553	**0.565**	0.538 (30)	**0.835**	0.808	0.825	0.738	0.729 (10)
12	24	0.648	0.641	**0.651**	0.622	0.588 (30)	0.905	0.889	**0.906**	0.75	0.774 (10)
24	24	0.696	0.695	**0.702**	0.658	0.639 (30)	**0.965**	0.957	**0.965**	0.785	0.787 (10)

where $f(\cdot)$ is a generalized logistic function which scales the $\tilde{d}_i$ value into range $[0, 1]$. To enable sparsity, the $\tilde{d}_i$ have a horseshoe prior with a slight modification of restricting $\tilde{d}_i$ to be a positive real number

$$\tilde{d}_j \sim \mathcal{N}^+(0, \tau^2 \cdot \lambda_j^2), \tag{4.10}$$

where λ_j and τ have the priors presented in Eq. 3.10 and Eq. 3.11, respectively. If the correlation weight d_i is near zero, the correlation terms between the corresponding cytosine and the other cytosines are diminished. This way not all of the cytosines have to follow the same methylation pattern as the rest. The horseshoe prior allows some d_j, $j = 1, ..., N_C$, to diminish to zero.

As LuxHS model includes cytosine-specific covariate coefficients $\boldsymbol{\beta}$, the hypothesis testing is also performed in cytosine-specific manner. Again, the Savage-Dickey density ratio estimate of the Bayes factor is used for testing. As result, LuxHS can detect differentially methylated cytosines, whereas LuxUS computes Bayes factors window-wise and can detect differentially methylated regions.

To demonstrate how LuxHS works, simulated and real data experiments were done. The results for a genomic window from real BS-seq data set are presented in Fig. 4.2. The figure demonstrates how the weight variable d behaves, when the data indicates that the cytosine does not follow the general correlation pattern. Also, in contrast to LuxUS the flexibility of LuxHS model is shown to be beneficial in the presented case.

The performance of LuxHS, LuxUS, LuxUS method applied for separate cytosines and RADMeth [21] were compared on simulated data. The simulated data was generated as genomic windows of length 1000 bp, each with 10 cytosines. The number of deviating cytosines in each window was varied from 0 to 2. The data was simulated from LuxHS model, and deviating cytosines were set to have correlation weights of $d = 0$.

Table 4.3. The AUROC and TPR values for detecting deviating cytosines for simulated data. TPR has been calculated with two thresholds (0.5 and 0.75) for the mean of posterior samples for d_j. N_D is the number of deviating cytosines in a genomic window, N_R is the number of samples, $N_{\mathrm{BS,Tot}}$ is the total number of reads and μ_B contains the means of the coefficients β used in the simulations. The $\mu_B = [-1.4, 2.3]$ value corresponds to methylation state difference of 0.5 between cases and controls. Table adapted from Publication II. Reprinted with permission.

	AUROC			TPR (0.5)			TPR (0.75)		
N_R	N_{BS}								
	6	12	24	6	12	24	6	12	24
	$N_D = 1$, $\mu_B = [-1.4, 2.3]$								
6	0.774	0.853	0.888	0.14	0.09	0.125	0.315	0.36	0.34
12	0.885	0.930	0.939	0.145	0.13	0.18	0.38	0.345	0.46
24	0.965	0.967	0.983	0.175	0.17	0.19	0.455	0.45	0.535
	$N_D = 2$, $\mu_B = [-1.4, 2.3]$								
6	0.741	0.786	0.804	0.103	0.11	0.118	0.36	0.318	0.358
12	0.833	0.863	0.869	0.153	0.145	0.11	0.423	0.408	0.313
24	0.893	0.899	0.937	0.15	0.16	0.183	0.458	0.458	0.5

In addition to the cytosines having no correlation with the rest of the cytosines, their differential methylation status was set to be the opposite from the other cytosines.

Results for selected simulation settings are shown in Table 4.2. The complete set of results is presented in Publication II. For the simulation settings with one deviating cytosine and with small methylation difference between cases and controls, making differential methylation analysis harder, LuxUS had the best overall performance measured in AUROC. However, when the methylation difference was increased, LuxHS had overall the best AUROC values.

With two deviating cytosines and small methylation differences between the cases and controls, LuxUS applied separately for each cytosine and LuxUS had the best overall performances. Again, when the methylation difference was set to be higher, LuxHS had the highest AUROC values. Occasionally, LuxUS method applied for separate cytosines had equal or slightly higher AUROC value. The success of the cytosine-wise LuxUS method suggests, that addition of two deviating cytosines per genomic window has muddled the correlation pattern severely and in such a case cytosine-wise analysis might be the best choice. Nevertheless, LuxHS method was able to outperform the LuxUS method applied for separate cytosines in most simulation settings where the methylation difference was high.

LuxHS method's ability to detect deviating cytosines was also measured using simulated data. For this purpose the estimated d_j values were examined. AUROC and true positive rate values were calculated for detecting the deviating cytosines. Results for selected simulated data sets

are presented in Table 4.3. Full results table can be found from Publication II. Even if the AUROC values were quite high, the TPR values remained rather low, suggesting that overall LuxHS estimated d_j appropriately, but it did not detect all deviating cytosines.

Both LuxUS and LuxHS models can be fitted using either MCMC sampling or variational inference approaches provided in Stan. The accuracies of the two approaches were compared. Based on these comparisons, both methods produce good results, but MCMC approach is slightly more accurate and more stable computationally. However, the computation times for the MCMC approach are considerably higher than for variational inference, so especially with high number of samples or covariates, the variational inference approach might be preferable. The analysis pipeline has been built in a way that promotes parallelization, which makes the tools faster in case there are multiple computing nodes available. Also, the preanalysis step can be used to filter the set of genomic windows for which the actual analysis is performed for. However, the preanalysis method is rather simplified, and the choice of its parameters can affect the results. Conservative parameter settings in the preanalysis phase could lead to accidentally filtering out differentially methylated cytosines. To allow as many differentially methylated cytosines as possible to proceed to the actual analysis, the parameters should be set in a rather liberal manner.

The LuxUS and LuxHS tools are interesting additions to the range of differential DNA methylation analysis tools with their two different ways of taking spatial correlation into account. One of the questions regarding the methods is whether the model assumptions match reality well enough. Although the good results from experiments with real BS-seq data suggest that the assumptions about the spatial correlation structure are reasonable. The LuxGLM model used as a basis can take into account technical parameters, such as bisulfite conversion efficiency, and allows usage of both continuous and discrete covariates. These features are available in LuxUS and LuxHS as well. Bayesian methods are utilized, so that the full posterior is retrieved as a result instead of just point estimates. Bayesian methods bring along another challenge, namely how the prior parameters should be chosen. In LuxUS and LuxHS the prior parameters can either be set to the proposed default values, or they can be defined by the user. The results of the analysis can be affected by the choice of priors. Nevertheless, the LuxUS model seems to not be very sensitive to e.g. σ_β^2 value (see Publication I and its supplementary materials for details). Finally, both LuxUS and LuxHS tools are freely available in GitHub[1].

[1]LuxUS: `https://github.com/hallav/LuxUS` and LuxHS: `https://github.com/hallav/LuxUS-HS`.

4.3 Publication III: Permutation-based significance analysis reduces the type 1 error rate in bisulfite sequencing data analysis of human umbilical cord blood samples

The aim of Publication III was to perform differential DNA methylation analysis on a relatively extensive RRBS-seq data set to find out whether there are pregnancy or delivery-related differences in cord blood methylation. The data set consisted of 200 cord blood samples, which had been RRBS-sequenced. The samples were collected from participants of the Finnish Diabetes Prediction and Prevention (DIPP) follow-up study between 1996 and 2006 in Turku University Hospital.

Publication III proposes an RRBS-seq analysis pipeline, which culminates in differential methylation analysis. The pipeline contains all relevant steps from raw sequencing data to FDR rate control. The preprocessing steps of BS-seq-type data have been described in Section 2.3. The pipeline utilizes open source tools and is freely available in GitHub[2]. The workflow begins with quality control and read trimming by Trim Galore tool [64], which combines FastQC [5] and Cutadapt [79] tools. Then alignment to reference genome is performed with Bismark [65], following with Bismark methylation calling and M-bias correction. The result of the methylation calling is a count matrix with total and methylated read counts for each cytosine in each sample. Single nucleotide polymorphisms are detected with bs-SNPer [33], and the counts for found SNPs were set to NA in the count matrix. The SNPs are also utilized for building a kinship matrix, which describes the genetic relatedness between the samples. Before differential methylation analysis, coverage filtering is performed on the count matrix.

