Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Hao Feng	Date

Statistical Methods for High-throughput Epigenomics Data

By

Hao Feng

Doctor of Philosophy

Biostatistics

Hao Wu, Ph.D.
Advisor

Zhaohui Steve Qin, Ph.D.
Committee Member

Peng Jin, Ph.D.
Committee Member

Karen Conneely, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Statistical Methods for High-throughput Epigenomics Data

By

Hao Feng
MS, Emory University, 2018
MSPH, Emory University, 2013
BS, University of Science and Technology of China, 2011

Advisor: Hao Wu, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2019

Abstract

Statistical Methods for High-throughput Epigenomics Data

By

Hao Feng

DNA methylation is an important epigenetic modification that has essential roles in biological and clinical processes including gene regulation, development and disease. Aberrant and unique DNA methylation patterns have been identified in various diseases such as cancer, making DNA methylation an ideal biomarker. Recently, various high-throughput technologies have emerged to measure genome-wide epigenomics profiles. However, due to the novelty of the technologies and special characteristics of the high-throughput DNA methylation data, there lacks rigorous and effective statistical methods to examine DNA methylation thoroughly.

The coherent theme of this dissertation is to develop novel statistical models and data analysis strategies for high-throughput epigenomics data. In particular, I propose several model-based methods for studying DNA methylation and its relationship to disease. My first research topic aims at classifying tumors into different subtypes based on their methylation profiles, which can facilitate the application of precision medicine on patients. In practice, the data obtained from clinical samples are mixed signals. The proportion of cancer cells in the mixture, known as the tumor purity, will bias the clustering results if not properly accounted for. In this work, I develop a model-based clustering method to infer tumor subtypes with the consideration of tumor purity.

Moving from solid tumor samples to blood assay, my second research topic aims at using cell-free DNA (cfDNA) methylation data to detect disease. Recent researches start to exploit the epigenetic information on cfDNA, which could have broad applications. In this work, I provide thorough reviews and discussions on the statistical method developments and data analysis strategies for using cfDNA epigenetic profiles, in particular DNA methylation, to construct disease diagnostic models.

Along the trajectory of studying cfDNA, my third research topic aims at investigating another type of epigenetic marker: 5-hydroxymethylcytosine (5hmC). Currently, little is known about the 5hmC epigenetic profile on cfDNA. Here, I investigate the genome-wide alteration of cfDNA 5hmC in young healthy subjects, old healthy subjects and late onset Alzheimers disease (AD) patients. This is the first investigation, both experimentally and computationally, to study the cfDNA 5hmC profile of neurodegenerative disease and its potential as a diagnostic biomarker.

Statistical Methods for High-throughput Epigenomics Data

By

Hao Feng
MS, Emory University, 2018
MSPH, Emory University, 2013
BS, University of Science and Technology of China, 2011

Advisor: Hao Wu, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2019

Acknowledgement

I want to express my sincere thanks to my dissertation advisor, Dr. Hao Wu. I started to work with Dr. Wu eight years ago. During the past, he guided me through coursework, through research, through ups and downs of my academic life. He is the person who introduced me to genomics research world. He is the person who always brought passion to our meetings. He is the person who always made himself available to me. I am grateful for his help along the way.

I want to thank my dissertation committee members, Steve Qin, Peng Jin and Karen Conneely. They generously offered their time, effort and guidance to train me during my graduate study.

I want to thank the Department of Biostatistics and Bioinformatics at Emory University for having me and nourishing me. The faculty members are supportive and accessible. I want to thank my classroom teachers and mentors: John Hanfelt, Yijuan Hu, Eugene Huang, Michael Kutner, Qi Long, Bob Lyles, Amita Manatunga, Limin Peng, Lance Waller and Tianwei Yu. My special thanks go to the awesome staff members at the Department of Biostatistics and Bioinformatics: Melissa Sherrer, Mary Abosi, Bob Waggoner and Angela Guinyard. They made my graduate study pleasant and delightful. I could not make this far without their help.

I want to thank my student peers Ziyi Li, Ben Li, Tianlei Xu and Li Chen. They made my graduate study life delightful and enjoyable. The time we spent together, inside and outside the classroom, will be the most cherished memory in my life.

Finally, I want to thank my parents for their unconditional support throughout my life. I also want to thank my wife, Yufei Li, for being extremely supportive and having unwavering belief in me. Without my family, I would not be the person I am today.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 DNA Methylation

DNA methylation is an important epigenetic modification of the DNA molecule, and plays a crucial role in many biological processes, including repression of gene transcription, maintenance of gene imprinting and X-chromosome inactivation (Bird, 2002; Hackett and Surani, 2013; Li et al., 1993). It involves the addition of a methyl group to the 5-position of a cytosine of CpG dinucleotides, with very rare cases that happen in CHG and CHH (H = A, T or C) (Lister et al., 2009). Methylation of a cytosine within a gene promoter region can repress gene expression by interfering with the binding of transcription factors or by binding proteins that inhibit transcription (Bird and Wolffe, 1999; Hendrich and Bird, 1998), while methylation within gene bodies has a heterogeneous relationship with gene expression (Cokus et al., 2008; Lister et al., 2008; Wang et al., 2013). Given its influence on gene expression, both the biological consequences and causes of changes in DNA methylation are of great interest.

Recently, methods for assessing methylation have improved substantially in terms of accuracy, genomic coverage, resolution and affordability. The measurement for DNA methylation can be categorized into two categories: array-based approach and sequencing-based approach. The array-based methods adopt the workflow of microarray, which measures the hybridization strength signal on pre-designed regions. Those regions are targeted to methylated or potentially methylated regions such as CpG islands, promoters, etc. For example, Illumina Infinium HumanMethylation27 BeadChip array probes the methylation level at $\sim 27,000$ CpG dinucleotides. The later verion Infinium HumanMethylation450 BeadChip array measures more than $450,000$ CpG sites. Current sequencing-based methods for methylation analysis can be further classified into two categories: enrichment- (Taiwo et al., 2012) and bisulfite-conversion-based methods (Harris et al., 2010). Enrichment-based methods such as

MeDIP-seq (Taiwo et al., 2012), MBD-seq (Serre et al., 2009; Rauch and Pfeifer, 2010) and methylCap-seq use different methyl-binding proteins or antibodies to enrich for methylated DNA fragments, followed by the application of next-generation sequencing of the fragments and alignment to a reference genome to estimate methylation levels at a $100 \sim 200$-bp resolution. In contrast, bisulfite-conversion-based methods such as whole-genome bisulfite sequencing (BS-seq or MethylC-seq) (Lister et al., 2008; Frommer et al., 1992) and reduced representation bisulfite sequencing (RRBS) (Meissner et al., 2008; Smith et al., 2009) allow estimation of methylation proportions at a single-nucleotide resolution. Treatment of DNA with sodium bisulfite induces deamination and conversion of unmethylated cytosines to uracil, which will be amplified as thymine, while methylated cytosines are protected by the methyl group and remain unchanged. Bisulfite sequencing data can be analyzed by counting the number of sequencing reads for each CpG site where either a thymine or a cytosine is observed. The count of thymine represents the number of sequenced DNA strands that are unmethylated (U) and the count of cytosine represents the number of DNA strands that are methylated (M) at this CpG site. By taking the ratio of methylated number (M) to the total number of reads (M + U), the proportion of methylated DNA can be calculated as $\frac{M}{M+U}$. By this process, DNA methylation proportions can be estimated at single-nucleotide resolution with genome-wide coverage via BS-seq, or with limited coverage ($5 - 10\%$ of all CpG sites genome-wide) via RRBS.

1.2 Tumor Subtype Classification

Classifying tumor samples into subtypes based on different types of clinical or molecular data is a key step in understanding cancer etiology and designing personalized treatment for cancer patients (Chung et al., 2002; Hoadley et al., 2014; Ogino et al., 2012). Originally, classification of cancer subtype was mostly based on clinical histological information. For example, according to the size and the appearance of malig-

nant cells under a microscope, lung carcinomas are categorized into two main classes: non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC) (Leong et al., 2014). With the advances of high-throughput technologies, tumor subtype classification has been performed more frequently using molecular signals such as DNA sequence variants, gene expressions or DNA methylation. For example, the PAM50 gene expression assay was used to categorize breast tumors into five intrinsic subtypes: luminal A, luminal B, human epidermal growth factor receptor2 (HER2) enriched, basal-like and normal-like (Parker et al., 2009). Similarly, glioblastoma multiforme was classified into four molecular subtypes: classical, neural, proneural and mesenchymal, where the former two were characterized by higher expression of epidermal growth factor receptor (EGFR) and neuron maker genes, respectively (Verhaak et al., 2010).

Aberrant DNA methylation pattern has been identified as a hallmark in different types of cancers (Das and Singal, 2004; Hansen et al., 2011). DNA methylation profile has widely been used to perform categorization on clinical presentation and patient prognosis (Stefansson et al., 2015; Zhuang et al., 2012). Clustering of lung cancer cell lines using DNA methylation markers showed that NSCLC and SCLC cell lines had different DNA methylation patterns (Virmani et al., 2002). DNA methylation profiling was also used in clustering of acute lymphoblastic leukemia (ALL) patients and served as a complementary method for diagnosis of ALL (Nordlund et al., 2015). These results suggest that each cancer subtype carries unique DNA methylation signature that can help to identify the subtypes.

1.3 Cell-Free DNA

Prognosis and diagnosis play vital roles in the prevention and treatment of diseases. Traditionally, various types of surgical biopsies such as bone marrow or needle biopsies are performed in clinical setting, especially for cancer diagnosis (Sgouros, 1993).

However, due to the invasive nature of the procedure and the potential sampling bias of tumor biopsy, surgical biopsy is often not a preferred choice. As an alternative to surgical biopsy, researchers and clinicians have been looking for molecular biomarkers for disease diagnosis. These biomarkers, either genetic or epigenetic, carry various indicative features for biological or disease states. They help achieve disease state detection, subtypes classification, progression prediction, and response-to-treatment characterization (AJ et al., 2001). The scope of molecular biomarker discovery has been greatly expanded during the last two decades, due to the advances of high-throughput genomics technologies such as microarray and next-generation sequencing (NGS). For example, based on gene expression microarray data, Prediction Analysis for Microarrays (PAM) identified a subset of gene biomarkers for cancer class prediction (Tibshirani et al., 2002). The PAM50 panel, which tests a group of 50 selected genes, has become the *de facto* gold standard for breast cancer subtype classification and metastasis prediction (Parker et al., 2009). Zilliox et al. created the gene expression barcode (Zilliox and Irizarry, 2007), which was trained on public gene expression microarray data, and can predict for a number of diseases given a new microarray dataset.

In recent years, disease diagnosis based on molecular biomarkers in specimen, including blood, urine, and cerebral spinal fluid, has gain tremendous attention. For example, the practice to look for traces of cancer DNA by interrogating biomarkers on plasma-isolated cell-free DNA (cfDNA) or circulating tumor DNA (ctDNA) is known as liquid biopsy (Crowley et al., 2013). As a safer, cheaper, and quicker alternative to surgical biopsy, the liquid biopsy has great potential in clinical practice. Cell-free DNA are short DNA fragments (around 160-180 base pairs) existing in plasma. When normal cells undergo apoptosis in a healthy individual, DNA fragments from the cells are shredded and released to blood stream. Thus, cfDNA is a mixture of DNA fragments from different cell types. In cancer patients, the cfDNA includes

some circulating tumor DNA (ctDNA), which are DNA fragments released from cancer cells (Schwarzenbach et al., 2011). As a hallmark of cancer, the ctDNA carries tumor-specific genetic variants such as copy number variation and point mutations. After capturing and sequencing the cfDNA, ctDNA can be distinguished from normal cfDNA by tumor-specific genetic variants. Presence of non-trivial amount of ctDNA is an indication of cancer.

The essence of using cfDNA for cancer diagnosis is to detect abnormal cfDNA segments. Here the abnormality is defined as the presents of unusual genetic variants. This principal, however, is only applicable to mutation-rich diseases - the ones with high rate of genetic alteration such as cancer. For many mutation-poor diseases and disorders not associated with high level of genetic alteration, other approaches are needed. Recently, researchers start to explore the cfDNA epigenetics information such as DNA methylation or nucleosome position to look for biomarkers for diagnosis (Kang et al., 2017; Xu et al., 2017; Song et al., 2017; Li, Zhang, Lu, You, Song, Luo, Zhang, Nie, Zheng, Xu et al., 2017). In analogous to the liquid biopsy, these approaches try to define abnormality based on the epigenetic profiles, and then construct model for disease prediction.

1.4 Overarching Goal and Outline

Selecting genomic biomarkers for disease prediction and classification showed good potential and gained popularity over the last decade. In the past, researchers identified genetic and genomic biomarkers for a variety of diseases, including asthma (Allen et al., 2003), breast cancer (Van De Vijver et al., 2002; Weigelt et al., 2003), cervical cancer (Wong et al., 2003), leukemia (Ross et al., 2003), ovarian cancer (Petricoin III et al., 2002), diabetes (Bell and Polonsky, 2001; Gloyn et al., 2003), leprosy (Bell, 2004), cardiovascular disease (Brindle et al., 2002), among other disease types. In terms of high-throughput platforms, there has been a success in using genetic variant

(Wray et al., 2007; MacArthur et al., 2014; Paynter et al., 2009), gene expression (Zilliox and Irizarry, 2007; Van De Vijver et al., 2002; Weigelt et al., 2003; Wong et al., 2003; Ross et al., 2003), and protein (Petricoin III et al., 2002; Weissinger et al., 2007; Schirmer et al., 2003) information for exploring disease-associated biomarkers. These biomarker studies have provided powerful insights into the perception of diseases. The identified biomarkers can guide future biomedical researches for disease etiology. Applied clinical tools for disease detection and early diagnosis is now available, with the identified biomarkers across the platforms. Biomarkers can potentially benefit targeted disease therapy and personlized medicine for each patient.

Although the usage of genetic, gene expression or proteomic biomarker have been studied, there is a lack of methods and tools for using epigenomics information for disease diagnosis and classification. Here I built statistical models to investigate the usage of epigenomic biomarker. This can bring new insights and enhance clinical practice of various disease types. Translating the epigenomic knowledge into routinely applied diagnostic tools will strengthen our understanding of epigenomics, and bring benefits for patients. With the increasing of high-throughput platforms usage and the decreasing of the cost of assay, the size of epigenomics data will grow over the next decades. I can expect a wide application of epigenomic biomarker for disease diagnosis and health status mornitoring.

Aiming at exploring epigenomic biomarkers for disease prediction and classification, in this dissertation, I present some statistical methods for deciphering epigenomics data. The outline of this dissertation is as follows. In chapter 2, I propose a statistical method *InfiniumClust* to perform cancer subtype clustering on DNA methylation data with the consideration of tumor purity. The sample clustering are performed through an Expectation-Maximization (EM) algorithm. In chapter 3, I focus on cfDNA methylation and explore how it can be used for disease prediction. I review the existing publications and investigate the statistical methods for cfDNA

methylation study. I conduct simulation and real data analysis, and provide some recommendations for data analysis strategies based on the results. In chapter 4, I investigate the genome-wide alteration of cfDNA 5hmC in young healthy subjects, old healthy subjects and late onset Alzheimer's disease (AD) patients. I constructed a classification machine learning model to predict AD from healthy old individuals. In chapter 5, I outlined several potential research directions that I want to pursue in the near future.

Chapter 2

Accounting for Tumor Purity Improves Cancer Subtype Classification from DNA Methylation Data

2.1 Introduction

Classifying tumor samples into subtypes based on different types of clinical or molecular data is a key step in understanding cancer etiology and designing personalized treatment for cancer patients (Chung et al., 2002; Hoadley et al., 2014; Ogino et al., 2012). Originally, classification of cancer subtype was mostly based on clinical histological information. With the advances of high-throughput technologies, tumor subtype classification has been performed more frequently using molecular signals such as DNA sequence variants, gene expressions or DNA methylation.

DNA methylation, as an important epigenetic modification of the DNA, carries signatures in different cancers. Aberrant DNA methylation pattern has been identified as a hallmark in different types of cancers (Das and Singal, 2004; Hansen et al., 2011). DNA methylation profile has widely been used to perform categorization on clinical presentation and patient prognosis (Stefansson et al., 2015; Zhuang et al., 2012). Clustering of lung cancer cell lines using DNA methylation markers showed that NSCLC and SCLC cell lines had different DNA methylation patterns (Virmani et al., 2002). DNA methylation profiling was also used in clustering of acute lymphoblastic leukemia (ALL) patients and served as a complementary method for diagnosis of ALL (Nordlund et al., 2015). These results suggest that each cancer subtype carries unique DNA methylation signature that can help to identify the subtypes.

A number of methods have been applied for clustering tumor samples based on high-throughput data, including nonparametric (K-means, agglomerative hierarchical clustering, etc.) and model-based methods (Houseman et al., 2008; Kuan et al., 2010). In particular, Non-negative Matrix Factorization (NMF) is a popular method for sample clustering based on gene expression data (Brunet et al., 2004). The method is based on matrix factorization with non-negative constraints, and was shown to have good performance. To systematically compare these methods, a recently developed

tool ClustEval evaluated the currently available clustering methods by using different datasets, varying parameters, and quality metrics. It suggested that no method performed the best in all settings (Wiwie et al., 2015).

Among all published results for cancer type classification, one important aspect is consistently ignored: the clinical tumor samples contain different types of cells as well as their adjacent normal cells. Due to the inclusion of normal cells in the tumor samples, the clinical tumor samples cannot be regarded as pure cancer cells. Previous studies have shown that tumor purities (the percentages of cancer cells in solid tumor samples) have a strong influence on the analysis of genomic data in cancer studies, and may bias the biological interpretation of results (Aran et al., 2015). Our exploratory analyses also show that applying traditional clustering methods such as K-means or NMF directly on the methylation profiles from tumor samples gives biased results (more details are provided in Section 2.2): samples with similar tumor purities tend to be clustered together. This is undesirable since there is no evidence showing associations between tumor purity and cancer subtypes. Thus, we believe it is of great necessity to consider tumor purity in the clustering procedure.

The importance of accounting for tumor purity in data analysis has been well recognized. For example, it is recommended to include purity in differential expression analysis (Aran et al., 2015). We recently developed InfiniumPurify, which incorporates purity in differential methylation (DM) analysis (Zheng et al., 2017). However, up to date there is no clustering method available with consideration of tumor purity. In this study, we developed a rigorous statistical method InfiniumClust to perform sample clustering on DNA methylation data with the consideration of tumor purity. InfiniumClust models the DNA methylation levels of a tumor sample as a mixture of normal and cancer data, where the mixing proportion is the tumor purity. The pure cancer data are further assumed to be from a mixture of different cancer subtypes. When tumor purities are known, the parameter estimation and sample clustering are

performed through an Expectation-Maximization (EM)-based algorithm. We performed extensive real-data based simulations and demonstrated good performances with InfiniumClust. We further applied InfiniumClust to DNA methylation data for 23 cancer types from The Cancer Genome Atlas (TCGA). Compared with existing clustering methods that ignore tumor purity, InfiniumClust provides less biased and more meaningful results. To our best knowledge, InfiniumClust is the first available tool for unsupervised clustering which taking tumor purity into account. InfiniumClust is currently available from R package ***InfiniumPurify***, which can be obtained at https://cran.r-project.org/web/packages/InfiniumPurify/index.html.

2.2 Materials and Methods

Sample clustering based on high-throughput data usually starts with feature (CpG site, gene, etc.) selection. It is a common practice to select a small number (such as 1000) of features with the largest variances, and use their data for clustering. Those features are highly heterogeneous, thus they contain information for subtypes. In contrast, using data for all features is not ideal because a large portion of them show no variation among samples, thus bringing noise to the clustering procedure. Under the above analyses, we first selected some highly variable CpG sites and then performed clustering based on these feature-reduced data.

The raw data for the clustering procedure are from Illumina Infinium DNA methylation $450k$ arrays, which report methylation beta values of more than $480,000$ CpG sites. Methylation beta values range from 0 to 1, hence they cannot be considered as froming a normal distribution. We first transformed the beta values using an arcsine transformation:

$$f(x) = arcsin(2x - 1)$$

Such transformation has previously been used in DM analysis (Park and Wu, 2016).

The transformed data follow the normal distribution better compared with the raw data, thus fitting our model assumption well. In addition, compared with commonly used logit transform, the arcsine transformation is more linear (especially at the boundaries). This is important since the methylation level from tumor sample is a mixture of the cancer and normal methylation level, and the signal mixing is at the original scale. A more linear transformation allows one to use a linear model for the transformed data with a better approximation.

2.2.1 The Data Model

Let $\mathbf{X}$, $\mathbf{Y}$ be $C \times N$ matrices of transformed methylation levels for normal and pure tumor cells, where $i = 1, 2, \ldots, C$ indexes CpG sites, and $j = 1, 2, \ldots, N$ indexes samples. Assuming tumor samples have K subtypes with proportions p_k, and satisfy $\sum_{k=1}^{K} p_k = 1$. Define a latent indicator variable $\mathbf{Z}$ as the membership of samples, i.e. $Z_{jk} = 1$ means the j^{th} pure tumor cell comes from the subtype k. Apparently, each sample can only belong to one cancer subtype, so

$$\sum_{k=1}^{K} Z_{jk} = 1$$

We assume that the transformed methylation level of CpG site i in normal cells j follows the normal distribution:

$$X_{ij} \sim N(\mu_{i0}, \sigma_{i0}^2)$$

The transformed methylation level at CpG site i in tumor cells j clustering into subtype k is assumed to follow a mixture of normal distributions:

$$Y_{ij} | Z_{jk} = 1 \sim N(\mu_{ik}, \sigma_{ik}^2)$$

where different subtypes have different means and variances. In practice, clinical tumor samples are affected by tumor purity, so methylation for pure tumor cells is unobserved. Instead, the observed data from clinical tumor samples, denoted by $Y_{ij}^{'}$, is from mixed cancer-normal tissues. For tumor sample j, let λ_j be the tumor purity, we have

$$Y_{ij}^{'} = \lambda_j Y_{ij} + (1 - \lambda_j) X_{ij}$$

Assuming that X_{ij} and Y_{ij} are independent, we have the distribution for the mixed signal as

$$Y_{ij}^{'} | Z_{jk} = 1 \sim N(\lambda_j \mu_{ik} + (1 - \lambda_j)\mu_{i0}, \lambda_j^2 \sigma_{ik}^2 + (1 - \lambda_j)^2 \sigma_{i0}^2)$$

The data model shows that due to the presence of cancer/normal sample mixing, directly clustering $Y_{ij}^{'}$ could lead to biased results. In the next section, we present a model-based clustering algorithm where tumor purities are considered.

2.2.2 Model-Based Clustering Method

For our method presented below, the tumor purities λ_j are assumed to be known. There are a number of methods available for purity estimation (Ahn et al., 2013; Bao et al., 2014; Carter et al., 2012; Yoshihara et al., 2013), and an informative review is presented by Wang et al. (2016). After obtaining the tumor purities λ_j from existing methods, the clustering problem model transforms into a K-component normal mixture model.

We develop the following method, termed InfiniumClust, to cluster methylation beta values from $450k$ arrays. In the clustering problem, the input data are $Y_{ij}^{'}$ and λ_j. Denote the parameter set to be estimated as

$$\boldsymbol{\theta} = (p_1, p_2, \ldots, p_{K-1}; \mu_{11}, \ldots, \mu_{CK}; \mu_{10}, \ldots, \mu_{C0}; \sigma_{11}^2, \ldots, \sigma_{CK}^2; \sigma_{10}^2, \ldots, \sigma_{C0}^2)$$

In detail, $p_1, p_2, \ldots, p_{K-1}$ are mixing proportions, $\mu_{11}, \ldots, \mu_{CK}; \sigma_{11}^2, \ldots, \sigma_{CK}^2$ are means

and variances of each mixing cancer component and $\mu_{10}, \ldots, \mu_{C0}; \sigma^2_{10}, \ldots, \sigma^2_{C0}$ are means and variances of the normal cells. Under these setups, the clustering problem can be performed through the following EM algorithm.

First, the conditional likelihood for observing the methylation status of sample j is

$$p(Y_{ij}^{'}|Z_{jk}=1) = p_k\phi(Y_{ij}^{'}; \lambda_j\mu_{ik} + (1-\lambda_j)\mu_{i0}, \lambda_j^2\sigma_{ik}^2 + (1-\lambda_j)^2\sigma_{i0}^2)$$

where ϕ is the probability density function (pdf) of standard normal distribution. Treating Z_{jk} as missing data, the joint likelihood of the observed and missing data for CpG site i and sample j is

$$p(Y_{ij}^{'}|Z_j) = \prod_{k=1}^{K}\{p_k\phi(Y_{ij}^{'}; \lambda_j\mu_{ik} + (1-\lambda_j)\mu_{i0}, \lambda_j^2\sigma_{ik}^2 + (1-\lambda_j)^2\sigma_{i0}^2)\}^{Z_{jk}}$$

So the complete data log-likelihood for parameters is

$$\ell(\boldsymbol{\theta}; \boldsymbol{Y}', \boldsymbol{Z}) = \sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{K} Z_{jk}(log[p_k] + log[\phi(Y_{ij}^{'}; \lambda_j\mu_{ik} + (1-\lambda_j)\mu_{i0}, \lambda_j^2\sigma_{ik}^2 + (1-\lambda_j)^2\sigma_{i0}^2)])$$

In EM algorithm, the E-step involves calculating the conditional expectation of the complete data log-likelihood, which gives the objective Q function as

$$\begin{aligned} &Q(\boldsymbol{\theta}|\boldsymbol{\theta^{(t)}}) = E_{\boldsymbol{\theta^{(t)}}}\{\ell(\boldsymbol{\theta}; \boldsymbol{Y}', \boldsymbol{Z})|\boldsymbol{Y}'\} \\ &= \sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{K}\{E_{\boldsymbol{\theta^{(t)}}}(Z_{jk}|\boldsymbol{Y}')(log[p_k] + log[\phi(Y_{ij}^{'}; \lambda_j\mu_{ik} + (1-\lambda_j)\mu_{i0}, \lambda_j^2\sigma_{ik}^2 + (1-\lambda_j)^2\sigma_{i0}^2)])\} \end{aligned}$$

E-step calculates the expected value of Z_{jk} conditional on the observed data and the parameter values at the current step, denoted by $\boldsymbol{\theta^{(t)}}$. At current iteration step t,

denote the expected value of Z_{jk} as Z_{jk}^t. E-step gives

$$
\begin{aligned}
Z_{jk}^{(t)} &\equiv E_{\boldsymbol{\theta^{(t)}}}(Z_{jk}|\boldsymbol{Y'}) \\
&= \frac{p_k^{(t)} \prod_{i=1}^{C} \phi(Y'_{ij}; \lambda_j \mu_{ik} + (1-\lambda_j)\mu_{i0}, \lambda_j^2 \sigma_{ik}^2 + (1-\lambda_j)^2 \sigma_{i0}^2)}{\sum_{k=1}^{K} \{p_k^{(t)} \prod_{i=1}^{C} \phi(Y'_{ij}; \lambda_j \mu_{ik} + (1-\lambda_j)\mu_{i0}, \lambda_j^2 \sigma_{ik}^2 + (1-\lambda_j)^2 \sigma_{i0}^2)\}}
\end{aligned}
$$

The M-step maximizes the conditional expectation of the objective function $Q(\boldsymbol{\theta}|\boldsymbol{\theta^{(t)}})$ with respect to current-step parameters. For the updates of μ_{ik} and μ_{i0}, we have

$$A - B\mu_{ik}^{(t+1)} - C\mu_{i0}^{(t+1)} = 0$$

and

$$A\mu_{i0}^{(t+1)} = B$$

where

$$A = \sum_{j=1}^{N} \frac{Z_{jk}^{(t)} \lambda_j Y'_{ij}}{\lambda_j^2 \sigma_{ik}^{2(t)} + (1-\lambda_j)^2 \sigma_{i0}^{2(t)}}$$

$$B = \sum_{j=1}^{N} \frac{Z_{jk}^{(t)} \lambda_j^2}{\lambda_j^2 \sigma_{ik}^{2(t)} + (1-\lambda_j)^2 \sigma_{i0}^{2(t)}}$$

$$C = \sum_{j=1}^{N} \frac{Z_{jk}^{(t)} \lambda_j (1-\lambda_j)}{\lambda_j^2 \sigma_{ik}^{2(t)} + (1-\lambda_j)^2 \sigma_{i0}^{2(t)}}$$

Combining the $K+1$ equations, we can obtain the updates for μ_{ik} and μ_{i0}. In practice, we update the μ_{ik} and μ_{i0} one-by-one for each CpG site and for each group, by solving one equation at a time. So the maximization of μ is essentially a conditional maximization step in the Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993). For the updates for $\sigma_{ik}^2, \sigma_{i0}^2$, the M-step is more challenging. The partial derivatives of Q function with respect to σ_{ik}^2 or σ_{i0}^2 show that the updates of $\sigma_{ik}^2, \sigma_{i0}^2$ exist on both numerator and denominator of a summation over sample j from 1 to N. Therefore, the closed form solutions for updating σ_{ik}^2 and σ_{i0}^2 do not

exist and they have to be solved numerically. We adopted the optimize function in R directly on objective likelihood function $Q(\boldsymbol{\theta}|\boldsymbol{\theta^{(t)}})$ to update σ_{ik}^2 and σ_{i0}^2.

The update for p_k can be achieved by solving the following partial derivative

$$\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial p_k} = \sum_{i=1}^{C}\sum_{j=1}^{N}\{\frac{Z_{jk}^{(t)}}{p_k} - \frac{Z_{jk}^{(t)}}{1-\sum_{l=1}^{K-1} p_l}\} = 0$$

thus

$$p_k^{(t+1)} = \frac{\sum_{j=1}^{N} Z_{jk}^{(t)}}{\sum_{k=1}^{K}\sum_{j=1}^{N} Z_{jk}^{(t)}}$$

The EM algorithm starts with initial values obtained from the K-means clustering directly performed on $Y_{ij}^{`}$. The results from the EM procedure provide the posterior probabilities of each sample being in each subtype, which can be used to determine the subtype assignments.

2.3 Results

For all the results presented in this section, the estimated purities are provided by InfiniumPurify (Zhang et al., 2015) and obtained from https://zenodo.org/record/253193.

2.3.1 DNA methylation subtypes are biased by tumor purities

We first explored the real data to check whether tumor purity tends to bias the clustering results by comparing the purity distributions among clusters. We focus our attention on breast cancer (BRCA) since it has mature clinical subtypes known as Luminal A, Luminal B, HER2-enriched, Basal-like and Normal-like, which are characterized by expressions of ER, PR and Her2. We also downloaded the consensus clustering results by NMF (cNMF) using DNA methylation $450k$ array data, and performed K-means clustering on the same data set. We tried two sets of estimated

tumor purities including the InfiniumPurify purities, which are estimated from DNA methylation $450k$ array data, and ABSOLUTE purities are based on SNP array data. The later ABSOLUTE purity estimates are actually the de facto gold standards provided by TCGA. These two types of tumor purities are shown to be highly correlated (Zhang et al., 2015). In the following analysis, we examined tumor purities distribution from both InfiniumPurify and ABSOLUTE on clusters of K-means, cNMF and PAM50, respectively.

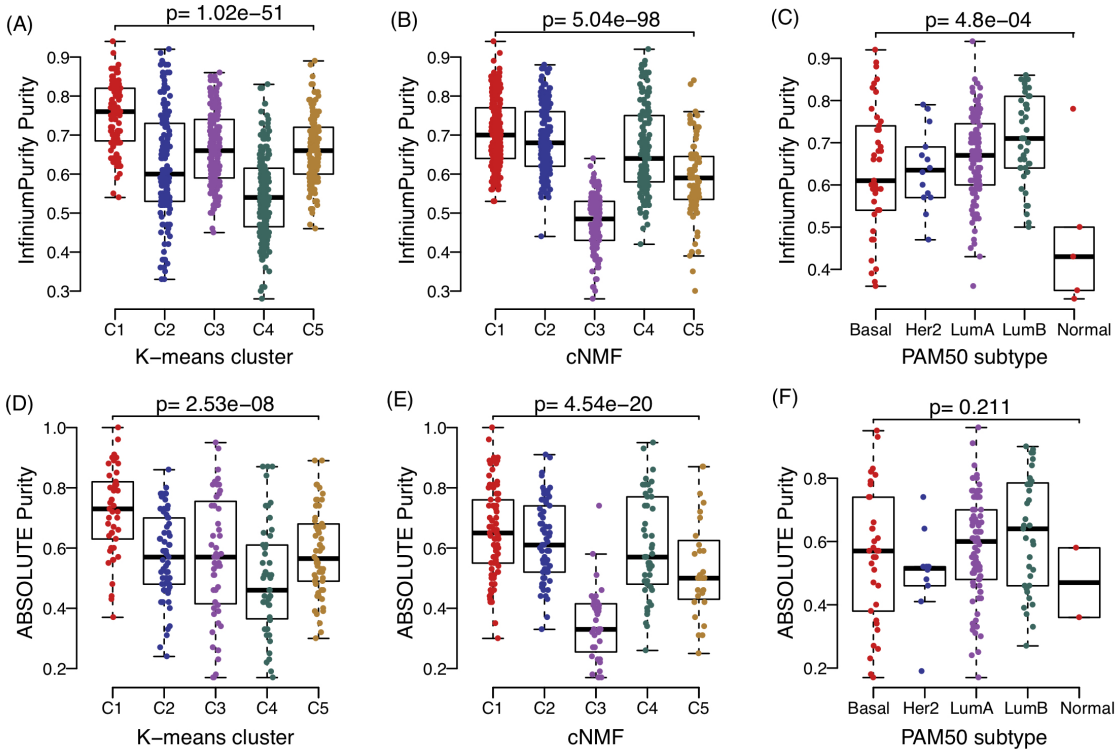


Figure 2.1: Purity distribution for different BRCA subtypes. (**A**-**C**) InfiniumPurify purity distribution on K-means clustering subtypes, cNMF subtypes and PAM50 molecular subtypes. (**D**-**F**) ABSOLUTE purity distribution on K-means clustering subtypes, cNMF subtypes and PAM50 molecular subtypes. P-values are from linear regression with ANOVA F-test

Figure 2.1 presents the purity distributions of different clusters obtained from the above three methods. We observed significant purity differences using InfiniumPurify among different subtypes from K-means, PAM50 and especially cNMF (Figure 2.1B). For example, the third cluster in cNMF has an averaged purity of 0.5, way below

other groups. Figure 2.1C shows the results from PAM50 subtypes, where the purity differences show significance (p-value=4.8×10^{-4}), for example, two luminal subtypes, especially luminal B, have much higher purities than basal-like samples. But the p-value is the smallest compared with K-means and cNMF (p-values are 1.02×10^{-51} and 5.04×10^{-98}, respectively).

Even though the number of tumor samples with ABSOLUTE purities is much less than that with InfiniumPurify purities (264 versus 746 samples), we still detected significant discrepancies in purities among different subtypes in Figure 2.1D and Figure 2.1E. For example, the ABSOLUTE purities of the third cluster are still way below other groups (Figure 2.1E), and the PAM50 subtypes still have the smallest purity differences compared with K-means, cNMF (Figure 2.1F). Overall, we consistently observe significant purity differences among different subtypes in typically used K-means and cNMF methods by purities from both InfiniumPurify and ABSOLUTE. These results demonstrate that tumor purities will bias the clustering results if not correctly accounted for, and a clustering method with consideration of purity is therefore needed.

2.3.2 Simulation

To evaluate InfiniumClust in cancer sample clustering, we conducted comprehensive simulation studies to compare the performance of our method with other available methods. In the simulation presented below, we used data from BRCA as template. Since methylation level ranges from 0 to 1, it is a natural choice to simulate methylation level from beta distribution. In detail, data are generated using the following scheme:

1. *Pure normal samples*: For sample j at CpG i, we generated its methylation level as $X_{ij} \sim Beta(\alpha_{i0}, \beta_{i0})$. Here α_{i0} and β_{i0} are the method of moments estimates (MME) from the beta values of a total of 96 normal BRCA samples.

2. *Pure tumor samples*: For sample j at CpG i, let $Y_{ij} \sim Beta(\alpha_{ik}, \beta_{ik})$ in subtype k, where α_{ik} and β_{ik} are estimated from the following procedure. We first convert α_{i0} and β_{i0} to mean and dispersion (denoted by m_i and d_i, respectively). The dispersion parameter in beta distribution represents the variance that is independent of mean. Because the pure tumor samples are more heterogeneous than normal, we multiply the above dispersions by 2 as the tumor dispersions (Neve et al., 2006; Zheng et al., 2014). For the mean of pure tumor samples, we randomly selected K normal samples from BRCA and used their beta values as the mean. With means and dispersions, we converted them back to α and β by the formulas $\alpha_{ik} = m_{ik}(\frac{1}{d_i} - 1)$ and $\beta_{ik} = (1 - m_{ik})(\frac{1}{d_i} - 1)$, and then generated beta values for K subtypes.

3. *Observed samples*: We generated tumor purity values λ_j, $j = 1, 2, \ldots, N$ uniformly from [0.05, 0.95]. Substituting X_{ij} (from (1)), Y_{ij} (from(2)) and λ_j into the formula $Y'_{ij} = \lambda_j Y_{ij} + (1 - \lambda_j) X_{ij}$. Then Y'_{ij} is the observed methylation level, which is a mixture of methylation level from pure cancer and normal samples.

We applied InfiniumClust and K-means to the simulated data, and compared their clustering performances. Because the group assignment of each sample is known, the accuracy is defined as the percentage of correctly clustered samples. Since the group indicators from all clustering methods are dummy variables, we use the following procedure to match the clustering results with the truth. Assuming there are N clusters, we first tabulate the group assignments for all samples from the truth and clustering results into a $N \times N$ table. Entry (**i**,**j**) in the table represents the number of samples belonging to the i^{th} group in truth, and predicted as the j^{th} group from clustering method. We then shuffle the rows and columns of the table, so that the sum of the diagonal elements achieves the maximum. Finally, the sum of the diagonal elements over the total number of samples is defined as the accuracy. For these simulations the data are generated from a 3-subtype mixture with mixing proportions

ratio 0.2:0.5:0.3, and all simulations are repeated for 20 times.

First, we evaluated the effect of CpG selection on the accuracy of InfiniumClust. Instead of selecting CpG sites with the top 1000 largest variances, we randomly selected 1000 CpG sites to run InfiniumClust. Results show that the accuracy could be significantly worse compared with choosing 1000 CpG sites with the highest variances (Figure 2.2A). This demonstrates that probes with larger variances due to different normal cell contaminations are more informative for clustering.

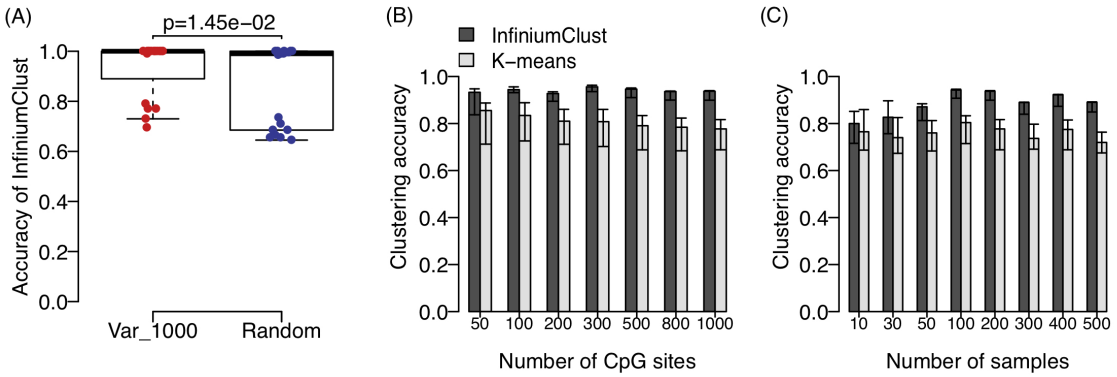


Figure 2.2: (**A**) Predicting accuracy on InfiniumClust by selecting 1000 CpG sites with the largest variance and 1000 randomly selected CpG sites. (**B**, **C**)Predicting accuracy in different number of CpG sites and sample sizes on InfiniumClust and K-means.

Based on the above analysis, we selected CpG sites with largest variances in tumor tissues and used their data as the input for clustering in the following analysis. First, we compared InfiniumClust and K-means at different numbers of selected CpG sites from 50, 100, 200, 300, 500, 800 and 1000, respectively. As shown in Figure 2.2B, overall the accuracies of InfiniumClust are much higher than those by K-means (around 0.94 versus 0.81) regardless of the number of CpG sites used. More importantly, InfiniumClust is robust against the numbers of CpG sites, but the accuracy of K-means gradually decreases with more CpG sites used.

We also tested the performance of InfiniumClust with varied sample sizes. If we have only 10 samples, the accuracies by InfiniumClust and K-means are almost the

same (0.8). But the accuracy of InfiniumClust increases to 0.9 when sample size increases from 30 to 500, while the accuracy of K-means roughly remains the same (0.76) (Figure 2.2C).

Compared with traditional clustering methods, InfiniumClust takes purity into consideration. So we further tested the effect of purity on the algorithm from the following aspects. First, we divided tumor samples into two groups (high purity and low purity) using the median of purities among samples as cutoff. Figure 2.3A shows that tumor samples with higher purities have higher chance to be clustered correctly. It is expected because samples with lower purities tend to be incorrectly clustered due to their higher normal cell contamination. They are likely to be clustered together since they are more similar to normal samples. Second, we examined the influence of accuracy of purity estimation on the algorithm. We randomly shuffled the purities of all tumor samples, and used them as input to implement InfiniumClust. As shown in Figure 2.3B, clustering accuracies significantly decreased to around 0.75, which is similar to the performance of K-means. This is not surprising because InfiniumClust uses shuffled purities, which is equivalent to ignoring tumor purities by K-means. Next, to test the robustness of our model against purity estimation, we added different levels of random noise with the Gaussian distribution to the purities of tumor samples. To be specific, tumor purities are added by a random noise of the Gaussian distribution with mean 0 and standard deviations from 0 to 1, step by 0.02. Note that the output purities could possibly be ranged out of [0, 1] after adding noise, so we set them as 0.01 if lower than 0 and 0.99 if larger than 1. As expected, the accuracy of InfiniumClust decreases with the increase of standard deviation, but still over 0.8 (K-means 0.75, Figure 2.3C). This indicates that InfiniumClust still has better performance than K-means even if estimated tumor purities are biased.