The data set included a large number of covariates related to the participants of the study. Information about the mother, pregnancy and delivery were available. Using Pearson correlation coefficients and Fisher's exact test p-values for continuous and binary covariates, respectively, correlated covariates were first discovered. If there were two or more covariates with significant correlation, only one of them was picked to be used in the analysis. Missing covariate values were imputed with medians over the non-missing values. To account for technical variation, the sequencing batch information and first two principal components derived from the methylation proportion data were included in the analysis as covariates.

The differential DNA methylation analysis was performed with PQLseq tool, which uses a GLMM model that includes a genetic random effect for taking population structures into account [108]. Information about relatedness is passed to the model through kinship matrix, which was estimated using the detected SNPs. After filtering and covariate selection, the analysis included 173 samples and 24 covariates for 2752981 cytosines in

[2]RRBS pipeline: `https://github.com/EssiLaajala/RRBS_workflow`.

total. Alternative analysis was performed with RADMeth tool [21], which is based on beta-binomial model. RADMeth does not allow continuous covariates, so the continuous covariates had to be binarized. For both models, the count data was transformed to pseudo-counts by adding $+1$ count to the methylated count and adding $+2$ to the total read count for each non-missing cytosine count. Using this transformation helps in avoiding problems caused by exact 0 or 1 methylation proportions.

After fitting the models and retrieving cytosine-specific p-values, spatial adjustment and FDR control to the p-values can be done. First, the spatial adjustment and Benjamini-Hochberg correction steps from RADMeth were applied on both PQLseq and RADMeth cytosine specific p-values. However, the p-values seemed to be inflated after testing these steps on p-values for a permuted covariate of interest, for which no differential methylation should be found. Tests revealed, that the PQLseq modeling part was working as expected and the problem was caused by the spatial adjustment step. Additionally, raw RADMeth p-values were found to be inflated and the cause is that it cannot handle situations, where there are no observations for one of the two classes of a binary covariate.

As a solution to the p-value inflation caused by the spatial adjustment step, a permutation-based empirical FDR control method was used. In short, each of the covariates are permuted P times in a way that there is no correlation between the permuted covariate and the other covariates. The analysis is repeated for all cytosines one permuted covariate at a time to estimate the null distributions of the p-values. Then for each permutation and covariate, the p-value threshold is set so that the number of discoveries for the permuted covariate would be less than 5% of the number of discoveries for the original covariate. The final p-value threshold is set to median over the P permutation thresholds. The number of permutations P used to produce the results in Publication III was 3.

Benjamini-Hochberg-corrected PQLseq p-values and spatially adjusted, empirically FDR-corrected p-values were calculated for each cytosine. Differentially methylated cytosines were combined into differentially methylated regions, if two or more cytosines with p-value smaller than 0.05 were within 2000 bp of each other, and at least 90% of them are differentially methylated in the same direction. As a result of the analysis, differentially methylated cytosines and regions were found with respect to sex, epidural anesthetic, 1 minute Apgar points, maternal height, age of the mother, gestational weight gain, smoking during pregnancy and maternal insulin-treated diabetes. The numbers of differentially methylated regions and cytosines for each covariate are presented in Table 4.4.

One of the sex-related differentially methylated regions was technically validated with pyrosequencing, and the measurement from the different sequencing method supported the finding from the RRBS-seq-based analysis. Detailed description of the pyrosequencing results can be found from

Table 4.4. Results from the differential DNA methylation analysis of the cord blood data set. The first column shows the adjusted p-value threshold for each covariate, calculated with the permutation method with $P = 3$. The second column shows the number of differentially methylated regions (defined using the threshold from first column) and the third column shows the number of differentially methylated cytosines with Benjamini-Hochberg corrected PQLseq p-value smaller than 0.05. Table adapted from Publication III. Reprinted with permission.

Covariate	**Adjusted p-value threshold**	**Number of DMRs**	**Number of DMCs**
Year of birth (sample collection year)	0	0	6
Smoking during pregnancy, mother	0	0	1
Sex	$4.67 \cdot 10^{-7}$	297	1426
Month of birth (cosine transformed)	0	0	0
Insulin-treated diabetes, mother	$1.13 \cdot 10^{-10}$	2	10
Induced labor	0	0	0
Height, mother	$6.55 \cdot 10^{-11}$	2	3
Gestational weight gain	$9.53 \cdot 10^{-13}$	1	0
Epidural anesthetic, delivery phase 1	$1.22 \cdot 10^{-15}$	1	0
Earlier miscarriage(s)	0	0	0
Caesarean section	0	0	0
BMI, mother (pre-pregnancy)	0	0	0
Birth weight	0	0	0
Apgar points low, 1 minute	$3.99 \cdot 10^{-12}$	2	2
Age, mother	$9.59 \cdot 10^{-12}$	2	0

Publication III and from a related study, where the same data set was used for differential DNA methylation analysis between children who later develop type 1 diabetes and healthy controls [68]. Also, 221 of the differentially methylated CpGs with respect to sex covariate have also been detected in earlier microarray-based differential DNA methylation analyses of cord blood [128, 80]. Based on these comparisons, it can be said that the findings are reproducible between technologies and studies.

In addition to the DNA methylation analysis pipeline, Publication III includes some recommendations for RRBS-seq data analysis. The usage of pseudo-counts [108], was found out to significantly reduce convergence problems encountered with PQLseq and RADMeth. The convergence problems are caused by exact 0 or 1 methylation states, as the logit link function values go to infinite. Concerning the read count filtering, a total read count requirement for minimum number of samples can be used for filtering the cytosines. However, when using models which take into account the coverage through e.g. binomial or beta-binomial likelihoods, it is recommended not to filter out the observations with total read counts below the threshold.

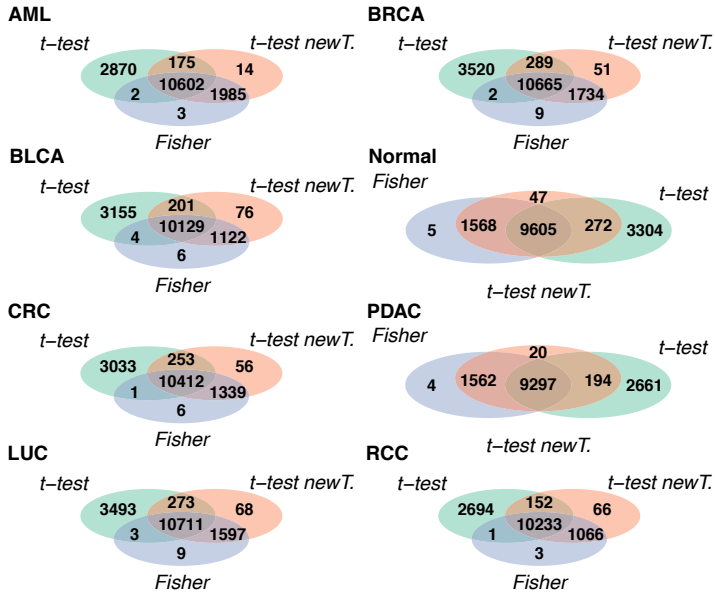

Figure 4.3. Venn diagrams showing the overlaps between DMRs sets found with the three DMR finding methods. The diagrams have been plotted for each of the eight classes separately. DMR sets from 100 data splits were pooled before plotting the diagrams. Results for thinned data with total read counts of 10^5. Figure adapted from Publication IV. Reprinted with permission.

4.4 Publication IV: Probabilistic modeling methods for cell-free DNA methylation based cancer classification

The aim of Publication IV was to experiment with different feature selection, feature extraction and probabilistic classification methods for improving cfMeDIP-seq based cancer classification. A blood plasma cfMeDIP-seq data set [101] worked as a basis for the experiments. The results were compared to the ones produced with a published method [101]. The data set consists of discovery and validation cohorts, with 189 and 199 samples respectively. In the discovery cohort, there are samples from seven cancer types and healthy controls, while the validation cohort consists of three cancer types and healthy controls.

In addition to testing various feature selection and classification methods, one of the aims was to study if the cancer type classification could be improved if the sequencing depth is low. For this purpose, the data was subsampled to simulate lower sequencing depth. This way, three different subsampled data sets were generated with total read counts, i.e. total number of cfMeDIP-seq reads per sample, of 10^6, 10^5 and 10^4 in each sample. The original total read counts per sample were of the magnitude of 10^7. The data was provided in preprocessed format, where for each sample the read counts for 505027 genomic windows, each of length 300 bp, had been computed.