We then evaluated the performance of InfiniumClust under different subtype proportions. We selected 200 samples, 1000 CpG sites with the largest variance to

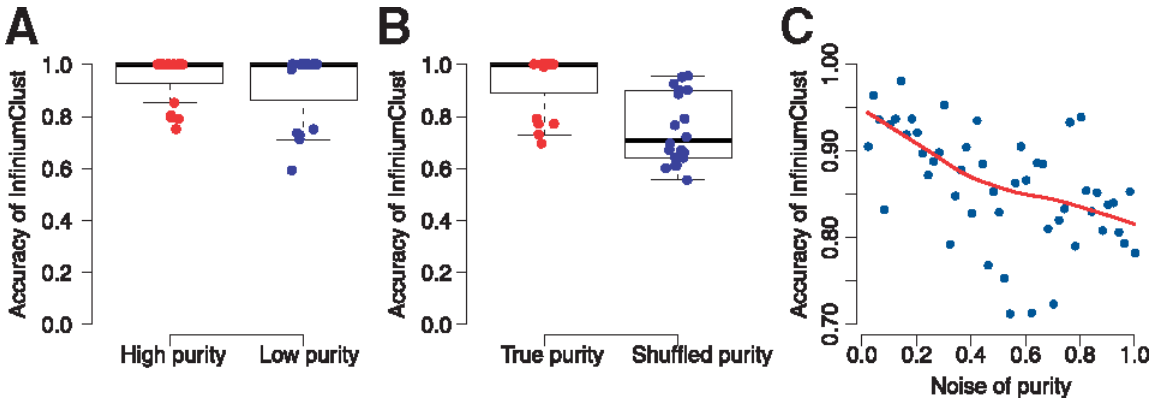


Figure 2.3: (**A**) Predicting accuracy of InfiniumClust on samples with high purity and samples with low purity. (**B**) Comparing the accuracy of InfiniumClust for the samples with the precise purity and samples with the shuffled purity. (**C**) Scatter plot of the noise of purity versus the accuracy of InfiniumClust, polynomial regression curve is displayed.

run InfiniumClust and K-means, with different proportion ratios of three subtypes. As shown in Figure 2.4A, InfiniumClust always performs well, on average about 0.9 for most scenarios. Even if proportions of subtypes are very unbalanced, e.g. 0.05:0.05:0.9, InfiniumClust still has good accuracy ($>$ .8). We also examined several proportions at four and five subtypes. InfiniumClust achieves accuracies as high as 0.9, whereas the accuracies of K-means are only around 0.7 (Figure 2.4B and C).

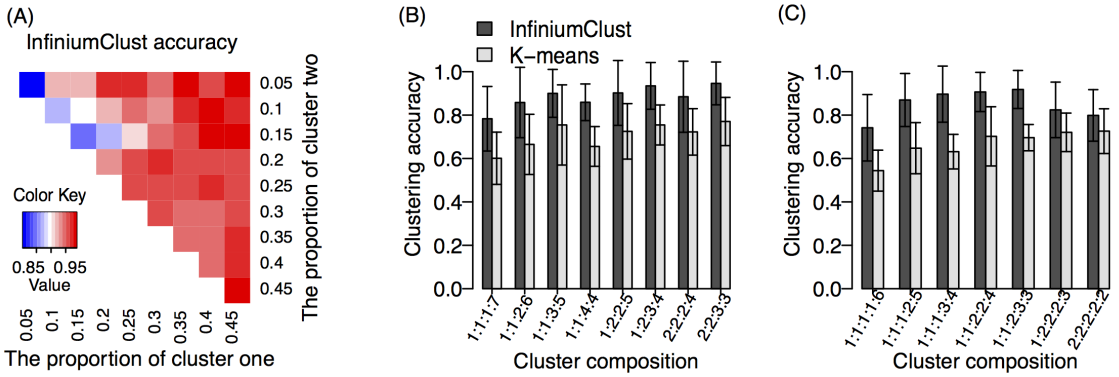


Figure 2.4: (**A**) Heatmap of different proportion of subtypes of $K = 3$ on InfiniumClust, where row indexes the proportion of the first subtype, column indexes the proportion of the second subtype. (**B**, **C**) Barplot of predicting accuracy in different proportion ratios of subtypes of $K = 4$ and $K = 5$.

Another possible procedure to account for the purity effect is to estimate pure can-

cer methylome and perform clustering (such as K-means) on purified data directly. Given methylation levels of normal-adjacent sample and estimated purity, the pure cancer methylome can be inferred by simply subtracting out the normal signal from the tumor data based on a linear equation with consideration of purity. One caveat in such approach is that there are much fewer normal samples in TCGA compared with tumor samples (only 674 normal-adjacent samples from 12 cancer types), i.e. only a small proportion of the tumor samples has corresponding normal controls. In this case, one can compute the average normal methylome and subtract that to obtain purified cancer methylome. We conducted simulations to evaluate the performance of this approach (termed as puKmeans hereafter). Overall, the accuracies of puKmeans are higher than K-means, but lower than InfiniumClust (Appendix 1, Fig. A4). This is expected, because puKmeans takes the point estimates of pure cancer methylome as inputs but ignores the variance, thus the information in data are not used in an optimal way. Furthermore, we found that puKmeans is sensitive to the number of CpG sites used in the clustering. In our simulation, using 100 CpG sites provides the best results, and using more CpG sites leads to lower accuracy. On the contrary, InfiniumClust is very robust against the number of CpG sites. These results demonstrate the advantage of the proposed model-based clustering method over a simplified approach to cluster on purified data.

Under all simulation scenarios, InfiniumClust achieves better accuracies. Moreover, InfiniumClust is robust against CpG site selections, sample sizes, biases in purity estimation and proportions of clusters. These results demonstrate the advantages of InfiniumClust in clustering tumor samples while considering tumor purity, as well as a well-constructed model.

2.3.3 Application of InfiniumClust to TCGA data

With the success of InfiniumClust on simulation data, we next tested InfiniumClust on real tumor samples. We analyzed all samples with both NMF clustering results and 450k array data (23 cancer types from TCGA). Data from 1000 CpG sites with the largest variance among tumor samples were used for InfiniumClust and K-means. The numbers of clusters of NMF used on these cancer types was the same number of clusters for InfiniumClust and K-means.

First, we explored purity distributions between correctly clustered and incorrectly clustered samples. Since the true clusters are unknown in real data, we used the consensus samples of the three methods (NMF, K-means and InfiniumClust) in each cancer type as a proxy for truth under the assumption that samples clustered into the same group by different methods tend to form a true cluster. The consensus sample is defined as follows. First consider two methods A and B that both cluster samples into N groups. The group indices from the clustering methods are dummy indicator variables that bear no biological meaning. To look at the agreements of the two methods, we first fix the group indices for method A, and then shuffle the group indices of method B to get the maximum overlap from all groups between the two methods. The overlapped samples in each cluster of the two methods are termed as consensus samples, while the rest are non-consensus samples. Similarly, for consensus samples among three methods, we first get consensus samples between any two methods, then compare the consensus samples with the results from the third method using the same procedure. The consensus samples obtained from this comparison are defined as consensus samples among three methods. All others are deemed non-consensus samples. For all 23 cancer types in TCGA, we found that 50% of the samples belong to the consensus group on the average. In general, we observed significant differences in purity levels between the two groups in most cancer types (Figure 2.5A, results for other cancer types are shown in Appendix 1, Fig. A5). Samples with higher purity

tend to be in the consensus group. The result is consistent with the simulation result that samples with higher purities tend to be clustered correctly.

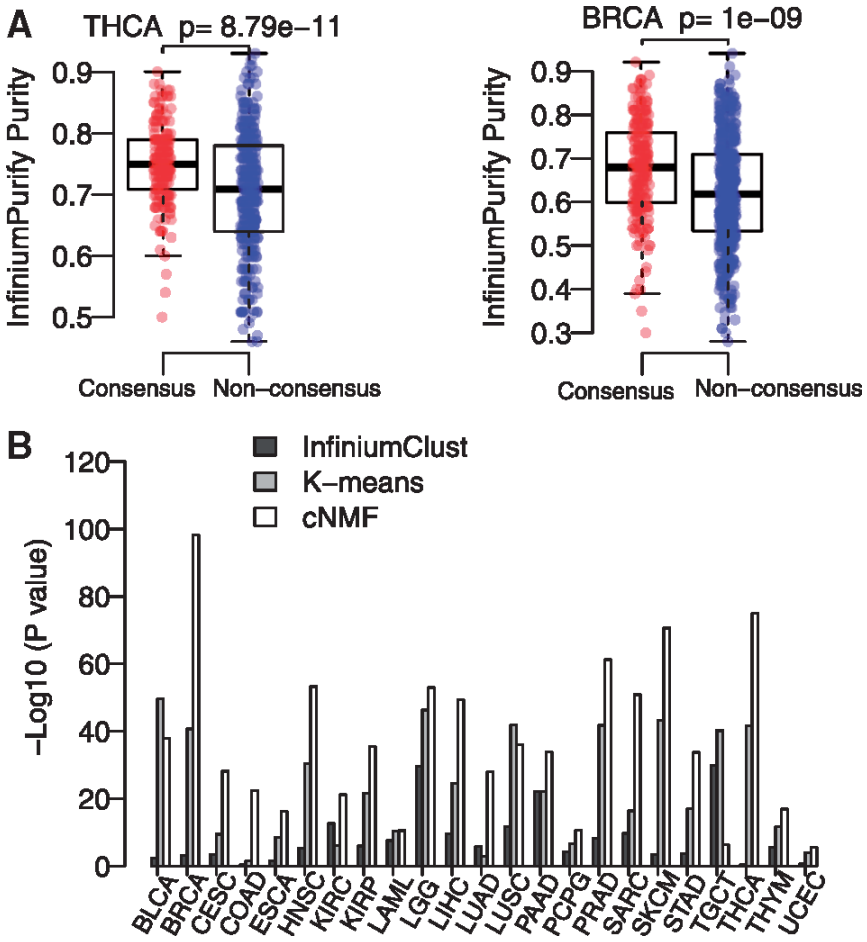


Figure 2.5: (**A**) The distribution of InfiniumPurify Purity in consensus samples versus non-consensus samples in THCA and BRCA. (**B**) Testing purity differences among clusters from three methods for 23 cancer types. P-values are from linear regression and F-test.

Next, we examined the purity difference among different clustering results in these three methods (Appendix 1, Fig. A5). We computed the p-values by testing purity difference among clusters from different methods. As shown in Figure 2.5B after $-log_{10}$ transformation, InfiniumClust gives much less significant p-values compared with the other two methods. These results indicate that InfiniumClust is less affected by the purities due to the inclusion of purities in clustering. They also show that compared with K-means, purity has stronger influences on NMF. Overall, these results

emphasize the risk of ignoring tumor purity when applying unsupervised clustering, and demonstrate good performances from InfiniumClust.

We also checked the overlap of clusters in these three methods. Appendix 1, Fig. A2 shows the pairwise overlaps from the three methods. The average overlap between clusters of InfiniumClust and K-means is 0.72 for 23 cancer types. In contrast, the overlaps between the clustering results from NMF and other two methods are much lower: the average overlap of NMF and InfiniumClust is 0.54. These results are expected because algorithm-wise, InfiniumClust and K-means are very similar (K-means and the normal mixture model perform similarly when data are normally distributed). The only difference is the consideration of purities in InfiniumClust. In contrast, NMF is based on a different method, thus it tends to produce different results.

To test the robustness of our method, we also used ABSOLUTE purities to repeat the above analysis. We selected eight cancer types with ABSOLUTE purities to compute purity distributions in consensus and non-consensus samples, the purity difference among different clusters, and the overlap of clusters in these three methods, respectively. The results are consistent with those using InfiniumPurify purities. In particular, we also observed significant differences in purity levels between the consensus and non-consensus in most cancer types; InfiniumClust gives much less significant results compared with the other two methods (K-means and NMF); especially. In BLCA and LUAD, the clusters overlap between InfiniumClust and K-means is even over 0.9 (Appendix 1, Fig. A3). These results further prove that InfiniumClust is less influenced by tumor purities. Therefore, we believe InfiniumClust provides robust results in real data.

We further applied puKmeans in TCGA tumor data (12 cancer types, 5185 tumor samples and 674 normal samples). As a comparison, we also conducted InfiniumClust and K-means clustering in these cancer-normal mixtures. Since it is difficult to eval-

uate the performances without a gold standard, we only performed some exploratory analyses. As shown in Appendix 1, Table A1, the average overlap between clusters of InfiniumClust and puKmeans is slightly higher than that between InfiniumClust and K-means for 12 cancer types (0.712 versus 0.696).

2.4 Discussion

In this study, we systematically investigated the impact of tumor purity as a confounding factor in unsupervised clustering of tumor samples, and proposed a statistical model to adjust the effect of purity in tumor sample clustering. We first found that under traditional K-means and NMF approaches, tumor purities bias the clustering results (samples with similar purities are likely to cluster together), and that tumor samples with low purities tend to be misclassified. We designed a model-based statistical method InfiniumClust for subtype classification based on DNA methylation data. In InfiniumClust, methylation levels from tumor samples at each CpG site are modeled as mixture of normal distributions. Parameter estimation and sample clustering is conducted by an EM type algorithm. Based on simulation, InfiniumClust achieved more robust and accurate results compared with K-means algorithm. When applying to real TCGA tumor samples, InfiniumClust obtained the least biased clusters compared to K-means and the well-established NMF method. These results reinforce our claim that purity difference may confound genomic analyses if not correctly accounted for. To the best of our knowledge, InfiniumClust is the first method for unsupervised clustering for cancer subtypes adjusted for tumor purity.

In our model, we assume a Gaussian distribution for the transformed methylation level in each CpG site: data from normal samples follow a single Gaussian distribution and data from tumor samples follow a mixture of Gaussian distributions. We validate the assumptions in real data, and demonstrate that they approximately hold even though there is mild violation (Appendix 1). However, according to our simulation

results, the clustering algorithm will still perform well even if some CpG sites do not satisfy the normality assumption.

The current version of InfiniumClust is specifically designed for Infinium 450k methylation array, which is the most widely used platform for DNA methylation. It is conceivable that the same principle and methods can be applied to data from other platforms, or even other types of genomics data. For example, gene expression or copy number variation perhaps play more direct roles in tumorigenesis, and their data from tumor samples are influenced by purities as well. We will pay further attention to model gene expression and copy number data with consideration of purities. It could even be possible to integrate all these data into a unified model to better improve the clustering accuracy.

Chapter 3

Disease Prediction by Cell-Free DNA Methylation

3.1 Introduction

Prognosis and diagnosis play vital roles in the prevention and treatment of diseases. Traditionally, various types of surgical biopsies such as bone marrow or needle biopsies are performed in clinical setting, especially for cancer diagnosis (Sgouros, 1993). However, due to the invasive nature of the procedure and the potential sampling bias of tumor biopsy, surgical biopsy is often not a preferred choice. In recent years, disease diagnosis based on molecular biomarkers in specimen, including blood, urine, and cerebral spinal fluid, has gain tremendous attention. For example, the practice to look for traces of cancer DNA by interrogating biomarkers on plasma-isolated cell-free DNA (cfDNA) or circulating tumor DNA (ctDNA) is known as liquid biopsy (Crowley et al., 2013). As a safer, cheaper, and quicker alternative to surgical biopsy, the liquid biopsy has great potential in clinical practice. cfDNA is a mixture of DNA fragments from different cell types. In cancer patients, the cfDNA includes some circulating tumor DNA (ctDNA), which are DNA fragments released from cancer cells (Schwarzenbach et al., 2011). As a hallmark of cancer, the ctDNA carries tumor-specific genetic variants such as copy number variation and point mutations. After capturing and sequencing the cfDNA, ctDNA can be distinguished from normal cfDNA by tumor-specific genetic variants. Presence of non-trivial amount of ctDNA is an indication of cancer.

Recently, researchers start to explore the cfDNA epigenetics information such as DNA methylation or nucleosome position to look for biomarkers for diagnosis (Kang et al., 2017; Xu et al., 2017; Song et al., 2017; Li, Zhang, Lu, You, Song, Luo, Zhang, Nie, Zheng, Xu et al., 2017). In analogous to the liquid biopsy, these approaches try to define abnormality based on the epigenetic profiles, and then construct model for disease prediction.

In this work, we focus on cfDNA methylation and explore how they can be used

for disease prediction. We systematically review the existing publications and investigate the statistical methods for cfDNA methylation study. We discuss two important aspects: marker selection and prediction model construction, under different scenarios. We conduct extensive simulation and real data analysis, and provide some recommendations for data analysis strategies based on the results.

3.2 The Cause of Alteration of cfDNA Methylation in Disease

DNA methylation is known to be highly tissue-specific (Schultz et al., 2015), which is an important basis for cfDNA methylation data analysis. Even though different tissues share exactly the same DNA sequence, the differences in their methylomes allow one to trace the tissue of origins of cfDNA, and subsequently use those information for disease prediction.

Considering cfDNA as a mixture of DNA segments from different tissues, the differences in cfDNA methylation between patients and healthy people could be from two sources. The first one is the alteration in one particular tissue type in disease, for example, the methylation level changes in certain cell types between breast carcinoma versus normal (Bloushtain-Qimron et al., 2008). The second is the change in mixing proportions in the composition of cfDNA, for example, hepatocellular carcinoma (HCC) patients have an increasing proportion of cfDNA fragments originates from apoptotic liver cells (Sun et al., 2015). It is important to note that both changes are usually not reflected in the methylation profiles in the blood sample, thus one cannot construct disease prediction model from blood data but have to rely on cfDNA.

It is well known that DNA methylation is highly tissue-specific (Kang et al., 2017; Avraham et al., 2014; Ghosh et al., 2010; Varley et al., 2013). Thus, both of these changes will lead to the marginal cfDNA methylation changes between cases and

controls. For disease prediction, the most straightforward idea is to detect differentially methylated loci (DML) or regions (DMRs) between cases and controls from the cfDNA methylation data, and use the methylation levels in those regions as predictors for diagnosis (Song et al., 2017; Li, Zhang, Lu, You, Song, Luo, Zhang, Nie, Zheng, Xu et al., 2017). Another family of approaches is to first trace the tissue-of-origin of cfDNA and estimate the mixing proportions, then construct a model to predict disease status based on the estimated proportions. This type of methods takes advantage of the tissue-specificity of epigenetic profiles such as DNA methylation or nucleosome position, and use signal deconvolution methods for proportion estimation (Kang et al., 2017; Sun et al., 2015; Ulz et al., 2016). We conduct detailed simulation studies to compare these two types of approaches under different scenarios (detailed in later section).

3.3 Existing Works

Table 3.1 lists the existing publications for using cfDNA epigenetic profiles in diseases diagnosis (Kang et al., 2017; Xu et al., 2017; Song et al., 2017; Li, Zhang, Lu, You, Song, Luo, Zhang, Nie, Zheng, Xu et al., 2017; Schultz et al., 2015; Sun et al., 2015; Ulz et al., 2016; Jensen et al., 2015; Lehmann-Werman et al., 2016; Tanić and Beck, 2017; Warton and Samimi, 2015; Lokk et al., 2014; Hatt et al., 2015; Guo et al., 2017; Snyder et al., 2016; Legendre et al., 2015). As discussed before, the prediction model construction can be roughly categorized into two classes: (1) using the marginal cfDNA epigenetic profile as predictors, or (2) using the mixing proportions as predictors. The first class includes (Xu et al., 2017; Song et al., 2017; Li, Zhang, Lu, You, Song, Luo, Zhang, Nie, Zheng, Xu et al., 2017; Schultz et al., 2015; Jensen et al., 2015; Lehmann-Werman et al., 2016; Lokk et al., 2014; Hatt et al., 2015; Legendre et al., 2015). For example, Xu et al. (2017) used 10 cfDNA methylation markers for diagnosis of hepatocellular carcinoma using logistic regression. Another

example is using Random Forest (RF) on a set of regions to classify cancer types (Song et al., 2017). The second class includes (Kang et al., 2017; Sun et al., 2015; Guo et al., 2017). For example, Kang et al. (2017) modeled proportion of tumor-derived cfDNA and used a probabilistic model to predict tumor burden and tumor type. Sun et al. (2015) used an external thousands-marker reference panel to solve for tissue proportions in HCC patients and healthy controls. The estimated proportions can potentially be used for disease diagnosis.

In addition to DNA methylation, there is attempt to use other cfDNA epigenetic information such as nucleosome position for disease prediction. Snyder et al. (2016) found that during apoptosis, genomic DNA protected by nucleosomes will be released to bloodstream, and the unprotected naked DNA will be degraded. The tissue-specific nucleosome positioning causes different fragmentation pattern in cfDNA, thus allows one to trace the tissue of origin which could be helpful for conducting disease prediction. However, due to the limited number of researches and data available, using cfDNA nucleosome position to predict disease will not be included in this review.

3.4 Methods

3.4.1 Marker Selection

Marker selection is the first step in disease prediction model construction. In cfDNA methylation studies, both the Whole Genome Bisulfite Sequencing (WGBS) and the human 450k/27k methylation array profile large number of CpG sites. A majority of these CpG sites are either irrelevant, noisy, or redundant for distinguishing the underlying disease status. Including all CpG sites as features in the model will have harmful impacts on traditional machine learning algorithms such as support vector machine (Li and Yu, 2008). Therefore, marker selection is a very important step to alleviate problem caused by bad markers. Typically, researchers select tens to thousands of

Disease	Epigenetic profile used	Data type	Sample size	Prediction method	Publication
Lung cancer, hepatocellular carcinoma, (HCC), pancreatic cancer, glioblastoma (GBM), gastric cancer, colorectal cancer, breast cancer patients	5hmC	hMe-Seal	49	Random Forest Mclust	Song et al. (2017)
Colorectal cancer, gastric cancer, pancreatic cancer, liver cancer, thyroid cancer	5hmC	hMe-Seal	350	Logistic regression	Li, Zhang, Lu, You, Song, Luo, Zhang, Nie, Zheng, Xu et al. (2017)
Hepatocellular carcinoma (HCC)	5mC	BS-seq	1933	Logistic regression	Xu et al. (2017)
Pregnant/non-pregnant plasma	5mC	BS-seq	27	NA	Jensen et al. (2015)
General cancer	5mC	Methylation 450k microarray/ BS-seq	87	Probalistic model for tumor burden	Kang et al. (2017)
Diabetes, multiple sclerosis, traumatic or ischemic brain damage, pancreatic cancer or pancreatitis	5mC	methylation 450k microarray	218	NA	Lehmann-Werman et al. (2016)
General disease	5mC	Methylation 450k microarray/ BS-seq	NA	NA	Tanić and Beck (2017)
Colorectal, breast, lung, pancreatic and ovarian cancers	5mC	Methylation 450k microarray/ BS-seq	NA	NA	Warton and Samimi (2015)
General disease-related pathogenic mechanisms	5mC	Methylation 450k microarray	N=4 (17 tissues)	NA	Lokk et al. (2014)
Colon, prostate, breast, lung cancer	nucleosome positioning	DNA sequencing	179	Coverage depth	Ulz et al. (2016)
Pregnancies/non-pregnancies	5mC	Methylation 450k microarray	22	NA	Hatt et al. (2015)
Lung, colorectal cancer	5mC	BS-seq	59	NA	Guo et al. (2017)
Tissue-specific methylation	5mC	MethylC-seq	N=4 (18 tissues)	NA	Schultz et al. (2015)
General cell types contribution	nucleosome positioning	DNA sequencing	60	Coverage depth	Snyder et al. (2016)
Prenatal, cancer, and transplantation assessments	5mC	BS-seq	83	Quadratic Programming	Sun et al. (2015)
Metastatic breast cancer	5mC	BS-seq	120	NA	Legendre et al. (2015)

Table 3.1: List of publications of using cfDNA epigenetics information to infer disease. The epigenetics information utilized are either DNA methylation (5mC), DNA hydroxymethylation (5hmC) or nucleosome position.

markers based on data from all CpGs. These makers could be CpG sites, CpG clusters (Kang et al., 2017), or fixed-size genomic bins (Sun et al., 2015). The selection criteria are typically based on the differentiating power of the markers, that is, selecting features showing significant differences among different tissues (Sun et al., 2015) or wide between-group methylation ranges (Kang et al., 2017). All existing publications utilize their own approach for selecting CpG sites. These approaches generally take following three aspects into consideration. First, some studies use differentially methylated loci (DML) as predictive markers. For example, Xu et al. (2017) used 10 highly-selective CpGs as the informative markers in diagnosis of hepatocellular carcinoma. It is a direct and reasonable approach because selected markers are discriminative for disease status. Second, some studies use regions instead of single CpG markers as features for prediction. For example, Song et al. (2017) used 5hmC signal within the gene body, Kang et al. (2017) used 100 bp up- and downstream CpG sites as regions, and Lehmann-Werman et al. (2016) used several adjacent CpG sites as the basic unit for features. The underlying assumption of choosing region instead of single CpG site as the feature is that adjacent CpG sites have similar methylation level, and pooling information from nearby CpG sites together can stabilize and enhance signals. Third, some studies borrow biological information from external data to select markers. For example, Xu et al. (2017) used solid tumor samples from The Cancer Genome Atlas (TCGA) to conduct preliminary marker selection. The intuition behind such approach is that features differ significantly between solid tumor and normal tissue would also be likely to demonstrate detectable methylation differences in the cfDNA of the same disease.

In order to select informative and discriminative markers for disease prediction, we suggest detecting DML from training data first. The criteria for selecting markers from this step can be relatively loose, in order to retain relatively large number of markers. Next, when external biological information such as markers from tissue-

specific methylation are available, one can use these locations to filter the markers from the previous step. Furthermore, one should consider pooling nearby CpG sites together to create regions if possible, instead of using single CpG site. This will help boost and stabilize the methylation signals. Finally, to determine the number of markers allowed in the final statistical model, one needs to conduct cross-validation (CV) to select the optimal number that minimize the prediction error. After all these steps above, data of the selected markers for all samples can be used either directly as features for disease prediction or for signal deconvolution (more details in subsection 3.4.3).

3.4.2 Data Generative Model

Once the markers are selected, the next step is to build statistical model to predict the disease status. The training data include the cfDNA methylation profiles (denoted as $\mathbf{Y}$, could be from the WGBS or the 450k/27k methylation array) for the selected markers, and disease status (denoted as $\mathbf{Z}$) for $\mathbf{N}$ subjects including N_0 patients and N_1 healthy people ($\mathbf{N} = N_0 + N_1$). $\mathbf{Y}$ is a matrix of $\mathbf{M}$ by $\mathbf{N}$, where $\mathbf{M}$ is the number of pre-selected biomarkers (CpG sites or regions). $\mathbf{Z}$ is a binary vector of length N, (1 for case and 0 for control). The goal of the problem is to use cfDNA methylation data ($\mathbf{Y}$) to predict disease status ($\mathbf{Z}$).

Suppose there are T tissues releasing DNA fragments into the cfDNA pool in plasma. Denote the methylation profiles for the M biomarkers in these T tissues as matrix $\mathbf{R}$. $\mathbf{R}$ is of dimension M by T, where each column represents the methylation levels of the M biomarkers from one tissue. It is important to note that due to biological variation, the $\mathbf{R}$ matrices are not exactly the same from different people. However, the marker selection step guarantees that the variation among individual for the same tissue are significantly lower than the difference among different tissues. Moreover, since there could be differential methylation in certain tissue types between

cases and controls, $\mathbf{R}$ in cases can be potentially different from the $\mathbf{R}$ in controls. In some situation, $\mathbf{R}$ can be obtained from methylomes of specific tissues or purified cell types (Sun et al., 2015). $\mathbf{R}$ could also be unknown or unavailable if we do not have external information about those biomarkers or the tissues of interests.

As described earlier, cfDNA is a pool of mixing DNA fragments from each of the T tissues. For each individual, the tissue proportion is a vector of length T. Each element in the vector is a number between 0 and 1, and all elements from the vector will sum up to 1. For these N individuals, the tissue proportions are represented as a T by N matrix $\mathbf{\Pi}$, where π_{ij} is the tissue proportion of the i^{th} tissue in the j^{th} individual, and $i = 1 \dots T$; $j = 1 \dots N$. It has the restriction of $\sum_{i=1}^{T} \pi_{ij} = 1$ and each $\pi_{ij} \in [0, 1]$.

Following the above notations, the expected values of the cfDNA methylation ($\mathbf{Y}$) is a mixture of the tissue specific methylation ($\mathbf{R}$):

$$E(\mathbf{Y}) = \mathbf{R\Pi}$$

We use the expectation notation E(.) here because the observed cfDNA methylation data $\mathbf{Y}$ contains random noises. For modeling and computational convenience, it is commonly assumed the random errors following normal distribution with mean 0. From this model, it is clear that the differences in either $\mathbf{R}$ or $\mathbf{\Pi}$ will cause E($\mathbf{Y}$) to differ between two groups. In the next section, we will discuss the possible statistical methods for using cfDNA methylation data $\mathbf{Y}$ to predict the disease.

3.4.3 Disease Prediction Approach

With training data, several methods can be applied for disease status prediction:

Directly using marker methylation to predict

As the most straightforward approach, one can directly use the observed cfDNA methylation ($\mathbf{Y}$) to predict disease status $\mathbf{Z}$, using an off-the-shelf machine learning (Xu et al., 2017; Song et al., 2017; Li, Zhang, Lu, You, Song, Luo, Zhang, Nie, Zheng, Xu et al., 2017) or model-based approach (Kang et al., 2017). The trained model can be evaluated using test data, and eventually used as a panel to diagnose new patients. This approach is easy and intuitive, and widely used in many existing publications (Kang et al., 2017; Xu et al., 2017; Song et al., 2017; Li, Zhang, Lu, You, Song, Luo, Zhang, Nie, Zheng, Xu et al., 2017). Since differences in either $\mathbf{R}$ or $\mathbf{\Pi}$ will cause E($\mathbf{Y}$) to differ between case and control, one does not need to know exactly the source of changes as long as $\mathbf{Y}$ can predict $\mathbf{Z}$.

Prediction based on tissue mixing proportions

To take a step further than using marker data directly, there are some researches to first estimate the mixing proportion $\mathbf{\Pi}$, and then using $\mathbf{\Pi}$ as predictor for diagnosis. The underlying assumption is that the disease is associated with the change of mixing proportions (which is related to cell death rates). The proportions estimation can be viewed as a dimension reduction step, which can potentially improve the signal to noise ratio in the data and lead to better prediction accuracy. An added benefit of this approach is that the results are more interpretable: disease is associated with the proportion change of certain cell type, which could be related to the cell death rate for that tissue.

The estimation of the mixing proportions can be achieved by using following two different procedures.

1. Reference-based method

   When external reference panel R is available, the estimation can be done by

regressing the mixed signal Y to purified tissue reference R. Since the regression coefficients are not totally free parameters and have to satisfy some constraints (between 0 and 1, sum up to 1 in each individual), the problem is a constrained linear regression in the following form:

$$\begin{cases} E(\mathbf{Y}) = \mathbf{R\Pi} \\ \sum_{i=1}^{T} \pi_{ij} = 1 \\ 0 \leq \pi_{ij} \leq 1 \end{cases}$$

This can be converted into an optimization problem to minimize the residual sum of squares. The optimization problem has a quadratic loss function and linear constraints, thus can be solved by *Quadratic Programming* (QP) algorithm.

With estimated proportions, we can train a Support Vector Machine (SVM) to predict $\mathbf{Z}$ from $\hat{\mathbf{\Pi}}$ from training data. When a new patient coming in, one can use the reference panel $\mathbf{R}$ to solve for new individuals tissue proportion $\hat{\pi}$ using QP, and then apply the trained SVM on $\hat{\pi}$ to predict disease status.

This approach is easy, intuitive, and computationally efficient. The only shortcoming is the requirement of $\mathbf{R}$. One can look for $\mathbf{R}$ in public data, but has to assume that it is not significantly different from the reference methylome of the population under study, which could be a strong assumption.

2. Reference-free method

When external reference panel $\mathbf{R}$ is unavailable, one can use *Non-negative matrix factorization* (NMF) algorithm to jointly solve for $\mathbf{R}$ and $\mathbf{\Pi}$. Briefly speaking, NMF is an algorithm that factorizes a matrix, say $\mathbf{V}$, into two matrices $\mathbf{W}$ and $\mathbf{H}$, such that:

$$\mathbf{V} = \mathbf{WH}$$

where all three matrices contain non-negative elements. Because $\mathbf{W}$ and $\mathbf{H}$ are both unknown, the factorization is solved by numerical approximation methods (Onuchic et al., 2016). To be specific, the estimator of $\mathbf{W}$ and $\mathbf{H}$ follows:

$$argmin_{\mathbf{W},\mathbf{H}}||\mathbf{V} - \mathbf{WH}||^2$$

where $0 \leq \mathbf{W} \leq 1$, $0 \leq \mathbf{H} \leq 1$, and $\sum_i \mathbf{H}_{ij} = 1$ for any j. After initialization of $\mathbf{W}$ and $\mathbf{H}$, a procedure is taken to estimate $\mathbf{W}$ given fixed $\mathbf{H}$, and then estimate $\mathbf{H}$ given fixed $\mathbf{W}$, iteratively, until it converges. NMF was traditionally used on chemometrics, signal processing, and image processing (Gao et al., 2005; Cichocki et al., 2006). Recently, NMF is gaining popularity among computation biological research community, especially in analyzing data from highly heterogeneous samples (Houseman et al., 2012, 2016; Cardenas et al., 2016). One major reason for this popularity is that the factorized matrices are in reduced dimensions and have better biological interpretation.

In estimating mixing proportions from cfDNA methylation, we factorize methylation matrix $\mathbf{Y}$ into two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$, while constraining each cell in matrix $\mathbf{H}$ takes value within [0, 1] and each column in matrix $\mathbf{H}$ sum up to 1. To be noticed, the original version of NMF only requires $\mathbf{W}$ and $\mathbf{H}$ to be non-negative, but does not have those added constrains. The algorithm was specifically customized by adding these new constraints to solve for $\mathbf{W}$ and $\mathbf{H}$ for DNA methylation study (Houseman et al., 2016; Li, Xiao, Shi, Yang, Wang, Wang, Marcia and Lu, 2017; Lutsik et al., 2017). Then $\mathbf{W}$ can be interpreted as a sudo-reference matrix comparable to the external reference matrix $\mathbf{R}$, and $\mathbf{H}$ can be interpreted as a sudo-tissue-proportion matrix similar to the tissue proportion matrix $\mathbf{\Pi}$. Matrix $\mathbf{H}$ can then be used for disease status prediction, resembling using QP-solved proportion matrix $\mathbf{\Pi}$.

NMF provides a flexible way to solve for tissue proportions when the external tissue reference information is unavailable. Under the context of cfDNA methylation study, assume we have training data for N individuals (with known disease status), and new patients (with cfDNA methylation data but disease status is unknown). Reference-free NMF-based approach is the following. First, we apply NMF on training data $\mathbf{Y}$ to factorize it into two matrices $\mathbf{W}$ and $\mathbf{H}$. Since the columns from $\mathbf{H}$ represent individuals with known disease status $\mathbf{Z}$, we train a *Support Vector Machine* (SVM) using $\mathbf{H}$ to predict $\mathbf{Z}$. Then we regress $\mathbf{W}$ on new patients data $\mathbf{Y^b}$ in regression to get testing datas tissue proportions $\mathbf{H^b}$. Finally, we apply the trained SVM on $\mathbf{H^b}$ to predict disease status for new patients.

To summarize the methods described above, a schematic illustration of plasma cfDNA methylation mixing procedure and deconvolution methods for disease detection and monitoring is shown in Figure 3.1. Besides directly using markers to for disease predication and monitoring, signal deconvolution methods can be categorized into either the reference-based approach when the tissue-specific reference profiles are known, or reference-free approach when the tissue-specific methylation reference profiles are unavailable.

3.4.4 Simulation

In order to evaluate and compare the aforementioned prediction approaches under different scenarios, we performed a series of simulations. In all simulations, data are generated to mimic the real data characteristics. The main simulation procedures are as following. We obtain the cfDNA WGBS methylation data of 32 healthy people as control samples and 29 hepatocellular carcinoma (HCC) patients as case samples from a previous study (Sun et al., 2015). Then we take the methylation levels for 1,013 CpG clusters (500 bp, each) from 14 different tissues as the reference panel $\mathbf{R}$, which

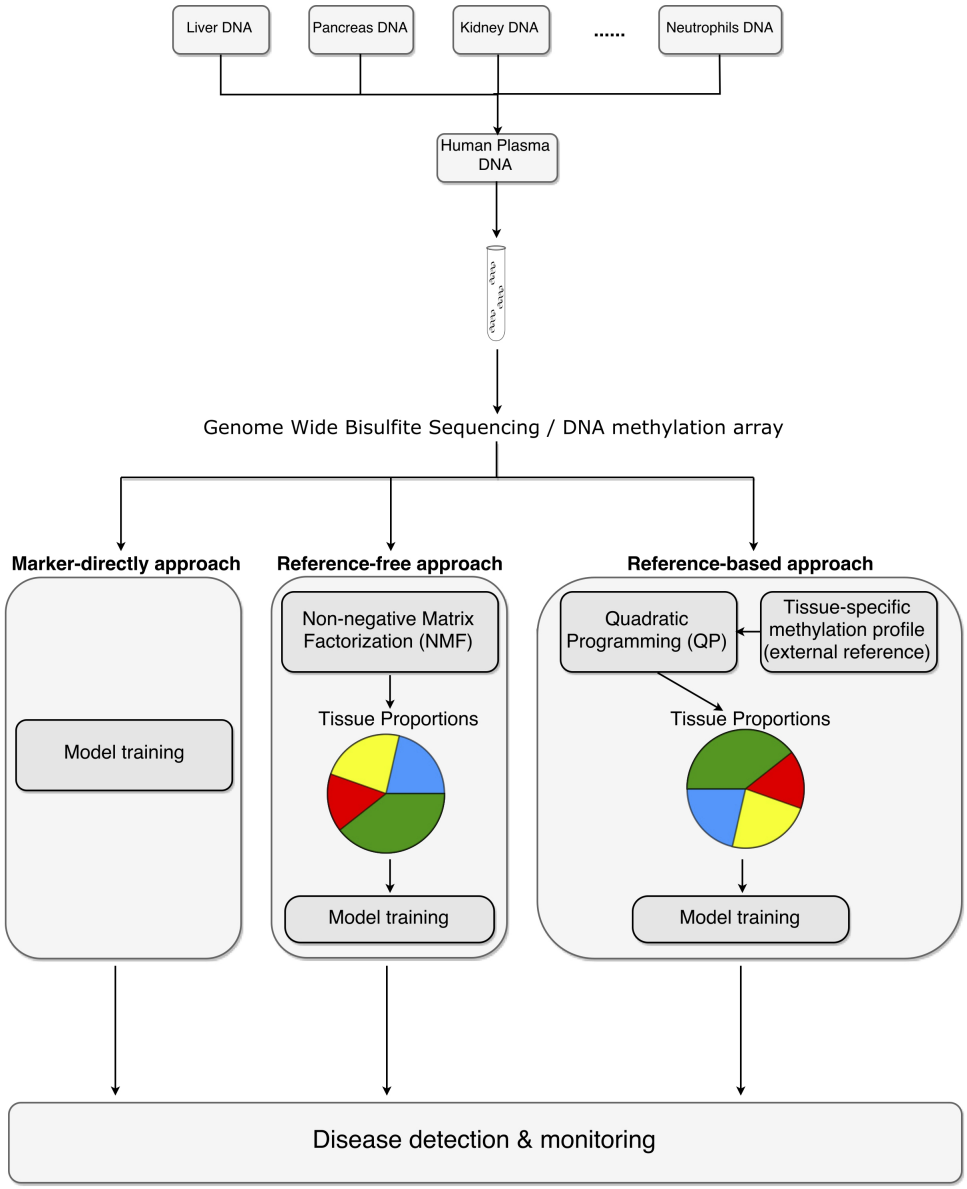


Figure 3.1: Schematic overview of plasma cfDNA methylation mixing procedure and deconvolution methods for disease detection and monitoring. One straightforward approach is to use marker directly for disease detection. Besides using biomarkers directly, signal deconvolution methods can be categorized into either the reference-free approach when the external tissue-specific methylation reference in unavailable, or the reference-based approach when the tissue-specific profile is known.

are reported by the authors and chosen based on tissue-specific methylation profiles. We further obtained the cfDNA methylation levels for all samples in these 1,013 regions as the methylation data of interest, and applied QP on the methylation data and reference panel to solve for tissue proportions for each patient. Next, we assume 32 healthy peoples tissue proportions come from a common Dirichlet distribution $Dir(\alpha_0)$, and 29 HCC cancer patients proportions from another common Dirichlet distribution $Dir(\alpha_1)$. We obtain the MLE of α_0 and α_1, as $\hat{\alpha_0}$ and $\hat{\alpha_1}$, respectively.

Using $\hat{\alpha_0}$ and $\hat{\alpha_1}$, we generate 50 controls tissue proportions $\mathbf{P}_0$ from $Dir(\hat{\alpha_0})$, and 100 cases' proportions $\mathbf{P}_1$ from $Dir(\hat{\alpha_1})$. To mimic the biological variation in reference panel for different person, we generate the noise-added reference panel $\mathbf{R}_i$ for each sample i base on the original reference panel $\mathbf{R}$. To be specific, we use the original reference $\mathbf{R}$ as the mean parameter in beta distribution, and then adjust the dispersion level based on simulation setting to control the noise level. Using higher dispersion will generate noisier reference panel $\mathbf{R}_i$. Then for each sample, we multiply $\mathbf{R}_i$ with the simulated mixing proportion to obtain the expected values for this individuals cfDNA methylation levels. The next step in simulation is adding noise to the simulated cfDNA methylation level, which is again based on beta distribution. We reparametrize the beta distribution $Beta(\alpha, \beta)$ into the following form:

$$Beta(\mu, \phi)$$

where $\mu = \frac{\alpha}{\alpha+\beta}$ is the mean and $\phi = \frac{1}{\alpha+\beta+1}$ is the dispersion. Here, we take $\mathbf{R}_i\mathbf{P}_i$ as the mean μ of $Beta$ distribution, and use different values for the dispersion ϕ to investigate the effect of noise levels on the performance of prediction.

3.4.5 Simulation Results

After obtaining the simulation data, we use leave-one-out cross validation (LOOCV) to evaluate and compare the classification accuracies from different methods. The classification accuracies from all simulations are summarized by the boxplots in Figure 3.2. Simulations are conducted under low (ϕ=0.17, Figure 3.2A), medium (ϕ=0.5, Figure 3.2B), and high (ϕ=0.67, Figure 3.2C) noise levels. Each simulation is repeated for 20 times. The methods under comparison include: marker directly predict approach (presented as marker), estimate tissue proportion approach using QP (QP), and the reference-free NMF approach (NMF). As a benchmark, we also include the results from using the true proportions as predictor (true prop).