For finding features for the cancer type classifiers, both feature selection and feature extraction methods were experimented with. Feature selection was done by statistical tests, which were used for finding differentially methylated regions. In the published method [101], the features to be used for classification were selected using moderated t-statistic, where log-transformed count data was used as input. The moderated t-tests were performed to the subsampled data with two log-transformations

$$\mathbf{X}_T = \log_2(c \cdot \mathbf{X} + s), \tag{4.11}$$

where $\mathbf{X}$ is the original count data, $c = 0.3$ and constant s is set to either $s = 10^{-6}$ or $s = 0.5$. $s = 10^{-6}$ is the same as in the published method and the latter is a new version of the transformation, which does not impose such a wide gap between zero and non-zero counts as $s = 10^{-6}$ does. Additionally, Fisher's exact tests were performed using the binarized version of the count data. Feature extraction was done with dimension reduction methods, specifically principal component analysis (PCA) and iterative supervised PCA (ISPCA) [94]. ISPCA utilizes the class labels of the training data set to find components that would best separate the classes from each other. Both binary and multiclass settings were tested with ISPCA.

In the previously published work [101], the cancer type classification was done with a binomial GLMnet model [31]. GLMnet utilizes elastic net regularization and fits the model with cyclical coordinate descent. For comparison, two types of logistic regression models were experimented with. The first one is similar to the GLMnet model: a set of DMRs were used as features and a sparsity inducing regularized horseshoe prior [93] was given to the coefficients. Principal components were also tested as features instead of DMR sites. The second model type is logistic regression with only two covariates: the numbers of hypermethylated and hypomethylated DMRs with non-zero counts. This type of a model had been found effective in T cell receptor analysis [25] and here it was tested whether such a simple, robust model would perform well, especially with the heavily subsampled data and very high-dimensional feature vectors. Probabilistic programming language Stan was used to fit the models. Feature selection, feature extraction and model fitting was done using training sets from 100 data splits, with 80% of the discovery cohort as training data and 20% of the discovery cohort as test data in each split. The classifiers were trained in one-vs-rest manner.

Next, the results for feature finding and classification are briefly described. The DMR sets found by each statistical testing method were compared by plotting Venn diagrams. These are presented in Fig. 4.3. Overall, there seemed to be high overlap between the three DMR sets produced by each method. However, the moderated t-test with the original transformation found a high number of DMRs unique to this method, while

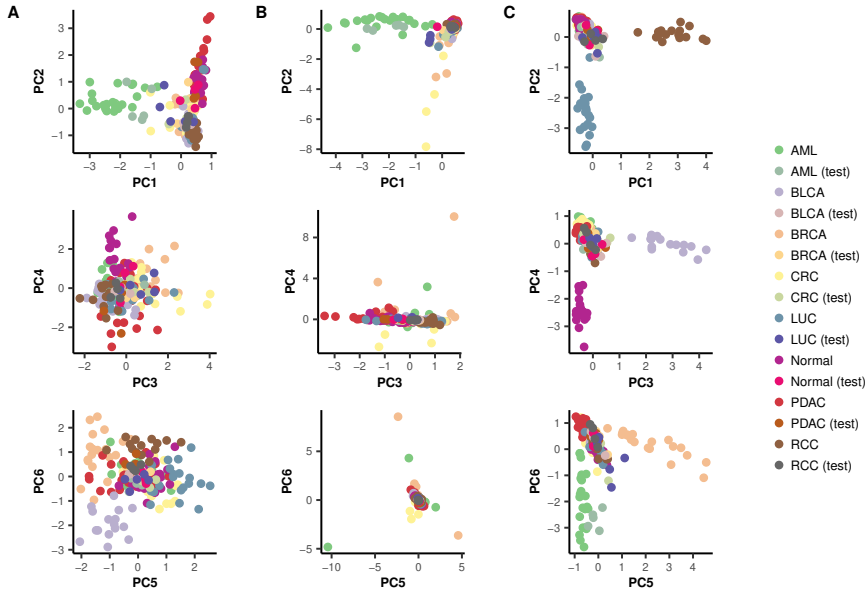

Figure 4.4. Visualization of results for (**A**) PCA using AML class DMRs as input, (**B**) binary ISPCA (AML class labels set to 1, other classes to 0) and (**C**) multiclass ISPCA. Results are shown for one data split. Each plot shows two principal components plotted against each other, and training and test samples for the data split have been plotted in different colors. Figure from Publication IV. Reprinted with permission.

the Fisher's exact test and moderated t-test with new data transformation shared a number of DMRs. The number of unique DMRs to these two methods only were low when compared to the moderated t-test with the original data transformation.

The results from PCA and ISPCA were also compared. PCA was run with a set of cancer class-specific DMRs as input, while for ISPCA all genomic windows were used. ISPCA was run with both binary and multiclass settings. An example of the results is shown in Fig. 4.4, where the results for subsampled data with total read count of 10^6 are presented for the acute myeloid leukemia (AML) class. The first principal component from the PCA and binary ISPCA seemed to clearly separate the AML samples from the other classes. The rest of the principal components seemed not to separate AML from other classes. Multiclass ISPCA showed expected behavior, and produced components which each separated one class from the others. The ISPCA approaches seemed to suffer from lack of information in the data when the subsampling of the data was more severe, as the method often produced no supervised components at all.

The classification performance was evaluated on the test data sets and on the separate validation cohort. Full description of the results can be found from Publication IV. Comparisons of the experimented methods against the method presented in earlier work for the discovery cohort is visualized

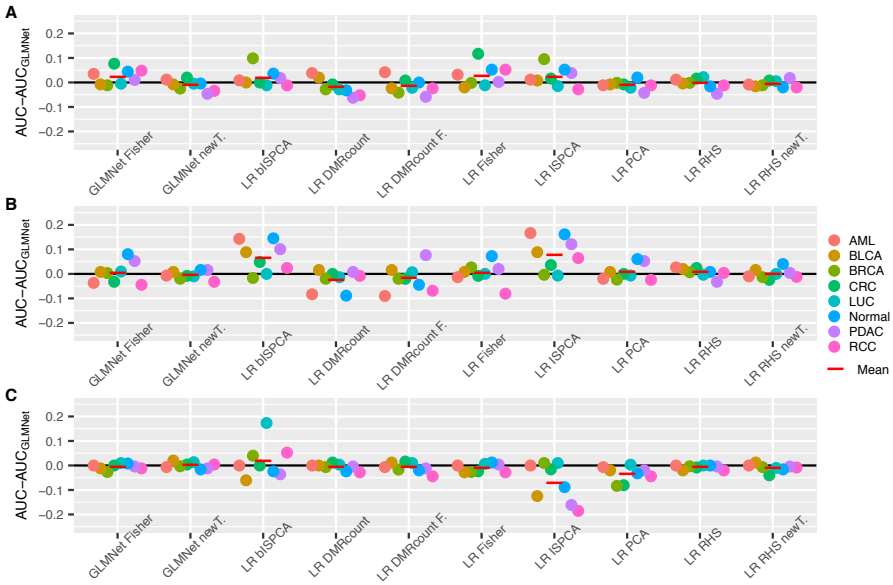

Figure 4.5. Differences between the class-specific AUROC values for the earlier published method based on GLMnet [101] and the new methods experimented in Publication IV. The red lines show the average performance across the classes. The **A**, **B** and **C** panels correspond to thinned data sets with 10^4, 10^5 and 10^6 total read counts, respectively. The results are presented for the discovery cohort set. Figure adapted from Publication IV. Reprinted with permission.

in Fig. 4.5. Our experiments showed that, as expected, the classification accuracy was lower when the sequencing depth was lower. The class-specific accuracies varied, implying that some classes are easier to classify than others. Overall, the original GLMnet method combined with moderated t-test DMRs performed well. But there were some cases where the new, experimented methods seemed to improve the classification results for the thinned data. These methods include the binary and multiclass ISPCA and Fisher's exact test for feature selection and logistic regression with DMR count variables for classification.

5. Discussion and conclusions

This thesis presents methods for two types of DNA methylation data analysis: differential DNA methylation analysis using BS-seq data and cell-free DNA methylation based cancer classification using cfMeDIP-seq data. For the former one, a data preprocessing pipeline and two analysis tools are proposed. For the latter one, different versions of logistic regression based classification methods and feature selection approaches were evaluated and compared on a data set with multiple cancer types. A feature common to all of the presented analysis methods is the usage of probabilistic modeling. Even if DNA methylation is one of the most studied epigenetic modifications of the DNA, research still remains to be done to understand its role in gene regulation and to utilize DNA methylation biomarkers e.g. in cancer diagnostics. For this purpose, advanced analysis methods are needed. This thesis aims to take a step, even if a small one, into that direction.

There are a couple of aspects in LuxUS and LuxHS tools proposed in Publication I and Publication II that could be improved or explored further. For example, different distance-dependent covariance functions could be experimented with. The models could be expanded to enable analysis of other methylation types, 5hmC and 5caC, in the similar manner as LuxGLM does [2]. This would make the tools more versatile. In Publication I and Publication II the focus was set on 5mC, as it is by far the most widely studied modification. The preanalysis step provided in LuxUS tool is quite simplified, and refining this step could enhance the actual GLMM analysis too. The two tools enable usage of both continuous and binary covariates and perform a Bayesian analysis using Stan, which provides access to summary tables and convergence diagnostics. The drawback of the methods is the long computation time, especially if the number of replicates and covariates is high. However, the computation can be parallelized genomic window-wise.

The contributions of Publication III include proposing a pipeline for BS-seq-based differential DNA methylation analysis, consisting of open source tools. The pipeline contains all relevant steps for the analysis, and

takes into account bias sources such as SNPs. Publication III also reports the inflation of spatially adjusted p-values, and suggests permutation-based empirical FDR control method. This method should be used with attention to sufficient number of permutations. Results from differential DNA methylation analysis of cord blood samples show, that there are sex-related methylation differences in the cord blood, part of which were novel findings and some were supported by earlier microarray data based studies.

In Publication IV, probabilistic modeling and different feature selection methods were tested for improving the classification accuracy of tumors based on data from cfMeDIP-seq experiments. Also, lower sequencing depths were simulated by subsampling the preprocessed count data. The results showed that the original method proposed for the task performed well for all of the classes and even if sequencing depth was lower. However, there were cases where the experimented methods could improve the classification. Making accurate predictions even with low sequencing depth could prove the utility of liquid tumor biopsies in clinical use.

In Publication IV, the focus was mainly in maximizing the classification accuracy. It remains as future work to investigate the biological meaning of the features chosen for the classifiers. For example, it would be interesting to find out if the PCA components could be given biological interpretations. The multiclass classification problem was simplified into multiple one-vs-rest classifiers, but building a multiclass classifier without transforming the problem into binary classification would be an interesting direction to investigate. The results in Publication IV do not include classification of early-stage and late-stage tumors into separate classes, which would have been possible for pancreatic ductal adenocarcinoma and lung cancer classes for which such data was available. It remains as future research to test the new classification methods for separating the early and late-stage tumors from each other. If the accuracy could be improved from the already impressive results presented in earlier work [101], this would further demonstrate the possibilities of liquid biopsies for tumor screening.

To conclude, this thesis began by introducing the topic of thesis and declaring the research questions related to the four publications. The second chapter described the epigenetic modification of interest in this thesis, DNA methylation, how it is measured and how the measurement data is processed. The third chapter explained the probabilistic and statistical methods utilized in the research. Finally, each of the publications was summarized and their contributions were discussed along with suggestions for future research.

References

[1] Alan Agresti. *Foundations of linear and generalized linear models*. Wiley series in probability and statistics. John Wiley & Sons Inc., Hoboken, New Jersey, 2015.

[2] Tarmo Äijö, Xiaojing Yue, Anjana Rao, and Harri Lähdesmäki. LuxGLM: a probabilistic covariate model for quantification of DNA methylation modifications with complex experimental designs. *Bioinformatics*, 32(17):i511–i519, 2016.

[3] Altuna Akalin, Matthias Kormaksson, Sheng Li, Francine E. Garrett-Bakelman, Maria E. Figueroa, Ari Melnick, and Christopher E. Mason. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome biology*, 13(10):1–9, 2012.

[4] Ethem Alpaydin. *Introduction to machine learning*. Adaptive computation and machine learning series. MIT Press, Cambridge, Massachusetts, third edition, 2014.

[5] Simon Andrews. Fastqc: A quality control tool for high throughput sequence data, 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 7 September 2021).

[6] S. Bach, I. Paulis, N. R. Sluiter, M. Tibbesma, I. Martin, M. A. van de Wiel, J. B. Tuynman, I. Bahce, G. Kazemier, and R. D. M. Steenbergen. Detection of colorectal cancer in urine using DNA methylation analysis. *Scientific reports*, 11(1):1–11, 2021.

[7] Guillermo Barturen, Antonio Rueda, José L. Oliver, and Michael Hackenberg. MethylExtract: high-quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000Research*, 2, 2013.

[8] Gustavo F. Bayón, Agustín F. Fernández, and Mario F. Fraga. Chapter 4 - bioinformatics tools in epigenomics studies. In Mario F. Fraga and Agustín F. Fernández, editors, *Epigenomics in Health and Disease*, pages 73–107. Academic Press, Boston, 2016.

[9] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[10] Stefano Bonnini, Livio Corain, Marco Marozzi, and Luigi Salmaso. *Nonparametric hypothesis testing : rank and permutation methods with applications in R*. Wiley Series in Probability and Statistics. Wiley, Chichester, England, 2014.

[11] Lutz P. Breitling, Rongxi Yang, Bernhard Korn, Barbara Burwinkel, and Hermann Brenner. Tobacco-smoking-related differential DNA methylation: 27k discovery and replication. *The American Journal of Human Genetics*, 88(4):450–457, 2011.

[12] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1):1–32, 2017.

[13] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. Handling sparsity via the horseshoe. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, pages 73–80, 2009.

[14] K. C. Allen Chan, Peiyong Jiang, Carol W. M. Chan, Kun Sun, John Wong, Edwin P. Hui, Stephen L. Chan, Wing Cheong Chan, David S. C. Hui, Simon S. M. Ng, Henry L. Y. Chan, Cesar S. C. Wong, Brigette B. Y. Ma, Anthony T. C. Chan, Paul B. S. Lai, Hao Sun, Rossa W. K. Chiu, and Y. M. Dennis Lo. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proceedings of the National Academy of Sciences*, 110(47):18761–18768, 2013.

[15] David P. Clark. *Molecular biology*. Elsevier Science & Technology, 2009.

[16] Joshua D. Cohen, Lu Li, Yuxuan Wang, Christopher Thoburn, Bahman Afsari, Ludmila Danilova, Christopher Douville, Ammar A. Javed, Fay Wong, Austin Mattox, Ralph H. Hruban, Christopher L. Wolfgang, Michael G. Goggins, Marco Dal Molin, Tian-Li Wang, Richard Roden, Alison P. Klein, Janine Ptak, Lisa Dobbyn, Joy Schaefer, Natalie Silliman, Maria Popoli, Joshua T. Vogelstein, James D. Browne, Robert E. Schoen, Randall E. Brand, Jeanne Tie, Peter Gibbs, Hui-Li Wong, Aaron S. Mansfield, Jin Jen, Samir M. Hanash, Massimo Falconi, Peter J. Allen, Shibin Zhou, Chetan Bettegowda, Luis A. Diaz, Cristian Tomasetti, Kenneth W. Kinzler, Bert Vogelstein, Anne Marie Lennon, and Nickolas Papadopoulos. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, 359(6378):926–930, 2018.

[17] Peter Congdon. *Applied Bayesian modelling*. Wiley Series in Probability and Statistics. Wiley, Chichester, England, second edition, 2014.

[18] Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel, Vilmos Adleff, Daniel C. Bruhm, Sarah Østrup Jensen, Jamie E. Medina, Carolyn Hruban, James R. White, Doreen N. Palsgrove, Noushin Niknafs, Valsamo Anagnostou, Patrick Forde, Jarushka Naidoo, Kristen Marrone, Julie Brahmer, Brian D. Woodward, Hatim Husain, Karlijn L. van Rooijen, Mai-Britt Worm Ørntoft, Anders Husted Madsen, Cornelis J. H. van de Velde, Marcel Verheij, Annemieke Cats, Cornelis J. A. Punt, Geraldine R. Vink, Nicole C. T. van Grieken, Miriam Koopman, Remond J. A. Fijneman, Julia S. Johansen, Hans Jørgen Nielsen, Gerrit A. Meijer, Claus Lindbjerg Andersen, Robert B. Scharpf, and Victor E. Velculescu. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, 570(7761):385–389, 2019.

[19] Matthew P. A. Davis, Stijn van Dongen, Cei Abreu-Goodger, Nenad Bartonicek, and Anton J. Enright. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods*, 63(1):41–49, 2013.

[20] Leticia De Mattos-Arruda, Regina Mayor, Charlotte K. Y. Ng, Britta Weigelt, Francisco Martínez-Ricarte, Davis Torrejon, Mafalda Oliveira, Alexandra Arias, Carolina Raventos, Jiabin Tang, Elena Guerini-Rocco, Elena

Martínez-Sáez, Sergio Lois, Oscar Marín, Xavier de la Cruz, Salvatore Piscuoglio, Russel Towers, Ana Vivancos, Vicente Peg, Santiago Ramon y Cajal, Joan Carles, Jordi Rodon, María González-Cao, Josep Tabernero, Enriqueta Felip, Joan Sahuquillo, Michael F. Berger, Javier Cortes, Jorge S. Reis-Filho, and Joan Seoane. Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma. *Nature communications*, 6(1):1–6, 2015.