As shown in Figure 3.2, using the true proportions as predictors achieves the highest accuracy in all simulation settings, as expected. When the noise level is low (Figure 3.2A), the prediction accuracies of all methods are reasonably good, with NMFs accuracy lower than the others. When the noise level increases, the three methods under comparison start to differ. At medium noise level, using marker directly to predict performs worse than QP ($p = 10^{-4}$, one-sided *t-test*) but better than NMF ($p = 10^{-3}$, one-sided *t-test*). At high noise level, using marker directly to predict performs worse than the other two methods ($p = 10^{-12}$ and 10^{-12} high; one-sided *t-test*). In particular, at high noise level (Figure 3.2C), directly using marker to predict perform rather poorly. This is because under our simulation setting, the methylation differences come from the differences in mixing proportions between cases and controls. The proportion estimation serves as a signal filtering step to extract better prediction features, which subsequently improves prediction. Across all noise levels, QP performs better than NMF (p= 10^{-12} low; 10^{-8} medium; 10^{-2} high; one-sided *t-test*). This is expected because QP utilize external information to help disease status prediction, which is supposed to outperform reference-free method NMF. We then conduct the following simulations to further investigate the QP and NMF methods

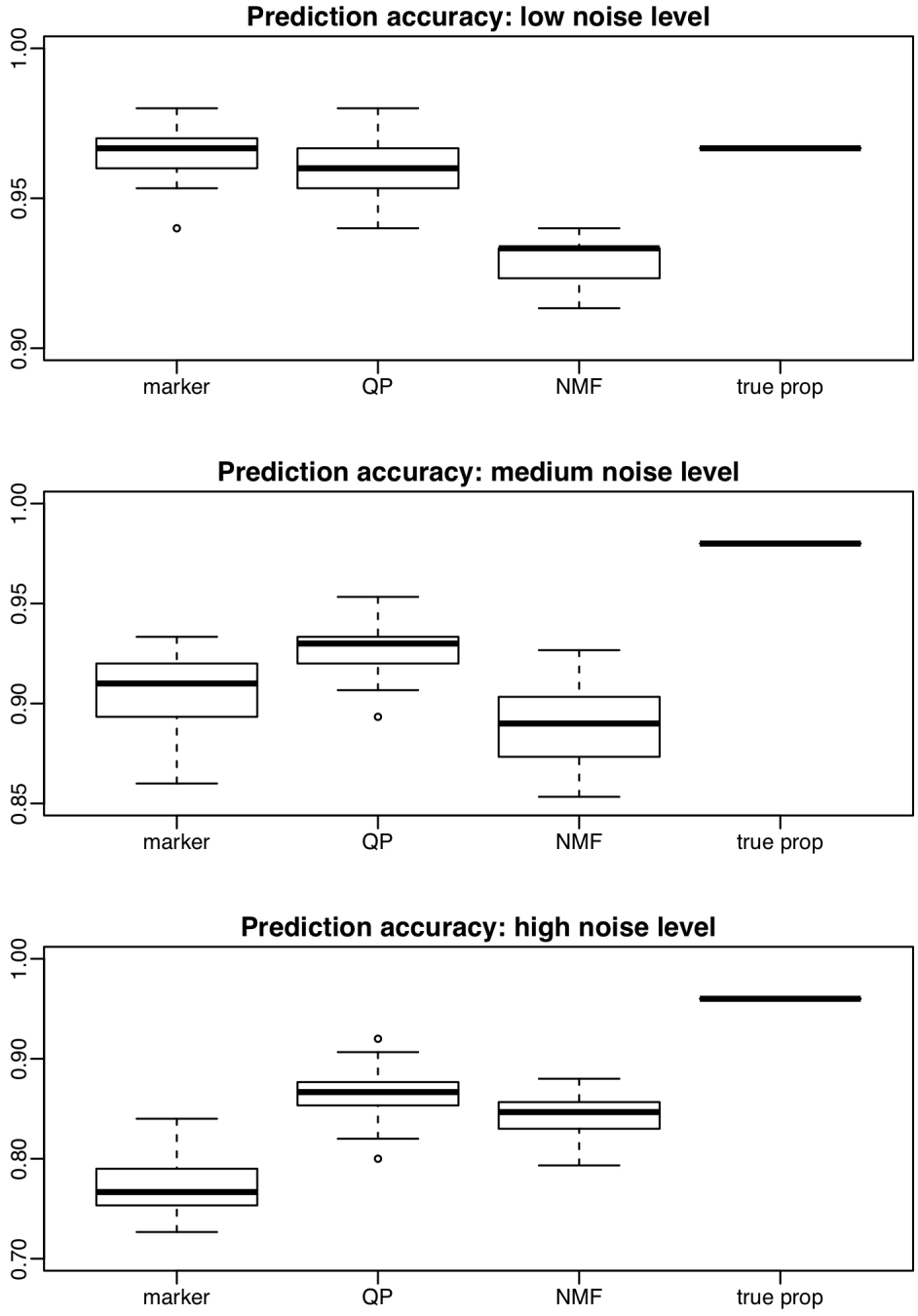


Figure 3.2: Boxplot of classification accuracies for multiple methods in simulations. Marker: Marker directly predict approach. QP: using tissue proportions solved from Quadratic Programming procedure for prediction. NMF: Non-negative matrix factorization (NMF) approach. True Prop: using simulated true proportion in classification. A total number of 20 simulations are conducted. A, low noise level; B, medium noise level; C, high noise level.

from other aspects.

Sample size consideration

Since the simulation above contains rather small sample size (150), we investigate how the size of training data will affect the results by increasing the sample size to 750 and 1500. Supplementary Figure B.1 shows that the total sample size has dramatic effect on prediction accuracy. As the sample size increase, the accuracies of all methods increase, across all noise levels. However, when the noise level is not low, NMF actually performs better than QP under larger sample size. This is because that when the noise level is high, the reference panel used for QP is noisy. In this case, it is suitable to use NMF for reference-free decomposition when the sample size allows. These results also provide some hints for sample size selection. For NMF, the gain of accuracy from 150 to 750 is dramatic, and then plateaued from 750 to 1,500. It is therefore advisable to have at least several hundred of samples to start using the NMF approach.

Other aspects in solving proportion

In either QP or NMF-based method, the estimated proportions are coordinates when projecting the original data into a lower dimensional space. The improvement in prediction accuracy using estimated proportions suggest that the coordinates contains cleaner signals for the outcome. We investigate how much impact the direction of the projection will have on prediction. We conduct simulations under different external reference $\mathbf{R}$, to see how the choice of reference will affect the classification results. We use a high variance reference in QP estimate tissue proportion approach to solve for tissue proportions. That is, more noise is added to the $\mathbf{R}$ used in QP. This mimic the situation that there is significant bias for the reference panel being used. We also try to use a random reference by randomly shuffling the entire $\mathbf{R}$ used in QP.

This mimic the extreme situation where the reference R is completely off. Results in Supplementary Figure B.2 indicate that using the high variance reference (QP high var) and randomly shuffled reference (QP random) both lead to a decrease of accuracy, where using random reference is much worse than all other methods (p=0.0159, *ANOVA*). Both QP and NMF project a matrix into lower dimensional space with either known or unknown coordinates. Whether the projection has predictive power for the outcome is important for the performance of the method. The results in Supplementary Figure B.2 illustrates that a bad projection direction leads to unfavorable prediction accuracy, and that using a more accurate external reference $\mathbf{R}$ will benefit the classification.

We also explore if directly solving proportions in ordinary least square (OLS) without any constraint will affect the prediction accuracy. Results in Supplementary Figure B.2 indicate OLS has comparable performance to QP. Of course, the OLS results will lose biological interpretation since without constraints, the regression coefficients cannot be interpreted as mixing proportions anymore. Thus, the QP is still a preferred method than OLS. When adopting reference-based approaches, we also compare QP with two other newly designed methods: Cibersort (Newman et al., 2015) and EpiDISH (Teschendorff et al., 2017). Cibersort (CBS) employs support vector regression (SVR), and EpiDISH uses robust partial correlations (RPC). Results in Supplementary Figure B.3 indicate CBS performs better than QP in high noise level, whereas RPC and QP are comparable overall. This indicates the reference-based algorithms that specifically designed for gene expression or DNA methylation data, where solved proportions constraint can be implemented *a posteriori*, can provide alternative means for QP.

Validation of NMF results

To validate if the NMF-solved reference matrix W is a good approximation to the true reference panel, we investigate the NMF results in simulation. Since the column orders of W is randomly generated from NMF, we first need to assign tissue types to the columns of W. To do so, we find matches based on pairwise correlations of the columns of W and the true references. In these two matrices, two columns with highest correlation are regarded to represent the same tissue. After this, we exclude these two matched columns and use the highest correlation on the remaining data to identify the second matched tissue. We iterate this procedure until all tissue types are determined. We found that overall, the estimations of the reference are accurate. The average correlation between estimated and true reference is above 0.83. Figure 3.3 shows the scatterplot for NMF-solved reference methylation versus the truth for four (out of 14) tissues. Such scatterplots for all tissues are available in Supplementary Figure B.4.

We further compare the NMF-solved proportion matrix H to the true proportions in simulation. Figure 3.4 shows the scatterplots of NMF-solved tissue proportions versus true proportions for the four tissues. In general, NMF-solved proportions correlate with true proportion well in most estimations, although in some tissues this relationship is weak. Possible reason for the inaccurate estimation in some cases is that the low abundance of certain tissue makes them difficult to estimate. The scatterplot for all tissues solved-proportion versus true proportions are available in Supplementary Figure B.5. Overall, reference-free approach has the capacity to elucidate compositions of heterogeneous cfDNA samples pertaining to their constituent homogeneous tissue types.

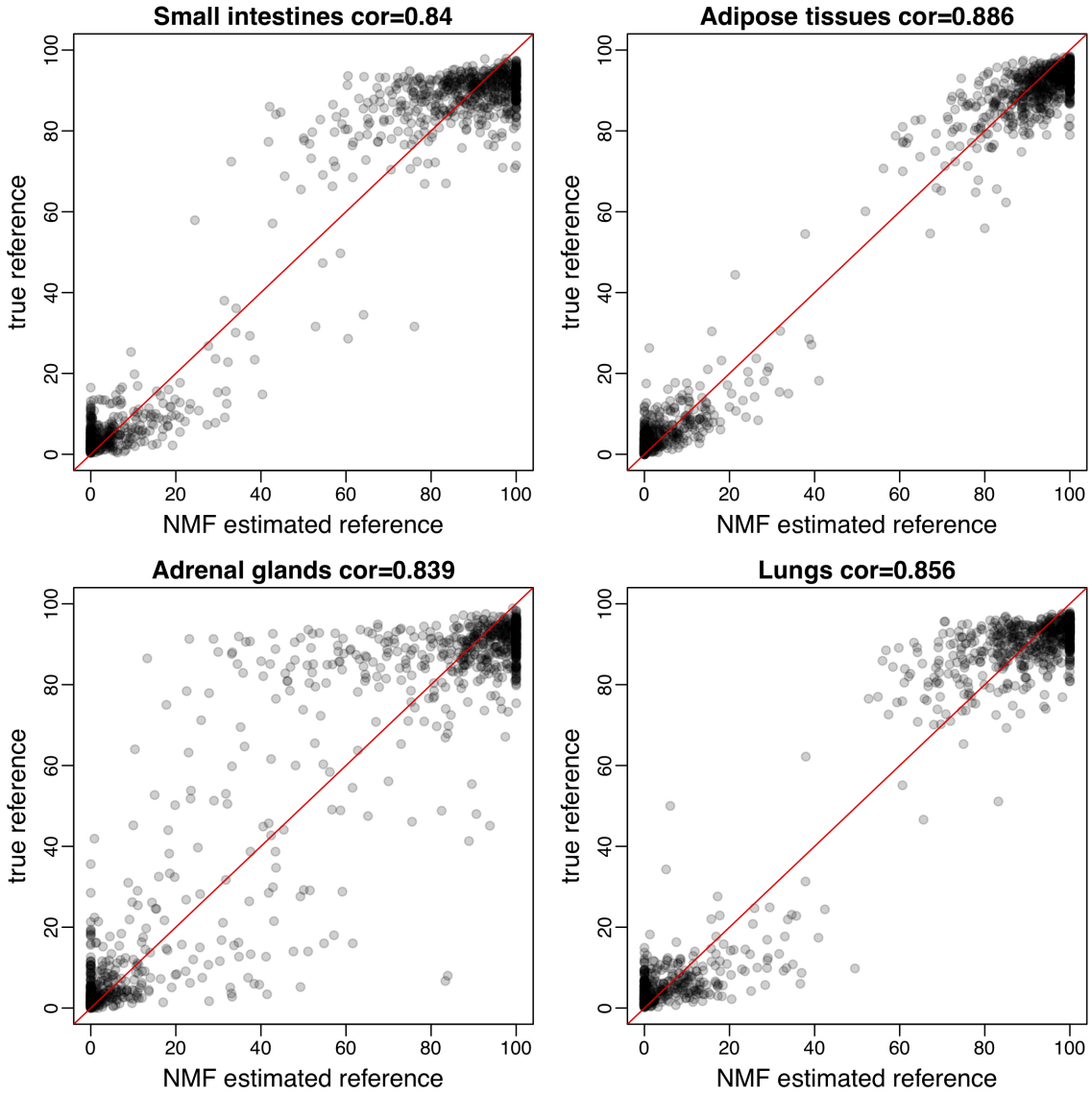


Figure 3.3: Scatterplots of NMF estimated reference methylation levels versus true reference methylation levels in 4 tissues. A, small intestines; B, adipose tissues; C, adrenal glands; D, lungs. Relatively strong correlations are observed. Spearmans correlation is shown in each panel.

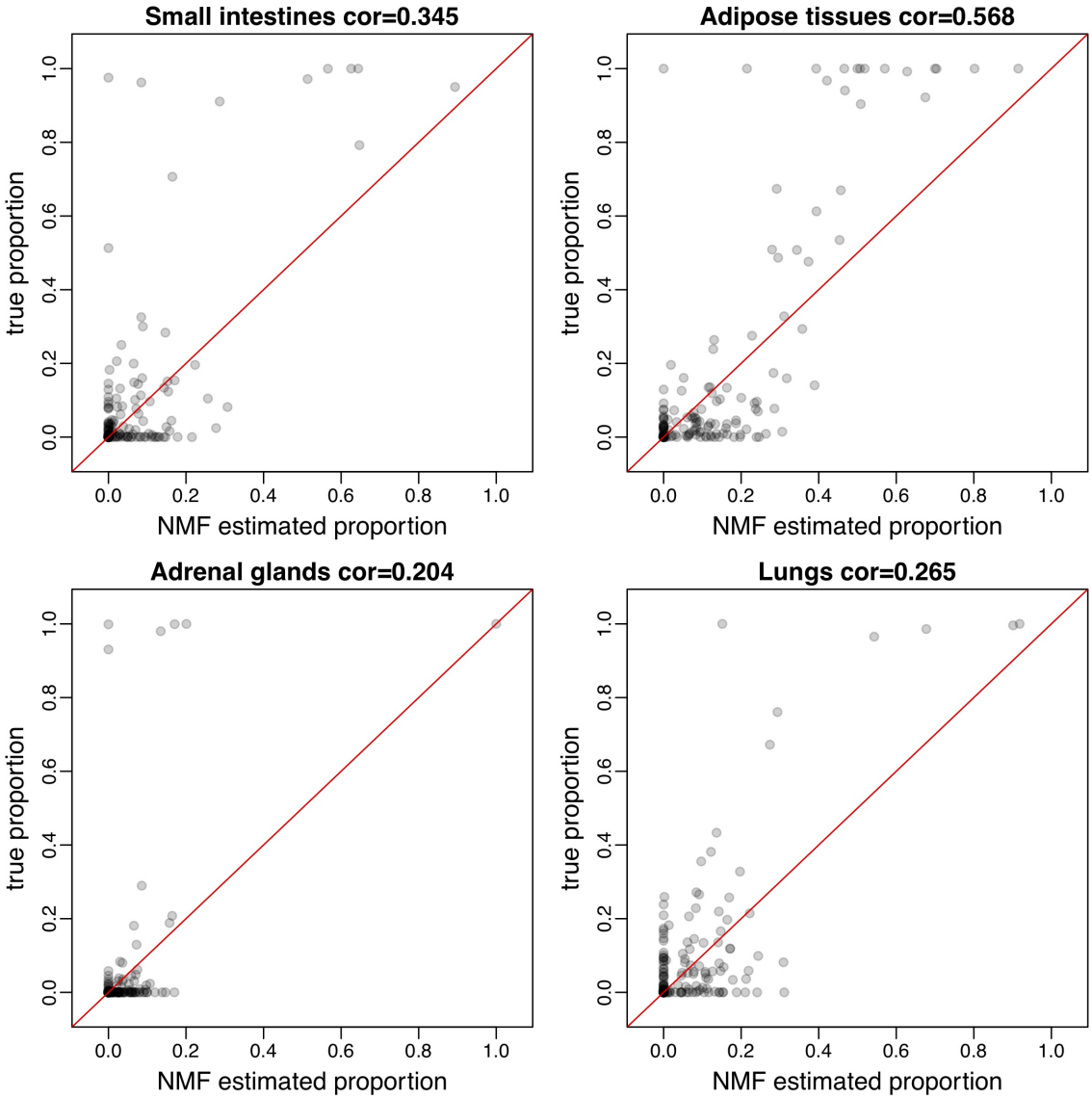


Figure 3.4: Scatterplots of NMF estimated tissue proportions versus true tissue proportions in 4 tissues. A, small intestines; B, adipose tissues; C, adrenal glands; D, lungs. Relatively strong correlations are observed. Spearmans correlation is shown in each panel.

3.5 Real data results

We further evaluate and compare the methods in real data. We obtain and process the cfDNA WGBS data for 27 hepatocellular carcinoma (HCC) patients, 32 healthy unpregnant control subjects, and 17 healthy pregnant subjects from Sun et al. (2015). This dataset is referred to as the WGBS dataset thereafter. The external reference panel is obtained from the same study, with reference data from the Roadmap Epigenomics Consortium (Kundaje et al., 2015) included. With this external reference panel known, we first apply QP to solve for tissue proportions. For each individual among HCC patients, healthy controls and pregnant subjects, the bar plots for estimated tissue proportions are shown in Figure 3.5. Each bar represents one person. To take a close look at the tissue proportions in a tissue-specific manner, the boxplot for liver and placenta tissue proportions among these 3 groups (HCC, control, pregnant) are shown in Figure 3.6. It demonstrates that HCC patients have an increased proportion of cfDNA originating from liver, which is consistent with the original study (Figure 7 in Sun et al. (2015)) and suggests that the cell death rates in liver are higher among HCC patients. Similarly, pregnant women show an increased proportion of cfDNA originating from placenta. The marked differences in these proportions indicate that the proportions will be predictive for the outcome.

The boxplots for all 14 tissue-type proportions, with one panel for each tissue type, are shown in Supplementary Figure B.6. We then apply NMF on real data to see if the NMF-solved result is similar to the truth. Although on average the correlation is not as ideal as in simulation, Supplementary Figure B.7 shows that NMF-solved reference correlates true reference well. NMF is effective for obtaining the underlying reference panel from real data.

We then apply 3 different methods to classify the HCC, control and pregnant subjects. The classification confusion matrices from leave-one-out cross-validation

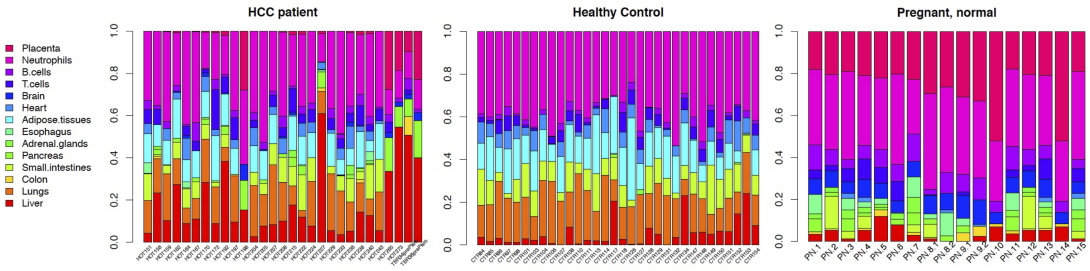


Figure 3.5: Barplot for the estimated 14 tissue proportions from real data for HCC patients, healthy controls and pregnant subjects, using Quadratic Programming (QP) with external reference available. HCC patients showed an increased proportion of cfDNA originating from liver, while pregnant controls showed an increased proportion of cfDNA originating from placenta.

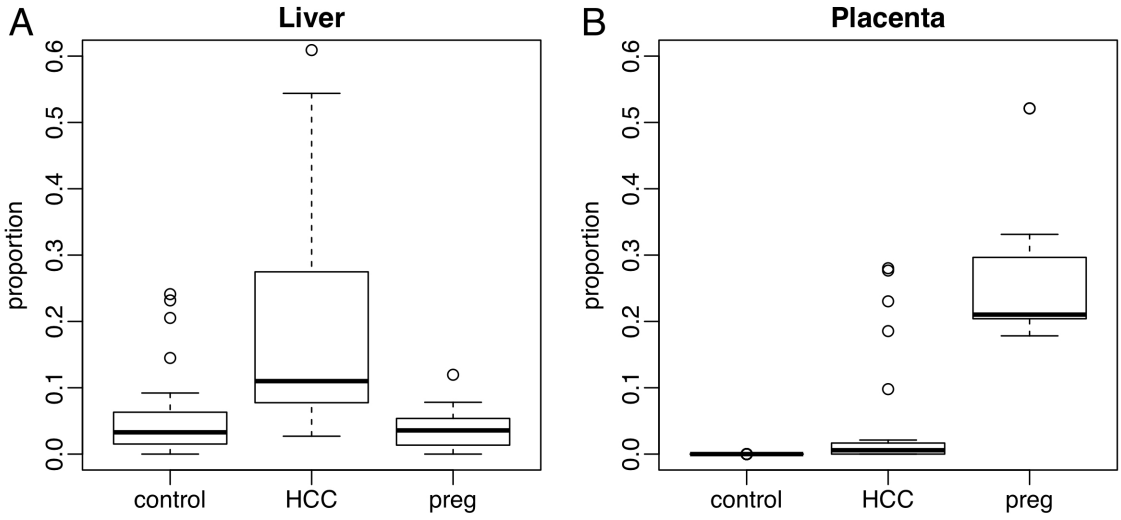


Figure 3.6: Boxplot of real data solved tissue proportions for liver and placenta, respectively, among 3 groups. A, tissue proportions for liver among 3 groups; B, tissue proportions for placenta among 3 groups.