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

[22] Silvia Domcke, Anais Flore Bardet, Paul Adrian Ginno, Dominik Hartl, Lukas Burger, and Dirk Schübeler. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature*, 528(7583):575–579, 2015.

[23] Thomas A. Down, Vardhman K. Rakyan, Daniel J. Turner, Paul Flicek, Heng Li, Eugene Kulesha, Stefan Gräf, Nathan Johnson, Javier Herrero, Eleni M. Tomazou, Natalie P. Thorne, Liselotte Bäckdahl, Marlis Herberth, Kevin L. Howe, David K. Jackson, Marcos M. Miretti, John C. Marioni, Ewan Birney, Tim J. P. Hubbard, Richard Durbin, Simon Tavaré, and Stephan Beck. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature biotechnology*, 26(7):779–785, 2008.

[24] Florian Eckhardt, Joern Lewin, Rene Cortese, Vardhman K. Rakyan, John Attwood, Matthias Burger, John Burton, Tony V. Cox, Rob Davies, Thomas A. Down, Carolina Haefliger, Roger Horton, Kevin Howe, David K. Jackson, Jan Kunde, Christoph Koenig, Jennifer Liddle, David Niblett, Thomas Otto, Roger Pettett, Stefanie Seemann, Christian Thompson, Tony West, Jane Rogers, Alex Olek, Kurt Berlin, and Stephan Beck. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature genetics*, 38(12):1378–1385, 2006.

[25] Ryan O. Emerson, William S. DeWitt, Marissa Vignali, Jenna Gravley, Joyce K. Hu, Edward J. Osborne, Cindy Desmarais, Mark Klinger, Christopher S. Carlson, John A. Hansen, Mark Rieder, and Harlan S. Robins. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature genetics*, 49(5):659–665, 2017.

[26] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[27] Robert Feil and Mario F. Fraga. Epigenetics and the environment: emerging patterns and implications. *Nature reviews genetics*, 13(2):97–109, 2012.

[28] Hao Feng, Karen N. Conneely, and Hao Wu. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, 42(8):e69–e69, 2014.

[29] Hao Feng, Peng Jin, and Hao Wu. Disease prediction by cell-free DNA methylation. *Briefings in Bioinformatics*, 20(2):585–597, 2019.

[30] Delphine Fradin, Sophie Le Fur, Clemence Mille, Nadia Naoui, Chris Groves, Diana Zelenika, Mark I. McCarthy, Mark Lathrop, and Pierre Bougneres. Association of the CpG methylation pattern of the proximal insulin gene promoter with type 1 diabetes. *PloS one*, 7(5):e36278, 2012.

[31] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[32] Marianne Frommer, Louise E. McDonald, Douglas S. Millar, Christina M. Collis, Fujiko Watt, Geoffrey W. Grigg, Peter L. Molloy, and Cheryl L. Paul. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences*, 89(5):1827–1831, 1992.

[33] Shengjie Gao, Dan Zou, Likai Mao, Huayu Liu, Pengfei Song, Youguo Chen, Shancen Zhao, Changduo Gao, Xiangchun Li, Zhibo Gao, Xiaodong Fang, Huanming Yang, Torben F. Ørntoft, Karina D. Sørensen, and Lars Bolund. BS-SNPer: SNP calling in bisulfite-seq data. *Bioinformatics*, 31(24):4006–4008, 2015.

[34] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*. CRC press, 2013.

[35] Christoph Grunau, S. J. Clark, and André Rosenthal. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic acids research*, 29(13):e65–e65, 2001.

[36] Hongcang Gu, Christoph Bock, Tarjei S. Mikkelsen, Natalie Jäger, Zachary D. Smith, Eleni Tomazou, Andreas Gnirke, Eric S. Lander, and Alexander Meissner. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nature methods*, 7(2):133–136, 2010.

[37] Weilong Guo, Petko Fiziev, Weihong Yan, Shawn Cokus, Xueguang Sun, Michael Q Zhang, Pao-Yang Chen, and Matteo Pellegrini. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC genomics*, 14(1):1–8, 2013.

[38] K. D. Hansen. bsseqdata: example whole genome bisulfite data for the bsseq package, 2016.

[39] Kasper D. Hansen, Benjamin Langmead, and Rafael A. Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome biology*, 13(10):1–10, 2012.

[40] R. Alan Harris, Ting Wang, Cristian Coarfa, Raman P. Nagarajan, Chibo Hong, Sara L. Downey, Brett E. Johnson, Shaun D. Fouse, Allen Delaney, Yongjun Zhao, Adam Olshen, Tracy Ballinger, Xin Zhou, Kevin J. Forsberg, Junchen Gu, Lorigail Echipare, Henriette O'Geen, Ryan Lister, Mattia Pelizzola, Yuanxin Xi, Charles B. Epstein, Bradley E. Bernstein, R. David Hawkins, Bing Ren, Wen-Yu Chung, Hongcang Gu, Christoph Bock, Andreas Gnirke, Michael Q. Zhang, David Haussler, Joseph R. Ecker, Wei Li, Peggy J. Farnham, Robert A. Waterland, Alexander Meissner, Marco A. Marra, Martin Hirst, Aleksandar Milosavljevic, and Joseph F. Costello. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature biotechnology*, 28(10):1097–1105, 2010.

[41] Katja Hebestreit, Martin Dugas, and Hans-Ulrich Klein. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13):1647–1653, 2013.

[42] Ellen Heitzer, Imran S. Haque, Charles E. S. Roberts, and Michael R. Speicher. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nature Reviews Genetics*, 20(2):71–88, 2019.

[43] Ellen Heitzer, Samantha Perakis, Jochen B. Geigl, and Michael R. Speicher. The potential of liquid biopsies for the early detection of cancer. *NPJ precision oncology*, 1(1):1–9, 2017.

[44] Stephane Heritier, Eva Cantoni, Samuel Copt, and Maria-Pia Victoria-Feser. *Robust methods in biostatistics*. Wiley Series in Probability and Statistics ; v.825. John Wiley & Sons, 2009.

[45] Holger Heyn. Chapter 1 - the role of the genetic code in the DNA methylation landscape formation. In Mario F. Fraga and Agustín F. Fernández, editors, *Epigenomics in Health and Disease*, pages 1–18. Academic Press, Boston, 2016.

[46] Matthew D. Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[47] Tianxiao Huan, Roby Joehanes, Ci Song, Fen Peng, Yichen Guo, Michael Mendelson, Chen Yao, Chunyu Liu, Jiantao Ma, Melissa Richard, Golareh Agha, Weihua Guan, Lynn M. Almli, Karen N. Conneely, Joshua Keefe, Shih-Jen Hwang, Andrew D. Johnson, Myriam Fornage, Liming Liang, and Daniel Levy. Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. *Nature communications*, 10(1):1–14, 2019.

[48] Yun Huang, William A. Pastor, Yinghua Shen, Mamta Tahiliani, David R. Liu, and Anjana Rao. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PloS one*, 5(1):e8888, 2010.

[49] Yun Huang and Anjana Rao. Connections between TET proteins and aberrant DNA modification in cancer. *Trends in Genetics*, 30(10):464–474, 2014.

[50] Aino-Kaisa Husu, Hannele Laivuori, Ritva Karhu, Tanja Saarela, and Kati Tihtonen. Äidin verestä otettavan sikiötestin (NIPT) vaikutus jatkotutkimuksiin osallistumiseen ja löydöksiin seulontapositiivisissa raskauksissa [The clinical impact of NIPT (non-invasive prenatal testing) on diagnostic testing after positive first trimester screening]. *Duodecim*, 136(3):315–22, 2020.

[51] Shinsuke Ito, Li Shen, Qing Dai, Susan C. Wu, Leonard B. Collins, James A. Swenberg, Chuan He, and Yi Zhang. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 333(6047):1300–1303, 2011.

[52] J. Edward Jackson. *A user's guide to principal components*. Wiley series in probability and mathematical statistics. Wiley, New York, 2004.

[53] Andrew E. Jaffe and Rafael A. Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome biology*, 15(2):1–9, 2014.

[54] Shengnan Jin, Dewen Zhu, Fanggui Shao, Shiliang Chen, Ying Guo, Kuan Li, Yourong Wang, Rongxiu Ding, Lingjia Gao, Wen Ma, Tong Lu, Dandan Li, Zhengzheng Zhang, Suili Cai, Xue Liang, Huayu Song, Ling Ji, Jinlei Li, Zhihai Zheng, Feizhao Jiang, Xiaoli Wu, Ju Luan, Huxiang Zhang, Zhengquan Yang, Charles R. Cantor, Chang Xu, and Chunming Ding. Efficient detection and post-surgical monitoring of colon cancer with a multi-marker DNA methylation liquid biopsy. *Proceedings of the National Academy of Sciences*, 118(5), 2021.

[55] Peter A. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492, 2012.

[56] Frank Jühling, Helene Kretzmer, Stephan H. Bernhart, Christian Otto, Peter F. Stadler, and Steve Hoffmann. Metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome research*, 26(2):256–262, 2016.

[57] Nina Kaminen-Ahola, Arttu Ahola, Murat Maga, Kylie-Ann Mallitt, Paul Fahey, Timothy C. Cox, Emma Whitelaw, and Suyinn Chong. Maternal ethanol consumption alters the epigenotype and the phenotype of offspring in a mouse model. *PLoS genetics*, 6(1):e1000811, 2010.

[58] Shuli Kang, Qingjiao Li, Quan Chen, Yonggang Zhou, Stacy Park, Gina Lee, Brandon Grimes, Kostyantyn Krysan, Min Yu, Wei Wang, Frank Alber, Fengzhu Sun, Steven M. Dubinett, Wenyuan Li, and Xianghong Jasmine Zhou. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome biology*, 18(1):1–12, 2017.

[59] Martin Kircher, Patricia Heyn, and Janet Kelso. Addressing challenges in the production and analysis of illumina sequencing data. *BMC genomics*, 12(1):1–14, 2011.

[60] Margaret Knowles and Peter Selby. *Introduction to the Cellular and Molecular Biology of Cancer*. Oxford University Press, 2005.

[61] Alexander Koch, Sophie C. Joosten, Zheng Feng, Tim C. de Ruijter, Muriel X. Draht, Veerle Melotte, Kim M. Smits, Jurgen Veeck, James G. Herman, Leander Van Neste, Wim Van Criekinge, Tim De Meyer, and Manon van Engeland. Analysis of DNA methylation in cancer: location revisited. *Nature reviews Clinical oncology*, 15(7):459–466, 2018.

[62] Christian Koelsche, Daniel Schrimpf, Damian Stichel, Martin Sill, Felix Sahm, David E. Reuss, Mirjam Blattner, Barbara Worst, Christoph E. Heilig, Katja Beck, Peter Horak, Simon Kreutzfeldt, Elke Paff, Sebastian Stark, Pascal Johann, Florian Selt, Jonas Ecker, Dominik Sturm, Kristian W. Pajtler, Annekathrin Reinhardt, Annika K. Wefers, Philipp Sievers, Azadeh Ebrahimi, Abigail Suwala, Francisco Fernández-Klett, Belén Casalini, Andrey Korshunov, Volker Hovestadt, Felix K. F. Kommoss, Mark Kriegsmann, Matthias Schick, Melanie Bewerunge-Hudler, Till Milde, Olaf Witt, Andreas E. Kulozik, Marcel Kool, Laura Romero-Pérez, Thomas G. P. Grünewald, Thomas Kirchner, Wolfgang Wick, Michael Platten, Andreas Unterberg, Matthias Uhl, Amir Abdollahi, Jürgen Debus, Burkhard Lehner, Christian Thomas, Martin Hasselblatt, Werner Paulus, Christian Hartmann, Ori Staszewski, Marco Prinz, Jürgen Hench, Stephan Frank, Yvonne M. H. Versleijen-Jonkers, Marije E. Weidema, Thomas Mentzel, Klaus Griewank, Enrique de Álava, Juan Díaz Martín, Miguel A. Idoate Gastearena, Kenneth Tou-En Chang, Sharon Yin Yee Low, Adrian Cuevas-Bourdier, Michel Mittelbronn, Martin Mynarek, Stefan Rutkowski, Ulrich Schüller, Viktor F. Mautner, Jens Schittenhelm, Jonathan Serrano, Matija Snuderl, Reinhard Büttner, Thomas Klingebiel, Rolf Buslei, Manfred Gessler, Pieter Wesseling, Winand N. M. Dinjens, Sebastian Brandner, Zane Jaunmuktane, Iben Lyskjær, Peter Schirmacher, Albrecht Stenzinger, Benedikt Brors, Hanno Glimm, Christoph Heining, Oscar M. Tirado, Miguel Sáinz-Jaspeado, Jaume Mora, Javier Alonso, Xavier Garcia del Muro, Sebastian Moran, Manel Esteller, Jamal K. Benhamida, Marc Ladanyi, Eva Wardelmann, Cristina Antonescu, Adrienne Flanagan, Uta Dirksen, Peter Hohenberger, Daniel Baumhoer, Wolfgang Hartmann, Christian Vokuhl, Uta Flucke, Iver Petersen, Gunhild Mechtersheimer, David Capper, David T. W. Jones, Stefan Fröhling, Stefan M. Pfister, and Andreas von Deimling. Sarcoma classification by DNA methylation profiling. *Nature communications*, 12(1):1–10, 2021.

[63] Keegan Korthauer, Sutirtha Chakraborty, Yuval Benjamini, and Rafael A. Irizarry. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. *Biostatistics*, 20(3):367–383, 2019.

[64] Felix Krueger. Trim Galore! A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, 2012. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (Accessed: 8 March 2022).

[65] Felix Krueger and Simon R. Andrews. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, 27(11):1571–1572, 2011.

[66] Felix Krueger, Benjamin Kreck, Andre Franke, and Simon R. Andrews. DNA methylome analysis using short bisulfite sequencing data. *Nature methods*, 9(2):145–151, 2012.

[67] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

[68] Essi Laajala, Ubaid Ullah, Toni Grönroos, Omid Rasool, Viivi Halla-aho, Mikko Konki, Roosa Kattelus, Juha Mykkänen, Mirja Nurmio, Mari Vähä-Mäkilä, Henna Kallionpää, Niina Lietzén, Bishwa R. Ghimire, Asta Laiho, Heikki Hyöty, Laura L. Elo, Jorma Ilonen, Mikael Knip, Riikka J. Lund, Matej Orešič, Riitta Veijola, Harri Lähdesmäki, Jorma Toppari, and Riitta Lahesmaa. Umbilical cord blood DNA methylation in children who later develop type 1 diabetes. *medRxiv preprint doi:10.1101/2021.05.21.21257593*, 2021.

[69] Louise Laurent, Eleanor Wong, Guoliang Li, Tien Huynh, Aristotelis Tsirigos, Chin Thing Ong, Hwee Meng Low, Ken Wing Kin Sung, Isidore Rigoutsos, Jeanne Loring, and Chia-Lin Wei. Dynamic changes in the human methylome during differentiation. *Genome research*, 20(3):320–331, 2010.

[70] Amanda J. Lea, Jenny Tung, and Xiang Zhou. A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. *PLoS genetics*, 11(11):e1005650, 2015.

[71] Wenshuai Li, Xu Zhang, Xingyu Lu, Lei You, Yanqun Song, Zhongguang Luo, Jun Zhang, Ji Nie, Wanwei Zheng, Diannan Xu, Yaping Wang, Yuanqiang Dong, Shulin Yu, Jun Hong, Jianping Shi, Hankun Hao, Fen Luo, Luchun Hua, Peng Wang, Xiaoping Qian, Fang Yuan, Lianhuan Wei, Ming Cui, Taiping Zhang, Quan Liao, Menghua Dai, Ziwen Liu, Ge Chen, Katherine Meckel, Sarbani Adhikari, Guifang Jia, Marc B. Bissonnette, Xinxiang Zhang, Yupei Zhao, Wei Zhang, Chuan He, and Jie Liu. 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. *Cell research*, 27(10):1243–1257, 2017.

[72] Xueqiu Lin, Deqiang Sun, Benjamin Rodriguez, Qian Zhao, Hanfei Sun, Yong Zhang, and Wei Li. BSeQC: quality control of bisulfite sequencing experiments. *Bioinformatics*, 29(24):3227–3229, 2013.

[73] Ryan Lister, Eran A. Mukamel, Joseph R. Nery, Mark Urich, Clare A. Puddifoot, Nicholas D. Johnson, Jacinta Lucero, Yun Huang, Andrew J. Dwork, Matthew D. Schultz, Miao Yu, Julian Tonti-Filippini, Holger Heyn, Shijun Hu, Joseph C. Wu, Anjana Rao, Manel Esteller, Chuan He, Fatemeh G. Haghighi, Terrence J. Sejnowski, M. Margarita Behrens, and Joseph R. Ecker. Global epigenomic reconfiguration during mammalian brain development. *Science*, 341(6146), 2013.