(LOOCV) are shown in Table 3.2.

Method		**Marker predict**			**NMF proportion predict**			**QP proportion predict**		
		HCC	control	preg	HCC	control	preg	HCC	control	preg
Truth	HCC	11	12	4	8	15	4	14	13	0
	control	3	29	0	4	28	0	3	29	0
	preg	0	0	17	0	0	17	0	0	17

Table 3.2: Classification confusion matrices for the WGBS data. HCC: hepatocellular carcinoma patients. control: healthy, unpregnant control people. preg: healthy pregnant women. Marker predict accuracy: 0.75. NMF predict accuracy: 0.70. QP predicted accuracy: 0.79.

As shown in Table 3.2, QP based method has the highest classification accuracy (79%). It is because QP takes advantage of accurate external reference information, which helps to extract the proportion used in classification. Directly using markers for predication also yield satisfying predication accuracy and performs better than NMF approach. This is because when sample size is relatively small, NMF solved reference and proportions is not as accurate as in relatively large samples. QP based method can outperform NMF approach under small sample size setting. We also applied two other reference-based algorithms, CBS and RPC here. Supplementary Table indicates all three reference-based methods (CBS, RPC and QP) performs similarly. The results show that the pregnant subjects can be easily separated with other groups, while separating HCC patients with healthy controls yields more misclassification. It is because pregnant subjects show a more profound change in estimated proportions for placenta ($\sim 20\%$ in proportion change on average) compared with the rest groups, thus the signal-to-noise ratio is very high. For HCC patients, even though the proportion from liver is significantly higher in liver, there is still non-trivial overlaps in proportions between HCC and normal control, leading to the misclassifications. Overall, as a non-invasive pre-screening procedure, the real data results are reasonably good and show promises that cfDNA methylation can potentially be used for disease diagnosis. We also analyze a set of reference-free real data for further comparison. We obtain and process cfDNA hydroxymethylation data for 15 healthy controls

and 18 colorectal cancer patients from Li, Zhang, Lu, You, Song, Luo, Zhang, Nie, Zheng, Xu et al. (2017). The data were generated from capture sequencing technology known as 5hmC-Seal, which has similar data characteristics as MeDIP-seq. This dataset is referred to as the 5hmC-seq dataset thereafter. Since there is no external reference panel available for this dataset, we can only apply either marker-directly approach or NMF approach for disease prediction. We summarize the sequencing read counts on each 2kb regions along the genome, and then use the counts as inputs for disease prediction. During each round of LOOCV, top 1000 DMRs are first identified in the training samples using DSS (Feng et al., 2014). We then use the log-transformed read counts from the top 1000 DMRs as the input data for both marker-directly approach and NMF approach. The model is trained using top 1000 DMRs or deconvoluted proportions, respectively for marker-directly approach and NMF approach. The prediction result from LOOCV is shown in Table 3.3. Overall, using marker and NMF yield similar prediction accuracies, although using marker performs slightly better (one more correct prediction). Based on our observation, the signal-to-noise ratio in this dataset is reasonably high. Thus, the DMR markers themselves already have good differential power to detect the cancer-normal difference. Therefore, using marker-directly approach yield decent accuracy.

Method		**Marker predict**		**NMF proportion predict**	
		Cancer	Normal	Cancer	Normal
Truth	Cancer	11	4	10	5
	Normal	0	18	0	18

Table 3.3: Classification confusion matrices for the 5hmC-seq data. Cancer: colorectal cancer patients. Normal: healthy controls. Marker predict accuracy: 0.88. NMF predict accuracy: 0.85.

3.6 Discussion and Conclusion

Recent studies have reported that cfDNA contains rich information of disease status and can be used to extract biomarkers and construct disease prediction model (Kang et al., 2017; Tanić and Beck, 2017; Guo et al., 2017). As a non-invasive alternative to surgical biopsy, cfDNA-based assay has great potential in disease diagnosis. The highly promising and sought-after liquid biopsy in cancer diagnosis depends on cfDNA sequence variants, thus can only be applied on diseases with high mutation rate such as cancer. Using cfDNA methylation overcomes such limitation and has much wider application. In this study, we review the published works of using cfDNA in disease diagnosis. We focus on the strategies for statistical method and data analysis, and conduct simulations to investigate several potential methods for cfDNA methylation deconvolution and prediction for disease. The advantages and disadvantages for the three general approaches are summarized in Table 3.4.

Method	**Advantages**	**Disadvantages**
Marker-directly	· Straightforward and easy to apply · Applicable on disease with no cfDNA tissue proportion change	· Results lack direct biological interpretation · Results contain no tissue proportion information
Reference-based	· Can estimate tissue proportions · Tissue proportions have biological interpretation	· Require external reference panel from pure tissues
Reference-free	· Does not require external reference panel from pure tissues · Can estimate reference panel and tissue proportions · Tissue proportions have biological interpretation	· Computationally more intensive

Table 3.4: Advantages and disadvantages of three cfDNA methylation disease predicting approaches.

cfDNA is a mixture of DNA fragments from multiple tissues, and the mixing proportions are potentially associated with disease status. The difference in proportions will lead to some marginal cfDNA methylation changes due to the tissue specificity of

the methylomes. The disease prediction can be achieved by either using methylation levels or estimated mixing proportions as predictors, with an off-the-shelf machine learning algorithm. Regardless of the downstream prediction approach, marker selection is a very important first step. We review the approaches for selecting marker in existing works, and make some recommendation. In general, we recommend selecting markers based on the training data as well as external biological information.

When there is no profound change in cell type specific methylomes between cases and controls, it is generally assumed that the changes of tissue proportions in the mixing pool of cfDNA is associated with disease status. If the reference methylome are available, reference-based methods like Quadratic Programming (QP) can produce reliable tissue proportion estimation. Simulation studies show that the accuracy of using estimated tissue proportions to predict disease status is higher than that of using marker directly. As an added advantage, the estimated proportions also provide more interpretable result. In contrast, the reference methylome could be unavailable under certain circumstances. For example, the subpopulation under this study is different from the previous one. Under this situation that the reference panel is different from the original one, NMF is a viable solution. NMF-based method provides a reference-free approach for solving both tissue proportion and tissue reference simultaneously. Simulation studies demonstrate that this method provide comparable results to reference-based approach.

Although the disease prediction accuracy in real data is reasonable, there could be complications in real practice. The prediction can be influenced by biological and/or technical artifacts such as genetic background, demographics, or batch effects, which is a difficulty faced by many other genome-based predictive assays. For example, it has been shown that batch effect or different data normalization methods can negatively affect the prognosis in cancer using gene expression data (Qi et al., 2015). To alleviate these problems, the training and test sample first need to be consistent: they must

be from the same population and experimental platform. If significant batch effects were observed, one needs to first perform data normalization using approaches such as ComBat (Johnson et al., 2007), or consider using alternative rank-based methods to stabilize the signal. Furthermore, there will be room for improving the results. First, larger training samples size can contribute to the improvement in prediction accuracy. We recommend to start with at least several hundred samples to construct a prediction model. We believe with advances of experimental technologies and data analysis method, more cfDNA methylation data will be generated from larger-scale studies, which will greatly improve the model. We also envision that if the sample size increased significantly (e.g., doubled or tripled), we should retrain the model to improve accuracy. It is possible that with the retrained model, some diagnoses for existing patients could be different. In that case, the ethical issues have to be carefully addressed. However, this is the nature of clinical research: with accumulation of data and evidence, diagnosis criteria could evolve. Second, both reference-based and reference-free methods are dimension reduction approach to project data into lower dimensional space: reference-based method projects the data matrix onto the known reference, and reference-free method jointly solves the reference and projection. The prediction accuracy will be related to the reference used (as we showed in our simulation study). It will be very interesting to develop novel statistical method to identify the optimal low dimension space to project to that can produce the best prediction accuracy.

3.7 Key Points

- cfDNA is a mixture of DNA fragments from multiple tissues, and the mixing proportions could potentially be associated with disease status. cfDNA screening has great potential to be a non-invasive procedure for disease testing.

- Prediction based on cfDNA methylation can be applied to diseases not associated with significant DNA sequence changes.
- One can predict disease based on cfDNA methylation levels, or the estimated mixing proportions.
- Marker selection is very important for disease prediction using cfDNA methylation. It should be done using both the training data as well as external biological information.
- Mixing proportion estimation can be performed with or without reference methylomes.

3.8 Methods Availability

The R scripts implementing the methods discussed in this work are available online at: https://github.com/haoharryfeng/cfDNAmethy, with instructions and an example dataset.

Chapter 4

Cell-Free 5hmC in Alzheimer's Disease Patients

4.1 Introduction

Aging can be characterized as a set of accumulated cellular changes, leading to the impairment in various biological functions and increased risks in developing age-related diseases, such as cancer and neurodegenerative diseases (López-Otín et al., 2013). The functional deterioration in biological processes can be caused by different reasons, from a certain decreased metabolism to the alteration of epigenetic patterns (López-Otín et al., 2013). To evaluate aging, several biomarkers, including telomere length, mitochondrial DNA deletion and protein alterations, have been proposed, but none of them is sensitive enough for clinical practice (Meissner and Ritz-Timme, 2010). Methylation of the fifth position of cytosine (5-methylcytosine, 5mC) is one of the well-characterized epigenetic hallmarks. It is involved in the gene regulation and associated with many age-related diseases (Smith and Meissner, 2013; Reik, 2007; Bird and Wolffe, 1999). It has been widely reported that the dynamic change of 5mC occur during aging (Fraga and Esteller, 2007; Christensen et al., 2009; Alisch et al., 2012). Interestingly, the results from a recent study even suggest that the methylation status in as few as three selected CpG sites in blood DNA are reliably enough for age prediction (Weidner et al., 2014).

Cell-free DNA (cfDNA) is short DNA fragments that commonly between 160180 base pairs and released from different types of dead or dying cells to blood stream (Crowley et al., 2013). Recently, cfDNA isolated from plasma has shown great potential in clinical practice and been recruited as the a novel disease diagnosis biomarker, which is non-invasion, faster and more economic compare to the regular surgical biopsy (Crowley et al., 2013). For example, in the plasma of cancer patients, the cfDNA includes some circulating tumor DNA (ctDNA), carrying tumor-specific genetic variants, including point mutations and copy number variation (Schwarzenbach et al., 2011). The high-throughput microarray or sequencing experiment can distin-

guish ctDNA from normal cfDNA by tumor-specific genetic variants, enabling ctDNA serve as a reliable indicator for cancer due to its nature of mutation-rich. In addition, other approaches, such as exploring the cfDNA epigenome information, are developed for investigating those mutation-poor diseases not associated with significant genetic changes. In these diseases, cfDNA epigenetic information such as DNA methylation or nucleosome positioning can potentially serve as diagnosis biomarkers (Kang et al., 2017; Xu et al., 2017; Hao et al., 2017). The uniqueness or abnormality in epigenetic profiles of cfDNA for these mutation-poor diseases are suitable for constructing statistical models for prediction.

For decades, 5mC was recognized as the only epigenetic mark on DNA until the rediscovery of 5-hydroxymethylcytosine (5hmC) in mouse Purkinje neurons and embryonic stem cells (ESCs) in 2009 (Kriaucionis and Heintz, 2009; Tahiliani et al., 2009). The novel ten-eleven translocation (Tet) proteins can act as writers and convert 5mC to 5hmC (Ito et al., 2011). 5hmC plays an important role in DNA demethylation process because it can be further oxidized to form 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), which can be quickly removed from the genome by thymine-DNA glycosylase (TDG) to initiate base excision repair (BER) (Ito et al., 2011; He et al., 2011; Maiti and Drohat, 2011). Overall, 5hmC is enriched in gene-bodies and cis regulatory elements, and the alteration of 5hmC has been widely reported to be involved in regulation gene expression (Li, Yao, Chen, Kang, Li, Cheng, Li, Lin, Wang, Wang et al., 2017; Cheng et al., 2018; Pastor et al., 2011; Song et al., 2011). Moreover, 5hmC can stably exist during aging and be dynamically recognized by certain groups of proteins (5hmC readers) (Spruijt et al., 2013; Szulwach et al., 2011). These founding suggest that 5hmC is not only an intermediate in the demethylation process, but rather has its own biological functions, thus is an ideal epigenetic mark for aging and age-related diseases. Indeed, two recent publications have revealed the great potential of cfDNA 5hmC severing as a novel biomarker in

cancer diagnosis (Li, Zhang, Lu, You, Song, Luo, Zhang, Nie, Zheng, Xu et al., 2017; Song et al., 2017). However, so far, little is known whether 5hmC is dynamically altered during aging and could be served as a novel biomarker for aging or age-related disorders.

In this study, we investigated the genome-wide alteration of cfDNA 5hmC in young healthy subjects (23-30 years old), old healthy subjects (68-76 years old) and late onset Alzheimer's disease (AD) patients (67-90 years old). In healthy subject groups, we identified 193 aging-related DhMRs (differentially hydroxymethylated regions), which associated with the genes involved in synaptic functions. However, these 193 aging-related DhMRs do not necessarily associated with pathology of AD. A set of 236 distinct disease-associated DhMRs were further identified by comparing cfDNA 5hmC in old healthy subjects and AD patients. Interestingly, we found the adjacent genes of the disease-associated DhMRs were highly enriched in various brain regions, such as cingulate gyrus, prefrontal cortex and cerebellum, suggesting these disease-associated DhMRs containing cfDNA is potentially released from brain. In addition, using cfDNA 5hmC data, we constructed a classification machine learning model to predict AD from healthy old individuals. Our cross-validation results showed reasonably prediction accuracy. As far as we know, our work is the first investigation, both experimentally and computationally, to study the cfDNA 5hmC profile of neurodegenerative disease.

4.2 Materials and Methods

4.2.1 Case materials

Frozen plasma from 20 healthy individuals, including 10 young (23-30 years) and 10 old (67-76 years) Caucasian females (Supplementary Table C.1). All individuals were disease-free, non-smokers who are not taking any medication. In addition, we also

collected plasma from 10 individuals who were diagnosed with late onset Alzheimer's disease (67-90 years) (Supplementary Table C.1).

4.2.2 Genomic DNA preparation

Genomic DNA was isolated from brain samples with standard protocols. Tissues were homogenized on ice and then treated with proteinase K (0.667 $\mu g/\mu l$) in 600 μl digestion buffer (100 mM Tris-HCl, pH 8.5, 5 mM EDTA, 0.2% SDS, 200 mM NaCl) at 55°C for overnight. The second day, 600 μl of Phenol:Chloroform:Isoamyl Alcohol (25:24:1 saturated with 10 mM Tris, pH 8.0, 1 mM EDTA) (P-3803, Sigma) was added to samples, mixed completely, and centrifuged for 10 min at 12,000 rpm. The aqueous layer solution was transferred into a new Eppendorf tube, and the genomic DNA precipitated with 600 μl isopropanol. The pellet was washed with 75% ethanol, air-dried, and eluted with Nuclease-Free Water (Ambion).

4.2.3 5hmC specific chemical labeling, affinity purification, and sequencing

5hmC enrichment was performed using a previously described procedure with an improved selective chemical labeling method (Szulwach et al., 2011). DNA libraries were generated following the Illumina protocol for "Preparing Samples for ChIP Sequencing of DNA" (Part# 111257047 Rev. A) using 2550 ng of input genomic DNA or 5hmC-captured DNA to initiate the protocol. All sequencing libraries were run on Illumina Hi-seq 2000 machines.

4.2.4 Bioinformatics analysis

FASTQ sequence file from each sample was aligned to the *Homo sapiens* reference genome (hg19) using Bowtie2. All subsequent data analyses were performed using R language and Bioconductor packages. We first cut the whole human genome (hg19)

into equal sized bins of 5kb. Numbers of reads overlapped with each bin were obtained to represent the level of 5hmC at the corresponding regions. Only regions with a mean 5hmC level greater than 1 across all samples were kept for follow-up analysis. Genome-wide bin counts are compared between Old samples and Young samples, and between Old samples and AD samples for DhMRs detection using Bioconductor package DSS (Wu et al., 2012). DSS uses a negative-binomial distribution to model the sequence read counts, and has a Wald test for testing differences in two groups. DhMRs are defined as the regions with false discovery rate (FDR) less than 0.2, considering the multiple testing. Given the gender bias, we removed all the DhMRs in sex chromosomes.

Different genomic features of DhMRs were obtained using HOMER (Hypergeometric Optimization of Motif EnRichment) software (Heinz et al., 2010). A chromosome distribution plot is generated by the generic plotting function in R. Enrichment analyses were performed by using Enrichr (http://amp.pharm.mssm.edu/Enrichr/), which contains more than 140 gene-set libraries (released by 01/23/2019) (Chen et al., 2013; Kuleshov et al., 2016).

4.2.5 cfDNA 5hmC AD biomarker selection and the prediction model

First, we cut the genome into 5kb bins and obtained the count of reads in each bin along the genome for each individual. Then we performed normalization to make each individual sample's profile comparable. Next, we selected the biomarker for predicting AD with the help of an external solid brain tissue's profile. To be specific, we obtained the data from brain samples of a study that has both AD and healthy individual. We then identified the DhMR (referred to as 'brain-specific DhMR' below) by comparing the profile of AD versus healthy samples. We overlapped the brain-specific DhMR with our cfDNA data to obtain the candidate bins that are the potential biomarkers for AD cfDNA.

Among these candidate bins, the predictive biomarkers were further filtered by the within-group coefficient-of-variation, which favored the biomarkers with good consistency. To be specific, for each candidate bins, we calculated the coefficient-of-variation for healthy group ($CV_{healthy}$) and AD group (CV_{AD}), respectively. Then we sum the values from two groups up to obtain the overall coefficient-of-variation (CV_a), for each bin.

$$CV_a = CV_{healthy} + CV_{AD}$$

By ordering CV_a from the smallest to the largest genome-wide and retaining the top 15 bins that have the smallest CV_a, we obtained the cfDNA 5hmC AD biomarkers. The idea behind this selection approach is to keep the biomarkers with good within-group consistency, which are potentially more suitable for classification algorithm. Then the values from the cfDNA 5hmC AD biomarkers can be fit into machine learning model for AD prediction. Here we adopted Random Forest algorithm, which is a supervised ensemble learning approach for classification, by constructing decision trees using the features given.

We conducted leave-one-out cross-validation (LOOCV) to evaluate the effectiveness of these biomarkers. In each round of LOOCV, 9 AD samples and 9 healthy samples were used for model training and the rest 1 AD sample and 1 healthy sample were left for model testing. In each round of LOOCV, top 15 bins with smallest CV_a were retained as cfDNA 5hmC AD biomarkers. Random Forest model was trained on the training samples. Prediction was performed on the left-out testing samples. The LOOCV was iterated through all samples.

4.3 Results

4.3.1 Identification and characterization of aging-related DhMRs in cfDNA

To explore the dynamic change of cfDNA 5hmC during aging, we firstly isolated the cfDNA from the plasma obtained from two age groups, including 10 young healthy subjects (23-30 years old) and 10 old healthy subjects (67-76 years old) (Supplementary Table C.1). Genomic DNA was then isolated from each cfDNA sample. By incorporating a previously established chemical labeling and affinity purification method, coupled with high-throughput sequencing technology (Song et al., 2011), we genome-widely profiled the cfDNA 5hmC distribution, which was then evaluated by counting normalized 5hmC mapped reads in 5-kb binned human genome (hg19). We found more bins had higher 5hmC reads in old group compared with those from young samples (Supplementary Figure C.1). This finding is consistent with our previous observation (Szulwach et al., 2011), confirming that aging resulted in a global increase of 5hmC in the old group.

To detect aging-related DhMRs, the number of 5hmC reads in each 5-kb bin of human genome (hg19) were statistically analyzed between young and old samples. Given the depletion of 5hmC in sex chromosome (Szulwach et al., 2011) and sex bias of the subjects in this study (Supplementary Table C.1), we excluded the DhMRs identified in X and Y chromosomes. In total, we identified 193 aging-related DhMRs, majority of which (172 out of 193) showed increased 5hmC level in old group, suggesting that the maintenance of 5hmC in certain loci may be important to against aging (Figure 4.1A). Genomic annotation of these 193 DhMRs revealed that 20% of DhMRs overlap with known genes, while 80% of DhMRs are found in non-coding regions, including intergenic regions, short and long interspersed nuclear elements (SINE and LINE), and long terminal repeats (LTR) (Figure 4.1B). These DhMRs,

including both gain-of-5hmC and loss-of-5hmC regions, are universally distributed on autosome (Figure 4.1C).

To further explore the biological relevance of these aging-related DhMRs, we performed enrichment analyses by using Enrichr (http://amp.pharm.mssm.edu/Enrichr/), which contains more than 140 gene-set libraries (released by 01/23/2019) (Chen et al., 2013; Kuleshov et al., 2016). Gene ontology (GO) and pathway analyses showed that the nearest genes of aging-related DhMRs are significantly involved in synaptic function and in Heparan sulfate biosynthesis (Figure 4.1D). Dysfunction of synapse has been suggested as one of the hallmarks for cognitive decline, which is associated with normal aging (Morrison and Baxter, 2012). Even subtle changes, for instance, could increase the risk of neuron death and contribute to the pathogenesis of neurogenerative diseases, such as Alzheimer's disease (AD) in human (Bell and Hardingham, 2011; Nicholson et al., 2004). In addition, it has been demonstrated that aging in human is accompanied by specific alterations in heparan sulfate biosynthesis (Feyzi et al., 1998), which links to many features of the pathogenesis of AD (Van Horssen et al., 2003). Together these results have supported the previous knowledge of aging and indicate the clinical importance of cfDNA 5hmC as a potential biomarker for aging and age-related neurogenerative diseases.