[74] Linjing Liu, Xingjian Chen, and Ka-Chun Wong. Early cancer detection from genome-wide cell-free DNA fragmentation via shuffled frog leaping algorithm and support vector machine. *Bioinformatics*, 37(19):3099–3105, 2021.

[75] Yaping Liu, Kimberly D. Siegmund, Peter W. Laird, and Benjamin P. Berman. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome biology*, 13(7):1–14, 2012.

[76] Yun Liu, Martin J. Aryee, Leonid Padyukov, M. Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, Klementy Shchetynsky, Annika Scheynius, Juha Kere, Lars Alfredsson, Lars Klareskog, Tomas J. Ekström, and Andrew P. Feinberg. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology*, 31(2):142–147, 2013.

[77] Henrik Madsen. *Introduction to general and generalized linear models*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, Boca Raton, Florida, 2010.

[78] Santiago Marco-Sola, Michael Sammeth, Roderic Guigó, and Paolo Ribeca. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature methods*, 9(12):1185–1188, 2012.

[79] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

[80] Mariana Maschietto, Laura Caroline Bastos, Ana Carolina Tahira, Elen Pereira Bastos, Veronica Luiza Vale Euclydes, Alexandra Brentani, Günther Fink, Angelica De Baumont, Aloísio Felipe-Silva, Rossana Pulcineli Vieira Francisco, Gisele Gouveia, Sandra Josefina Ferraz Ellero Grisi, Ana Maria Ulhoa Escobar, Carlos Alberto Moreira-Filho, Guilherme Vanoni Polanczyk, Euripedes Constantino Miguel, and Helena Brentani. Sex differences in DNA methylation of the cord blood are related to sex-bias psychiatric diseases. *Scientific reports*, 7(1):1–11, 2017.

[81] Tom R. Mayo, Gabriele Schweikert, and Guido Sanguinetti. M3D: a kernel-based test for spatially correlated changes in methylation profiles. *Bioinformatics*, 31(6):809–816, 2015.

[82] Angelika Merkel, Marcos Fernández-Callejo, Eloi Casals, Santiago Marco-Sola, Ronald Schuyler, Ivo G. Gut, and Simon C. Heath. gemBS: high throughput processing for DNA methylation data from bisulfite sequencing. *Bioinformatics*, 35(5):737–742, 08 2018.

[83] Fabian Model, Péter Adorján, Alexander Olek, and Christian Piepenbrock. Feature selection for DNA methylation based cancer classification. *Bioinformatics*, 17(suppl_1):S157–S164, 06 2001.

[84] Farshad Nassiri, Ankur Chakravarthy, Shengrui Feng, Shu Yi Shen, Romina Nejad, Jeffrey A. Zuccato, Mathew R. Voisin, Vikas Patil, Craig Horbinski, Kenneth Aldape, Gelareh Zadeh, and Daniel D. De Carvalho. Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. *Nature Medicine*, 26(7):1044–1047, 2020.

[85] Radford M. Neal. An improved acceptance procedure for the hybrid Monte Carlo algorithm. *Journal of Computational Physics*, 111(1):194–203, 1994.

[86] Mark S. Nixon and Alberto S. Aguado. *Feature extraction and image processing for computer vision*. Academic Press, Oxford, 3rd edition, 2012.

[87] Pier Vitale Nuzzo, Jacob E. Berchuck, Keegan Korthauer, Sandor Spisak, Amin H. Nassar, Sarah Abou Alaiwi, Ankur Chakravarthy, Shu Yi Shen, Ziad Bakouny, Francesco Boccardo, John Steinharter, Gabrielle Bouchard, Catherine R. Curran, Wenting Pan, Sylvan C. Baca, Ji-Heui Seo, Gwo-Shu Mary Lee, M. Dror Michaelson, Steven L. Chang, Sushrut S. Waikar, Guru Sonpavde, Rafael A. Irizarry, Mark Pomerantz, Daniel D. De Carvalho, Toni K. Choueiri, and Matthew L. Freedman. Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes. *Nature medicine*, 26(7):1041–1043, 2020.

[88] Yongseok Park and Hao Wu. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, 32(10):1446–1453, 2016.

[89] Brent S. Pedersen, David A. Schwartz, Ivana V. Yang, and Katerina J. Kechris. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics*, 28(22):2986–2988, 2012.

[90] Stella Pelengaris and Michael Khan. *The molecular biology of cancer*. John Wiley & Sons, 2006.

[91] Belinda Phipson, Stanley Lee, Ian J. Majewski, Warren S. Alexander, and Gordon K. Smyth. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The annals of applied statistics*, 10(2):946, 2016.

[92] Juho Piironen and Aki Vehtari. On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 905–913. PMLR, 2017.

[93] Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017.

[94] Juho Piironen and Aki Vehtari. Iterative supervised principal components. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 106–114. PMLR, 2018.

[95] Saifur Rahaman, Xiangtao Li, Jun Yu, and Ka-Chun Wong. CancerEMC: frontline non-invasive cancer screening from circulating protein biomarkers and mutations in cell-free DNA. *Bioinformatics*, 2021.

[96] Lovisa E. Reinius, Nathalie Acevedo, Maaike Joerink, Göran Pershagen, Sven-Erik Dahlén, Dario Greco, Cilla Söderhäll, Annika Scheynius, and Juha Kere. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PloS one*, 7(7):e41361, 2012.

[97] Sheldon M. Ross. Chapter 8 - hypothesis testing. In Sheldon M. Ross, editor, *Introduction to Probability and Statistics for Engineers and Scientists*, pages 297–356. Academic Press, Boston, fifth edition, 2014.

[98] Dirk Schübeler. Function and information content of DNA methylation. *Nature*, 517(7534):321–326, 2015.

[99] Heidi Schwarzenbach, Dave S. B. Hoon, and Klaus Pantel. Cell-free nucleic acids as biomarkers in cancer patients. *Nature Reviews Cancer*, 11(6):426–437, 2011.

[100] Shu Yi Shen, Justin M. Burgener, Scott V. Bratman, and Daniel D. De Carvalho. Preparation of cfMeDIP-seq libraries for methylome profiling of plasma cell-free DNA. *Nature protocols*, 14(10):2749–2780, 2019.

[101] Shu Yi Shen, Rajat Singhania, Gordon Fehringer, Ankur Chakravarthy, Michael H. A. Roehrl, Dianne Chadwick, Philip C. Zuzarte, Ayelet Borgida, Ting Ting Wang, Tiantian Li, Olena Kis, Zhen Zhao, Anna Spreafico, Tiago da Silva Medina, Yadon Wang, David Roulois, Ilias Ettayebi, Zhuo Chen, Signy Chow, Tracy Murphy, Andrea Arruda, Grainne M. O'Kane, Jessica Liu, Mark Mansour, John D. McPherson, Catherine O'Brien, Natasha Leighl, Philippe L. Bedard, Neil Fleshner, Geoffrey Liu, Mark D. Minden, Steven Gallinger, Anna Goldenberg, Trevor J. Pugh, Michael M. Hoffman, Scott V. Bratman, Rayjean J. Hung, and Daniel D. De Carvalho. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature*, 563(7732):579–583, 2018.

[102] Vivek Shukla, Xavier Coumoul, Tyler Lahusen, Rui-Hong Wang, Xiaoling Xu, Athanassios Vassilopoulos, Cuiying Xiao, Mi-Hye Lee, Yan-Gao Man, Mutsuko Ouchi, Toru Ouchi, and Chu-Xia Deng. BRCA1 affects global DNA methylation through regulation of DNMT1. *Cell research*, 20(11):1201–1215, 2010.

[103] Giulia Siravegna, Silvia Marsoni, Salvatore Siena, and Alberto Bardelli. Integrating liquid biopsies into the management of cancer. *Nature reviews Clinical oncology*, 14(9):531–548, 2017.

[104] Gordon K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

[105] Matthew W. Snyder, Martin Kircher, Andrew J. Hill, Riza M. Daza, and Jay Shendure. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell*, 164(1-2):57–68, 2016.

[106] Qiang Song, Benjamin Decato, Elizabeth E. Hong, Meng Zhou, Fang Fang, Jianghan Qu, Tyler Garvin, Michael Kessler, Jun Zhou, and Andrew D. Smith. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PloS one*, 8(12):e81148, 2013.

[107] Walter W. Stroup. *Generalized linear mixed models: modern concepts, methods and applications*. CRC press, 2012.

[108] Shiquan Sun, Jiaqiang Zhu, Sahar Mozaffari, Carole Ober, Mengjie Chen, and Xiang Zhou. Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. *Bioinformatics*, 35(3):487–496, 2018.

[109] Shuying Sun, Aaron Noviski, and Xiaoqing Yu. MethyQA: a pipeline for bisulfite-treated methylation sequencing quality assessment. *BMC bioinformatics*, 14(1):1–9, 2013.

[110] Zhifu Sun, Saurabh Baheti, Sumit Middha, Rahul Kanwar, Yuji Zhang, Xing Li, Andreas S. Beutler, Eric Klee, Yan W. Asmann, E. Aubrey Thompson, and Jean-Pierre A. Kocher. SAAP-RRBS: streamlined analysis and annotation pipeline for reduced representation bisulfite sequencing. *Bioinformatics*, 28(16):2180–2181, 2012.

[111] Zhifu Sun, Julie Cunningham, Susan Slager, and Jean-Pierre Kocher. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics*, 7(5):813–828, 2015.

[112] Oluwatosin Taiwo, Gareth A. Wilson, Tiffany Morris, Stefanie Seisenberger, Wolf Reik, Daniel Pearce, Stephan Beck, and Lee M. Butcher. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nature protocols*, 7(4):617–636, 2012.

[113] Michele M. Taylor and Susan K. Murphy. Chapter 3 - systems biology and the epigenome. In Rebecca C. Fry, editor, *Systems Biology in Toxicology and Environmental Health*, pages 43–56. Academic Press, Boston, 2015.

[114] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[115] Junko Tsuji and Zhiping Weng. Evaluation of preprocessing, mapping and postprocessing algorithms for analyzing whole genome bisulfite sequencing data. *Briefings in bioinformatics*, 17(6):938–952, 2016.

[116] Peter Ulz, Gerhard G. Thallinger, Martina Auer, Ricarda Graf, Karl Kashofer, Stephan W. Jahn, Luca Abete, Gunda Pristauz, Edgar Petru, Jochen B. Geigl, Ellen Heitzer, and Michael R. Speicher. Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nature genetics*, 48(10):1273–1278, 2016.

[117] Isabella Verdinelli and Larry Wasserman. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618, 1995.

[118] Simone Wahl, Alexander Drong, Benjamin Lehne, Marie Loh, William R. Scott, Sonja Kunze, Pei-Chien Tsai, Janina S. Ried, Weihua Zhang, Youwen Yang, Sili Tan, Giovanni Fiorito, Lude Franke, Simonetta Guarrera, Silva Kasela, Jennifer Kriebel, Rebecca C. Richmond, Marco Adamo, Uzma Afzal, Mika Ala-Korpela, Benedetta Albetti, Ole Ammerpohl, Jane F. Apperley, Marian Beekman, Pier Alberto Bertazzi, S. Lucas Black, Christine Blancher, Marc-Jan Bonder, Mario Brosch, Maren Carstensen-Kirberg, Anton J. M. de Craen, Simon de Lusignan, Abbas Dehghan, Mohamed Elkalaawy, Krista Fischer, Oscar H. Franco, Tom R. Gaunt, Jochen Hampe, Majid Hashemi, Aaron Isaacs, Andrew Jenkinson, Sujeet Jha, Norihiro Kato, Vittorio Krogh, Michael Laffan, Christa Meisinger, Thomas Meitinger, Zuan Yu Mok, Valeria Motta, Hong Kiat Ng, Zacharoula Nikolakopoulou, Georgios Nteliopoulos, Salvatore Panico, Natalia Pervjakova, Holger Prokisch, Wolfgang Rathmann, Michael Roden, Federica Rota, Michelle Ann Rozario, Johanna K. Sandling, Clemens Schafmayer, Katharina Schramm, Reiner Siebert, P. Eline Slagboom, Pasi Soininen, Lisette Stolk, Konstantin Strauch, E-Shyong Tai, Letizia Tarantini, Barbara Thorand, Ettje F. Tigchelaar, Rosario Tumino, Andre G. Uitterlinden, Cornelia van Duijn, Joyce B. J. van Meurs, Paolo Vineis, Ananda Rajitha Wickremasinghe, Cisca Wijmenga, Tsun-Po Yang, Wei Yuan, Alexandra Zhernakova, Rachel L. Batterham, George Davey Smith, Panos Deloukas, Bastiaan T. Heijmans, Christian Herder, Albert Hofman, Cecilia M. Lindgren, Lili Milani, Pim van der Harst, Annette Peters, Thomas Illig, Caroline L. Relton, Melanie Waldenberger, Marjo-Riitta Järvelin, Valentina Bollati, Richie Soong, Tim D. Spector, James Scott, Mark I. McCarthy, Paul Elliott, Jordana T. Bell, Giuseppe Matullo, Christian Gieger, Jaspal S. Kooner, Harald Grallert, and John C. Chambers. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature*, 541(7635):81–86, 2017.

[119] Jonathan C. M. Wan, Charles Massie, Javier Garcia-Corbacho, Florent Mouliere, James D. Brenton, Carlos Caldas, Simon Pacey, Richard Baird, and Nitzan Rosenfeld. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature Reviews Cancer*, 17(4):223–238, 2017.

[120] Nathan Wan, David Weinberg, Tzu-Yu Liu, Katherine Niehaus, Eric A. Ariazi, Daniel Delubac, Ajay Kannan, Brandon White, Mitch Bailey, Marvin Bertin, Nathan Boley, Derek Bowen, James Cregg, Adam M. Drake, Riley Ennis, Signe Fransen, Erik Gafni, Loren Hansen, Yaping Liu, Gabriel L. Otte, Jennifer Pecson, Brandon Rice, Gabriel E. Sanderson, Aarushi Sharma, John St. John, Catherina Tang, Abraham Tzou, Leilani Young, Girish Putcha, and Imran S. Haque. Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. *BMC cancer*, 19(1):1–10, 2019.

[121] Michael Weber, Jonathan J. Davies, David Wittig, Edward J. Oakeley, Michael Haase, Wan L. Lam, and Dirk Schuebeler. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature genetics*, 37(8):853–862, 2005.

[122] Yalu Wen, Fushun Chen, Qingzheng Zhang, Yan Zhuang, and Zhiguang Li. Detection of differentially methylated regions in whole genome bisulfite sequencing data using local Getis-Ord statistics. *Bioinformatics*, 32(22):3396–3404, 2016.

[123] Samuel S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938.

[124] Yuanxin Xi and Wei Li. BSMAP: whole genome bisulfite sequence MAPping program. *BMC bioinformatics*, 10(1):1–9, 2009.

[125] Yang Xie, Wei Pan, and Arkady B. Khodursky. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, 21(23):4280–4288, 2005.

[126] Rui-hua Xu, Wei Wei, Michal Krawczyk, Wenqiu Wang, Huiyan Luo, Ken Flagg, Shaohua Yi, William Shi, Qingli Quan, Kang Li, Lianghong Zheng, Heng Zhang, Bennett A. Caughey, Qi Zhao, Jiayi Hou, Runze Zhang, Yanxin Xu, Huimin Cai, Gen Li, Rui Hou, Zheng Zhong, Danni Lin, Xin Fu, Jie Zhu, Yaou Duan, Meixing Yu, Binwu Ying, Wengeng Zhang, Juan Wang, Edward Zhang, Charlotte Zhang, Oulan Li, Rongping Guo, Hannah Carter, Jian-kang Zhu, Xiaoke Hao, and Kang Zhang. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nature materials*, 16(11):1155–1161, 2017.

[127] Xi Yang, Di Liu, Fei Liu, Jun Wu, Jing Zou, Xue Xiao, Fangqing Zhao, and Baoli Zhu. HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC bioinformatics*, 14(1):1–4, 2013.

[128] Paul Yousefi, Karen Huen, Veronica Davé, Lisa Barcellos, Brenda Eskenazi, and Nina Holland. Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. *BMC genomics*, 16(1):1–12, 2015.

[129] Dmitri V. Zaykin. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of evolutionary biology*, 24(8):1836–1841, 2011.

[130] Michael J. Ziller, Fabian Müller, Jing Liao, Yingying Zhang, Hongcang Gu, Christoph Bock, Patrick Boyle, Charles B. Epstein, Bradley E. Bernstein, Thomas Lengauer, Andreas Gnirke, and Alexander Meissner. Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS genetics*, 7(12):e1002389, 2011.

[131] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Aalto-DT 124/2022

ISBN 978-952-64-0927-6 (printed)
ISBN 978-952-64-0928-3 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL THESES