4.3.2 Aging-related DhMRs in cfDNA are not necessarily associated with Alzheimer's disease

AD is a typical age-related neurodegenerative disease characterized by a progressive decline in cognitive functions, affecting more than 15 million people worldwide (Blennow, n.d.). In AD patient brains, loss of synapses and neurons is often observed in the cerebral cortex and some subcortical regions, which leads to severe atrophy of the affected regions, such as degeneration in the temporal lobe, frontal cortex and cingulate gyrus (Wenk et al., 2003). Majority of AD is late onset, typically ranging from

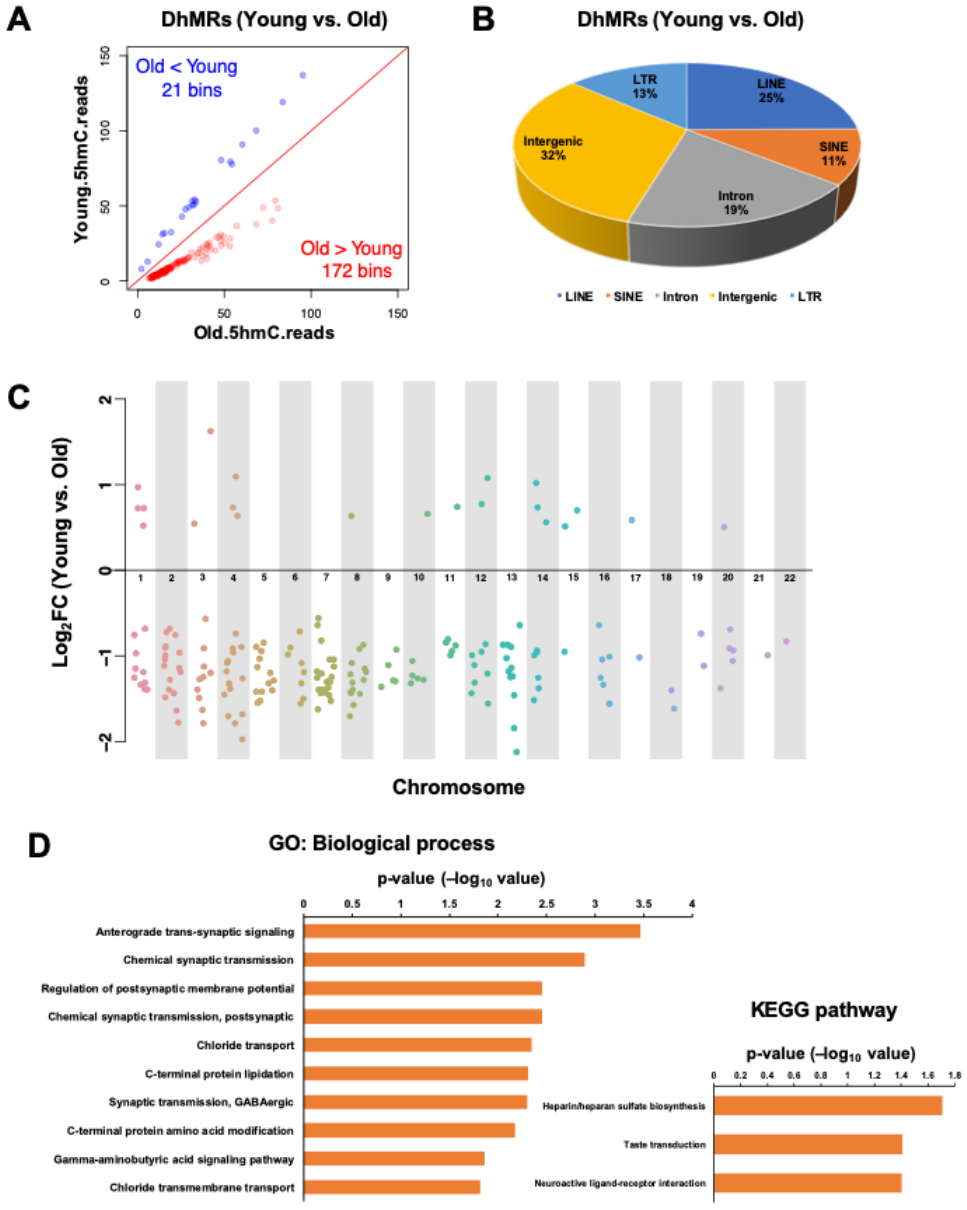


Figure 4.1: Identification and characterization of aging-related DhMRs in human cell-free DNA. (**A**) The number of 5hmC reads in each 5kb bin of human genome (hg19) were analyzed between healthy young (n =10) and old (n =10) individuals to find differentially hydroxymethylated regions (DhMRs). In total, 193 aging-related DhMRs were identified, and majority of the DhMRs (172 out of 193) showed increased 5hmC level in old group. (**B**) Genomic annotation of 193 aging-related DhMRs to show their percentage of each genomic region. (**C**) Chromosomal distribution of 193 aging-related DhMRs indicates they are relatively universally located in autosome. (**D**) Gene ontology (GO) and pathway analyses showed that the nearest genes of aging-related DhMRs are significantly involved in synaptic function and in several pathways including Heparan sulfate biosynthesis.

60 to 65 years, and associates with no clear genetic association or cause (Bekris et al., 2010). To test whether aging-related DhMRs are potentially associated with AD, we genome-widely profiled 5hmC in the cfDNA obtained from AD patient plasma and examined the number of 5hmC reads of AD cfDNA in those aging-related DhMRs. Based on 5hmC changes in each group, we divided aging-related DhMRs into four subgroups, including 45 "Increase (Young < Old < AD)" group, 127 "Peak (Young < Old > AD)" group, 21 "Valley (Young > Old < AD)" group, and 0 "Decrease (Young > Old > AD)" group (Figure 4.2A). Human phenotype association analysis of DhMRs in each subgroup highlighted a number of diseases or phenotypes related to aging, such as carotid artery diseases, stroke, blood pressure and glomerular filtration rate (Figure 4.2B). However, we did not see a clear associated between these aging-related DhMRs and AD. In addition, although there are certain trends of 5hmC change in each subgroup, none of them showed statistical difference in 5hmC changes between old and AD samples, suggesting these aging-related DhMRs in cfDNA are not necessarily associated with AD.

4.3.3 Identification and characterization of disease-associated DhMRs of Alzheimer's disease

To explore disease-associated DhMRs of AD, we therefore only focus on comparing AD samples to old samples, given their matched ages to minimum aging effect. The number of 5hmC reads in each 5-kb bin of human genome (hg19) were statistically analyzed between AD and old samples. In contrast to what we observed in the comparison between young and old groups (Supplementary Figure C.1), we found a large portion of bins had higher 5hmC reads in AD samples (76,497 bins) than those from old samples (19,901 bins), suggesting the development of AD is associated with the overall increase of 5hmC (Supplementary Figure C.2A).

Following the analysis mentioned above, we identified 236 disease-associated DhMRs,

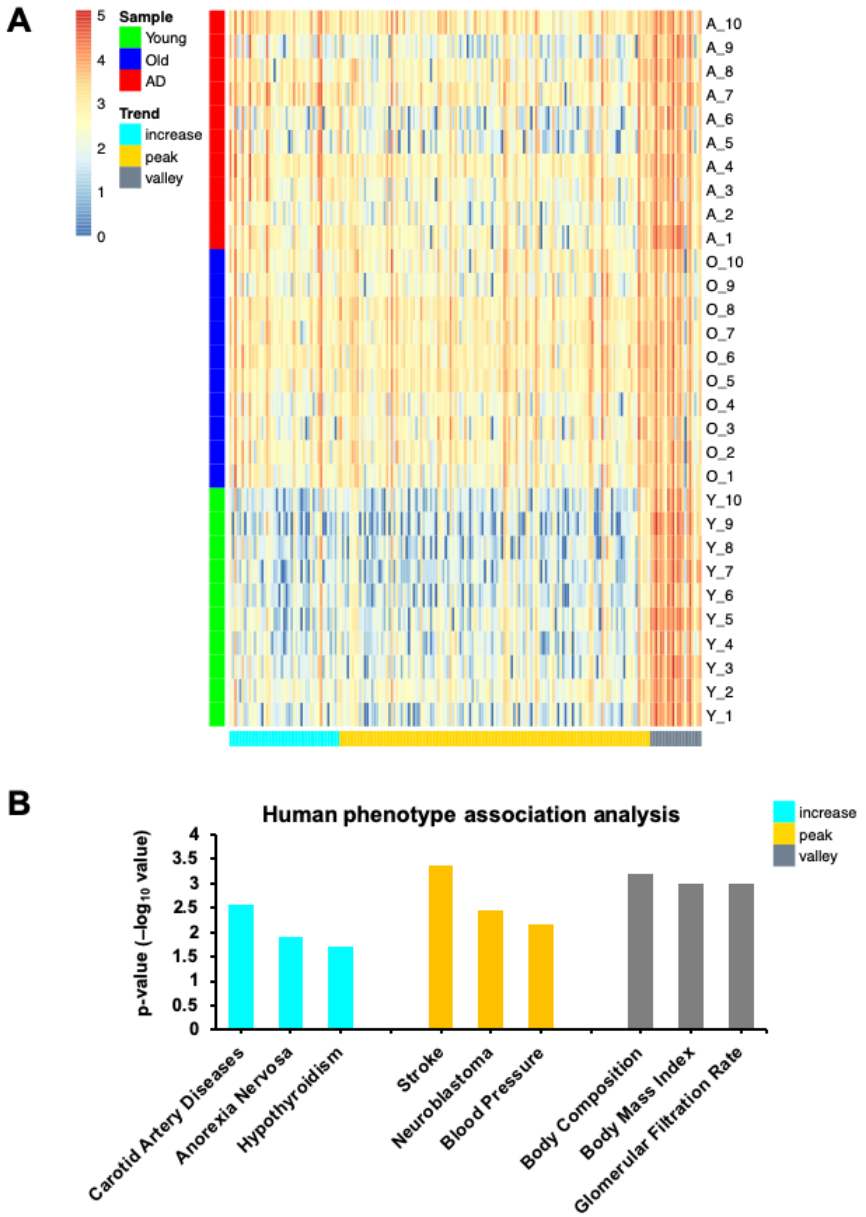


Figure 4.2: Aging-related DhMRs is not necessarily associated with Alzheimer's disease (AD). (**A**) The number of 5hmC reads of AD cfDNA were examined in the 193 aging-related DhMRs. Based on 5hmC changes in each group, aging-related DhMRs were further divided into four subgroups, including 45 "Increase (Young < Old < AD)" group, 127 "Peak (Young < Old > AD)" group, 21 "Valley (Young > Old < AD)" group, and 0 "Decrease (Young > Old > AD)" group. (**B**) Human phenotype association analysis of DhMRs in each subgroup highlighted a number of diseases or phenotypes related to aging, such as carotid artery diseases, stroke, blood pressure and glomerular filtration rate.

where 180 DhMRs have significantly higher 5hmC level in AD; while others have significantly lower 5hmC level in AD (Figure 4.3A and Supplementary Figure C.2B). Genomic annotation of these 236 disease-associated DhMRs indicated that 16% of DhMRs overlap with known genes, while 84% of DhMRs are found in non-coding regions, including intergenic regions, SINE, LINE, and LTR (Figure 4.3B). Similar to those aging-related DhMRs, GO analysis showed that the nearest genes of disease-associated DhMRs are also significantly involved in synaptic function; while pathway analysis highlighted Neuroactive ligand-receptor interaction (Figure 4.3C and Supplementary Figure C.2C).

Given the cfDNA is mixture of DNA fragments from various tissues, it is usually difficult to distinguish the resource of the cfDNA. To further explore the tissue specificity of the DhMRs, we performed enrichment analysis by using Enrichr with the ARCHS4 Tissues library, which contains the genes that highly expressed in human tissues. Interestingly, we found the adjacent genes of the aging-related DhMRs were only enriched in cingulate gyrus, while those of disease-associated DhMRs were highly enriched in different regions of central nervous system, including cingulate gyrus, prefrontal cortex, cerebellum and spinal cord, suggesting these disease-associated DhMRs containing cfDNA is potentially released from brain (Figure 4.3D).

4.3.4 Prediction of Alzheimer's disease using cfDNA 5hmC biomarker

cfDNA has the potential to serve as the biomarker for disease diagnosis, and here we built a prediction model to use the cfDNA 5hmC biomarker to predict Alzheimer's disease. The goal of this computational model is to distinguish AD patients and healthy old individuals, using their cfDNA 5hmC profile.

We utilized an external dataset from an AD study on brain tissues to assist our cfDNA 5hmC biomarker selection and conducted cross-validation to evaluate the prediction accuracy (Figure 4.4). Here, we first identified DhMR when comparing AD

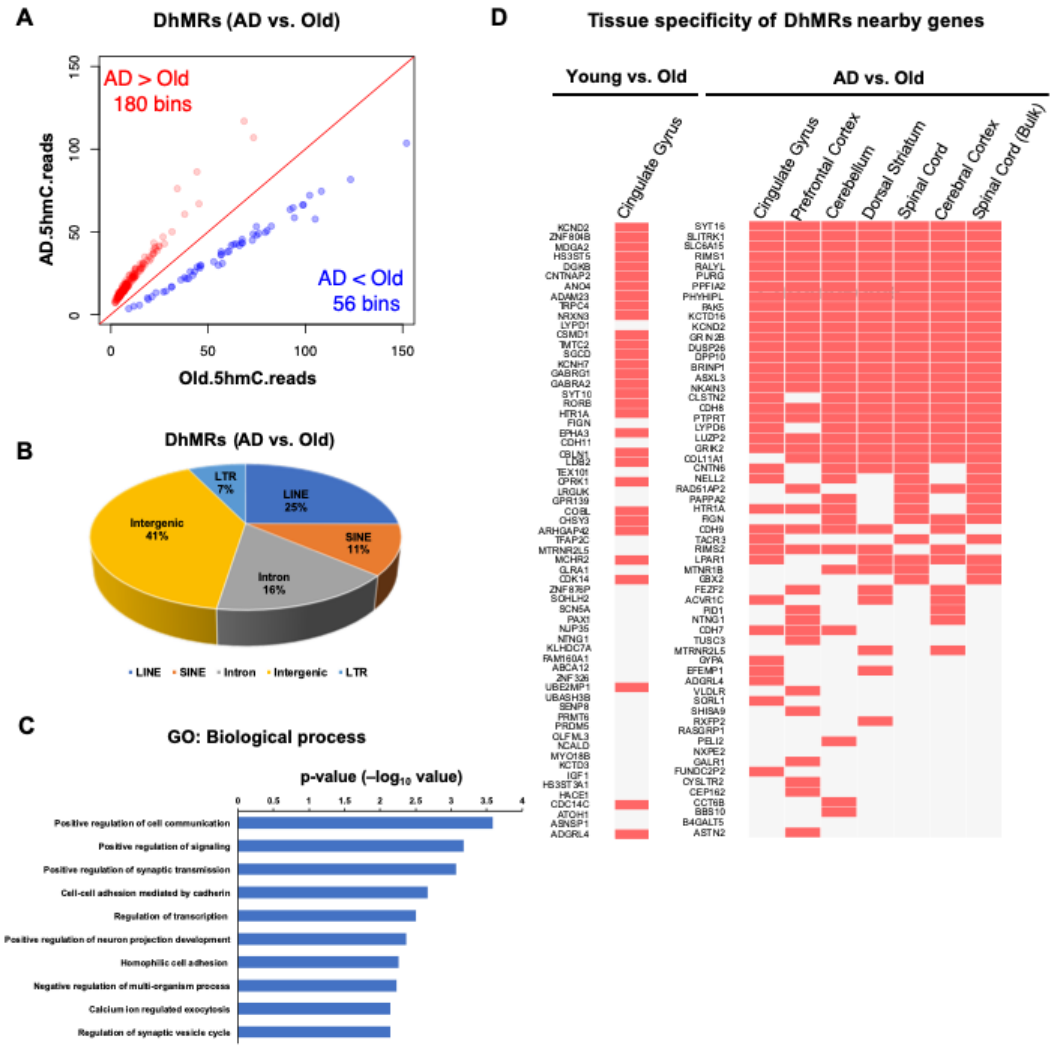


Figure 4.3: Identification and characterization of DhMRs associated with AD. (**A**) The number of 5hmC reads in each 5kb bin of human genome (hg19) were analyzed between AD patients (n =10) and healthy old (n =10) individuals to find differentially hydroxymethylated regions (DhMRs). In total, 236 disease-associated DhMRs were identified. (**B**) Genomic annotation of 236 disease-associated DhMRs to show their percentage of each genomic region. (**C**) Gene ontology (GO) analyses showed that the nearest genes of disease-associated DhMRs are significantly involved in synaptic function. (**D**) Enrichment analysis by using Enrichr with its ARCHS4 Tissues library were performed to explore the tissue specificity of the disease-associated DhMRs. The adjacent genes of the aging-related DhMRs were only enriched in cingulate gyrus, while the disease-associated DhMRs were highly enriched in various brain regions.

versus healthy samples using brain tissues data. Next, we overlapped genomic regions from these DhMR with our cfDNA data and retained biomarkers that showed good consistency. After the selection of biomarkers, the prediction model was constructed using Random Forest algorithm, which is a supervised ensemble learning approach for classification based on decision trees. This machine learning model was built on 5kb genomic regions as biomarkers. Iteratively, using leave-one-out cross-validation (LOOCV), we split the samples into training and testing. The model from each round of LOOCV will be trained on 9 AD samples and 9 healthy old samples. Testing accuracy is evaluated on the left-out 1 AD sample and 1 healthy old sample. This procedure is iterated through all samples. Results from LOOCV showed our models the prediction accuracy is 80% (Table 4.1).

	Predicted AD	Predicted healthy
True AD	9	1
True healthy	3	7

Table 4.1: AD prediction accuracy from leave-one-out cross-validation (LOOCV). Iteratively, Random Forest model was trained with training samples (9 AD and 9 healthy), and then applied on testing samples (1 AD and 1 healthy) to obtain the prediction. Accuracy: 80%.

4.4 Discussion

Epigenetic information, such as DNA methylome, in cfDNA has been investigated to pursue alternative biomarkers for clinical diagnosis of various diseases, including cancers and aging-related disorders (Feng et al., 2018; Jung and Pfeifer, 2015). In the present study, for the first time, we investigated the potential of cfDNA 5hmC in severing as a clinical biomarker. We successfully detected the 5hmC signals from cfDNA that isolated from plasma and identified 193 regions with dynamic 5hmC changes (aging-related DhMRs). These DhMRs reflects the changes between young and old healthy individuals and are strongly associated with the genes playing impor-

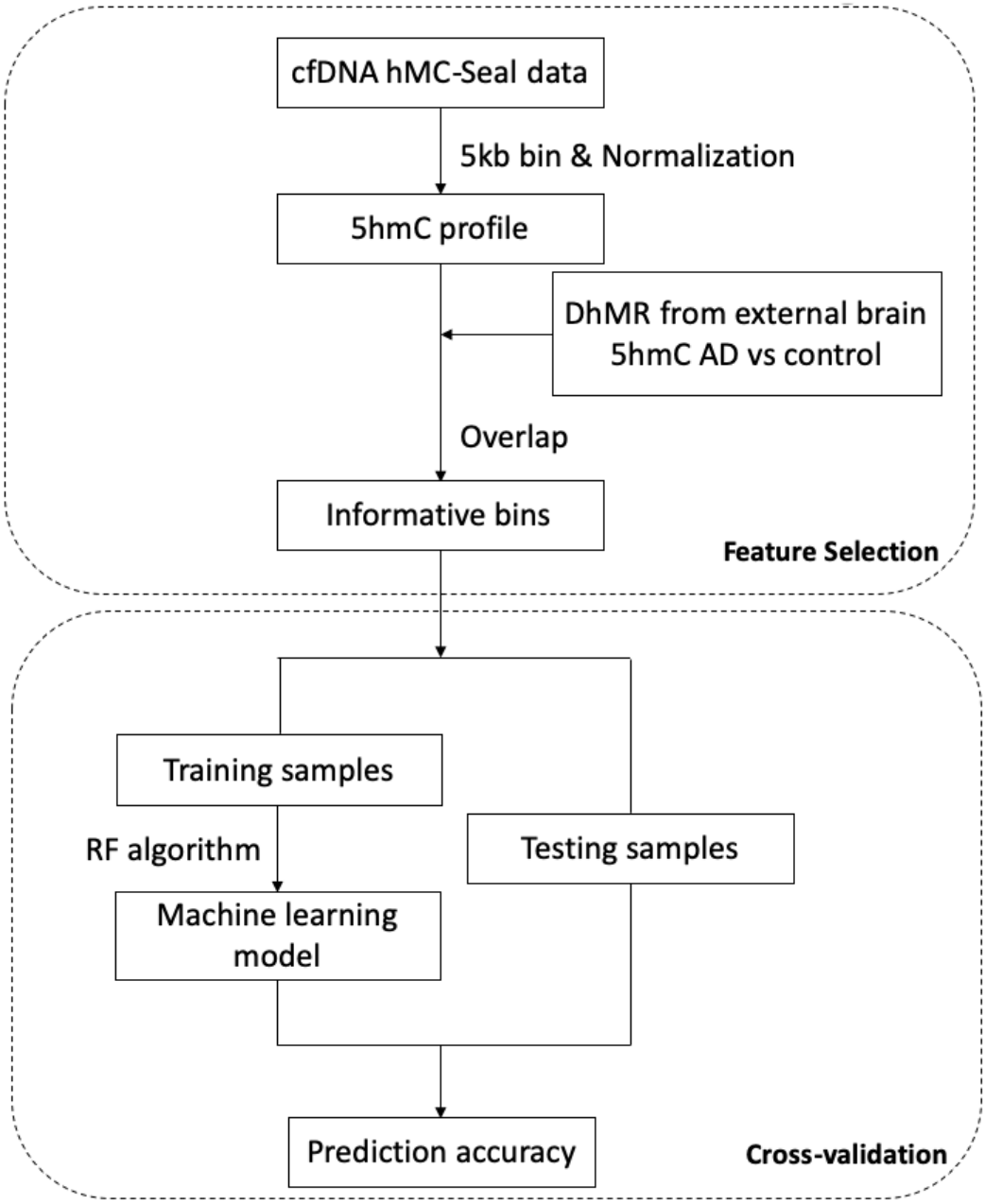


Figure 4.4: Schematic overview of cfDNA 5hmC AD prediction model. In the feature selection, external brain samples were utilized to help find the candidate features. In cross-validation, leave-one-out cross-validation (LOOCV) was adopted to evaluate the performance of AD prediction model. RF: Random Forest.

tant roles in synaptic functions. Although the various changes during human aging often contribute to the pathology of AD (López-Otín et al., 2013), we found these aging-related DhMRs identified in healthy groups do not necessarily associated with AD. Alternatively, by comparing cfDNA 5hmC in AD and old healthy individuals, we further identified a set of 236 distinct disease-associated DhMRs. Interestingly, we found these disease-associated DhMRs were associated with the genes that highly expressed in central nervous system, potentially reflecting the striking changes in the brain during the development of AD. Furthermore, we proposed a prediction model to use the cfDNA 5hmC biomarker to predict Alzheimer's disease. Cross-validation showed reasonably accuracy in predicting AD patients from healthy old subjects.

In recent years, disease diagnosis using molecular biomarkers from plasma-isolated cfDNA gained tremendous attention. For example, researchers examined specimens like plasma, blood, urine and cerebral spinal fluid to identify traces of biomarkers for cancer and other diseases. This cheaper, safer and less invasive alternative approach exhibits potentials in clinical practice. Here, using cfDNA 5hmC marker, our study confirmed the merit of cfDNA in aging and AD research. Our disease prediction model also verified the capability of using cfDNA 5hmC markers to prediction AD subject from health old group.

It is known that 5hmC is highly enriched in the brain (Globisch et al., 2010), and the distribution of 5hmC across the genome in human and mouse brain has been explored by our group and others. The distribution of 5hmC is tissue- and cell-type-specific. For example, it has been reported that 5hmC is particularly enriched in synaptic genes with a tissue-specific differential distribution at exon-intron boundaries, suggesting a potential role for 5hmC in RNA splicing in brain (Khare et al., 2012). In addition, 5hmC is consistently found to be enriched at euchromatin in both mouse and embryonic stem cells (ESCs) and neuronal cells, but is preferentially distributed at cis-regulatory elements, where the enrichment is more significant in

human than in mouse (Szulwach et al., 2011; Chen et al., 2014; Ficz et al., 2011; Pastor et al., 2011; Wu et al., 2011).

Although the etiology of sporadic AD has not been elucidated, it has been suggested that alterations in the level of epigenetics could be involved in its pathophysiology (Berger et al., 2009; Dupont et al., 2009; Waddington, 1939; Waseem Bihaqi et al., 2012; Irier and Jin, 2012). Indeed, our previous studies have shown an age-dependent acquisition of this modification in genes linked to neurodegenerative disease (Song et al., 2011; Szulwach et al., 2011). Other earlier studies suggested that numbers of epigenetic marks are changed in AD brain, including the decreased 5mC and DNMT1 levels in AD-vulnerable neurons (Mastroeni et al., 2010) and a global 5mC and 5hmC increase in the middle frontal gyrus and middle temporal gyrus (Coppieters et al., 2014). Those global increased levels of 5mC and 5hmC were correspondingly correlated and also positively associated with typical markers of AD, including amyloid beta, tau, and ubiquitin load (Coppieters et al., 2014). Interestingly, although the massive aging-associated changes in the brain are caused by oxidative stress, a mice study suggested that the aging-associated increase of 5hmC in hippocampus is independent of oxidative stress and possibly caused by the altered Tet enzymatic activities (Chen et al., 2012). A recent study examined the epigenetic alteration during AD progression by analyzing the hippocampus and parahippocampal gyrus of preclinical AD and late-stage AD patients and found significantly increased levels of Tet1, 5mC, and 5hmC, but decreased 5fC and 5caC levels (Bradley-Whitman and Lovell, 2013). Notably, in support of these observations, our current results revealed a substantial global increase of 5hmC in AD cfDNA and that the majority of the disease-associated DhMRs are gain-of-5hmC. Our results also indicated that those disease-associated DhMRs are significantly associated with the genes that highly expressed in AD-vulnerable brain regions, suggesting a great potential of cfDNA 5hmC serving as a reliable biomarker for disease prediction.

Our prediction model was built by borrowing information from brain tissue study. This step stabilized the signals and helped with feature selection. Results showed improved performance by utilizing the external brain tissue study information. The interpretation is two-fold. First, cfDNA data signal is noisy; therefore, feature selection and information extraction are vital for properly utilizing and interpreting the cfDNA data. Second, the 5hmC signals from cfDNA is associated with signals in brain. Studying such association is still an open question.

In summary, we investigated the genome-wide alteration of cfDNA 5hmC in young healthy subjects, old healthy subjects and late onset Alzheimer's disease (AD) patients. We identified the dynamical alteration of 5hmC during aging process and in age-related disorders. We constructed a classification model to predict AD from healthy old individuals. This is the first investigation to study the cfDNA 5hmC profile of aging and in neurodegenerative disease.

Chapter 5

Conclusion and Future Research Plan

5.1 Conclusion

In this dissertation, I present some statistical models and data analysis strategies for high-throughput epigenomics data. The coherent theme here is to use epigenomic biomarkers for disease prediction and classification. In my first project, I focus on real-world solid tumor samples and utilize DNA methylation data to conduct cancer subtype clustering. During this procedure, I take the tumor purity into consideration to solve the mixture signal problem. Moving from solid tumor samples to blood assay, I focus on cfDNA in the second project. I use cfDNA methylation information to predict disease. This liquid biopsy approach could potentially have broad applications for early diagnosis. Along the trajectory of studying cfDNA, my third research project aims at investigating 5hmC in cfDNA environment. Here I particularly focus on cfDNA 5hmC's association with aging and AD.

Epigenomics has been shown to be associated with transcriptome, and is further linked to phenotypes. Epigenomics study offers great potentials in elucidating the underlying mechanism for disease. Currently, however, there are still many questions that have not been answered. For example, how is the epigenome heterogeneity varies from cell-to-cell or cell-type-to-cell-type? At the cellular level, how will the epigenetic affect transcriptome? With my experience in epigenomics research and the advancement in sequencing technology, I will further investigate and address these fundamental questions in epigenomics. Therefore, I am proposing my future research plan, with more dedicate investigation into single-cell epigenomics.

5.2 Future Research Plan

Nowadays, the advancement in biotechnology is a prime driven force of new biological discoveries. Single-cell sequencing technology is one of the emerging ones. Traditionally, the NGS was conducted on bulk samples, which contains hundreds of thousands

of cells. The signal we observed is the averaged signal across a cell population. The advances of single-cell sequencing technology now allow us to inspect the DNA methylation profile of each cell, one by one, the basic unit of living creatures. Single-cell epigenomics study presents a promising direction to a better understanding of disease etiology and leads to new drug targets and strategies for personalized treatment.

The data from single-cell DNA methylation experiments offer great opportunities as well as challenges for data analyses. Unlike the traditional bulk DNA methylation sequencing data, the single-cell DNA methylation sequencing data lack well-developed statistical methods and tools. For example, there is not yet a consensus on statistical method for cell clustering. Moreover, there is a need of tools to impute missing single-cell methylation values. To address the problem of the shortage of statistical methods and tools, and serve the immediate needs, I am planning to work on the following research directions.

5.2.1 Single-cell Methylation Missing Value Imputation

Due to the low capture efficiency and stochastic CpG methylation, the overall missing fractions are high in single-cell methylation data. The sparsity of single-cell methylation data is a major hurdle to effectively study cellular heterogeneity. With large proportion of CpG missing, it is difficult to investigate cell-to-cell heterogeneity in the genome-wide scale. Imputing the missing values will enable us to study the regulation and the dynamics of DNA methylation in single cells at the whole genome scale, and to help reveal new linkages between epigenetic and other genomics features.

Here I will use histone modification data to help imputing the single-cell methylation status. Based on previous research, DNA methylation level is diverse at various functional genomic regions. I will develop statistical model for missing methylation level imputation, with genomic information incorporated. Within each functional genomic region, methylation/unmethylation transition probability can be estimated.

With the estimated probability, a Markov model can be utilized for imputation. The schematic overview of this method is shown in Figure 5.1.

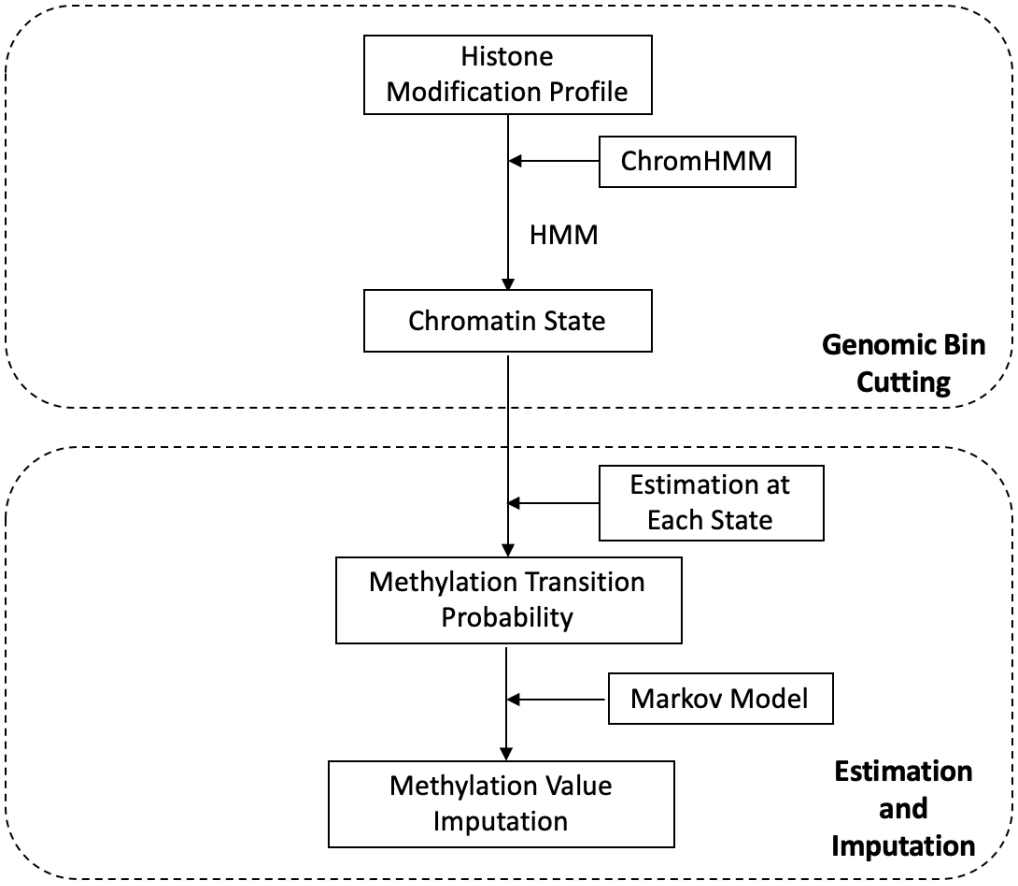


Figure 5.1: Schematic overview of single-cell methylation value missing imputation. Histone modification data can help us cut the genome into multiple regions. Transition probability of methylation/unmethylation can be estimated from each region. Then Markov Model can be constructed to impute the missing value form nearby CpG site.

5.2.2 Cell Clustering Using Single-Cell Methylation Data

Although the notion of a cell type is intuitively clear, it still lacks a consistent and rigorous definition. In the past, researchers use the size or the shape to define cell types. Later, researchers start to use the existance of proteins on the surface of cells to define the cell type. Now, single-cell methylation data provides a first time data-driven, coherent and unbiased approach that can help investigating the natural

groupings of cell population. Besides providing a deeper understanding of basic unit of living organism or tissue, clustering can provide references for disease studies. Therefore, I will develop methods for unsupervised clustering of single cell methylome data. Feature selection and dimension reduction will be integrated to unsupervised clustering procedure. The method will also allow continuous cell grouping to construct the pseudotime of cells.

5.2.3 Cell Type Methylome Profile Construction

In biological and clinical research, it is often of great interest to inspect the methylome characteristics of each of the cell type composition of a certain tissue or organ. Single-cell methylation data can bridge the gap between a cell populations signals and individual cellular behaviors. The data can be used to accurately construct cell subtypes methylome. To achieve this, I will jointly model scBS-seq data and bulk BS-seq data, and account for the discrepancy between them. Figure 5.2 provide a schematic illustration of how to the pure cell type methylome. On the top level, the observed methylome from bulk BS-seq can be modeled as an aggregation of methylomes from multiple cell types, with different weights on each pure cell type. On the second level, the methylome from each cell type is an aggregation of methylomes from individual cells. Hierarchical modeling can help obtaining the latent cell type methylomes, with incorporation of biological and technical noise. I aim to achieve a more accurate and robust estimation of cell-type methylome profile than estimating from bulk or scBS-seq alone. This will also provide an useful database reference for researchers to study the composition of his/her's own mixture sample.

5.2.4 Methods for Jointly Profiled Single-Cell Data

As single cell methylation sequencing technology evolves, parallel single-cell sequencing protocols are emerging. These parallel single-cell sequencing technologies allow for

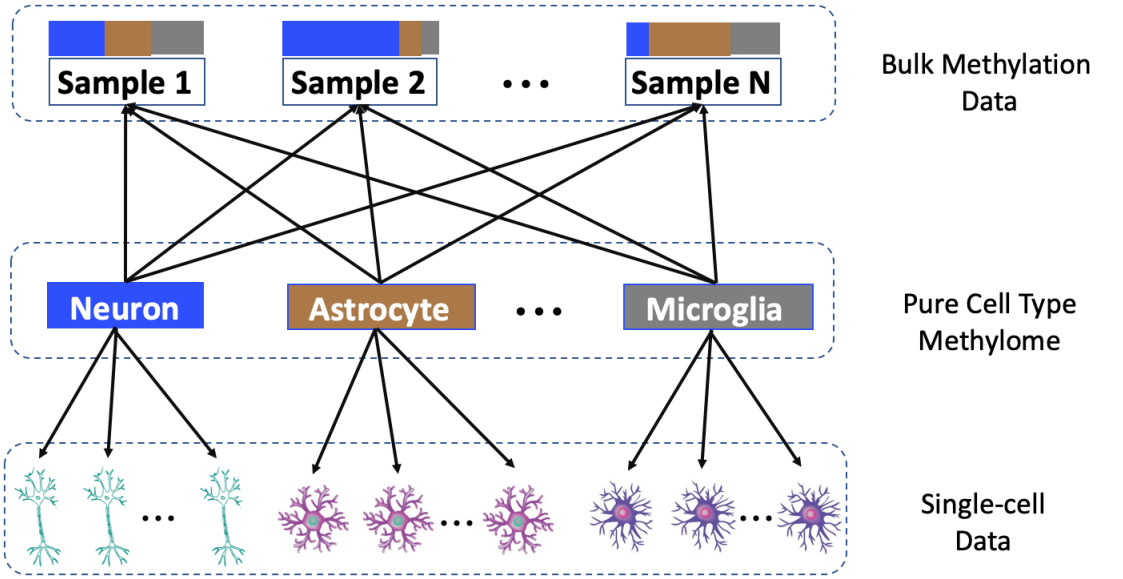


Figure 5.2: Schematic overview of cell type mixture problem in single cell methylation, using brain as an example. Bulk samples are mixtures of pure cell type profiles, while each pure cell type generate multiple cell's profile in the single-cell level.

joint profiling of two or more genomic features. For example, scM&T-seq measures single-cell genome-wide methylome and transcriptome for each cell (Angermueller et al., 2016). scNMT-seq (single-cell nucleosome, methylation and transcription sequencing) measures chromatin accessibility, DNA methylation and transcriptome jointly (Clark et al., 2018). These biotechnology improvements allow the discovery of associations across multi-omics measure, at single-cell level. Here, I will segment genome into functional regions, model the multi-omics measures, and assess the association between each pair of the measures. In addition, I will develop statistical methods for the prediction of methylation level, using other single- or multi-omics measures.

Appendix A

Appendix for Chapter 2

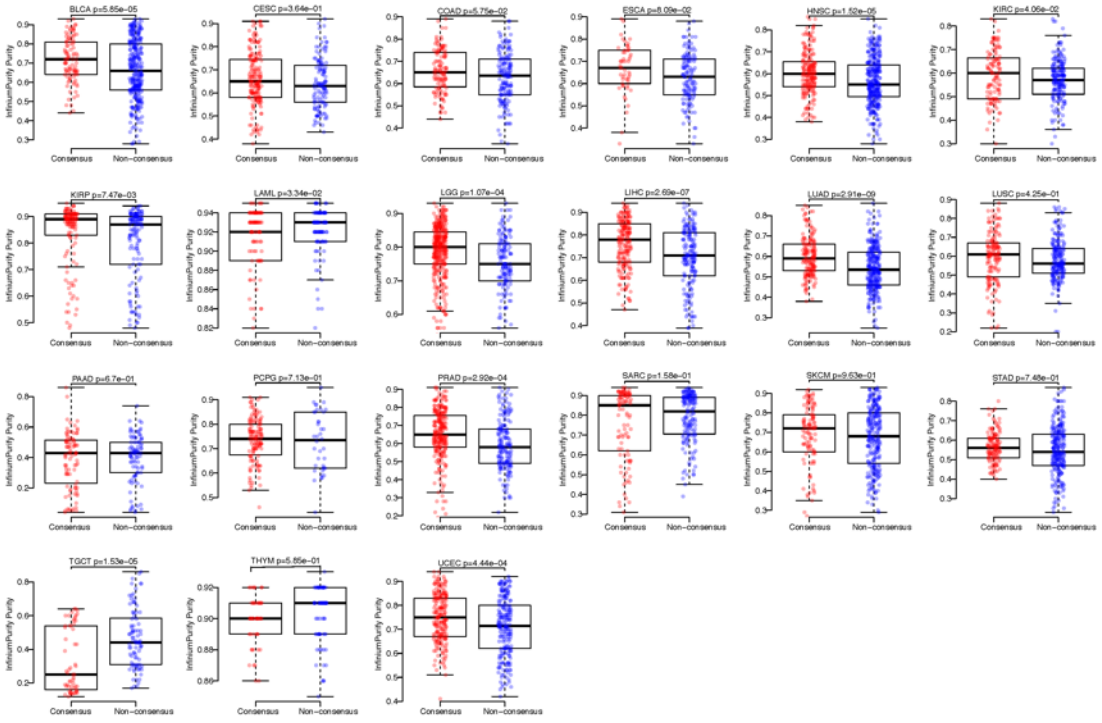


Figure A.1: InfiniumPurify Purity differences between consensus and non-consensus samples for 21 cancer types from TCGA.

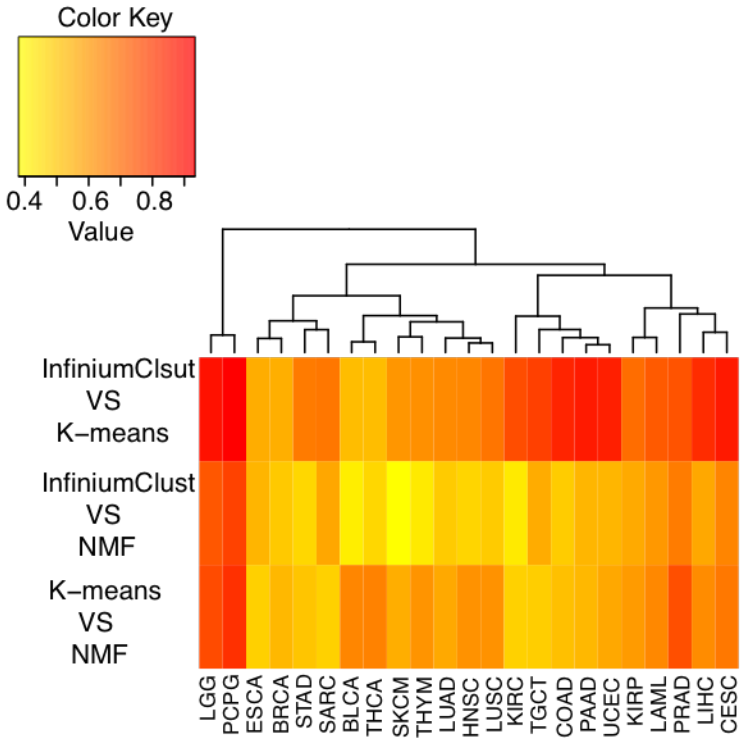


Figure A.2: Overlap matrices of clusters in the three methods by using InfiniumPurify purity in 23 cancer types. Red: highly overlapped.

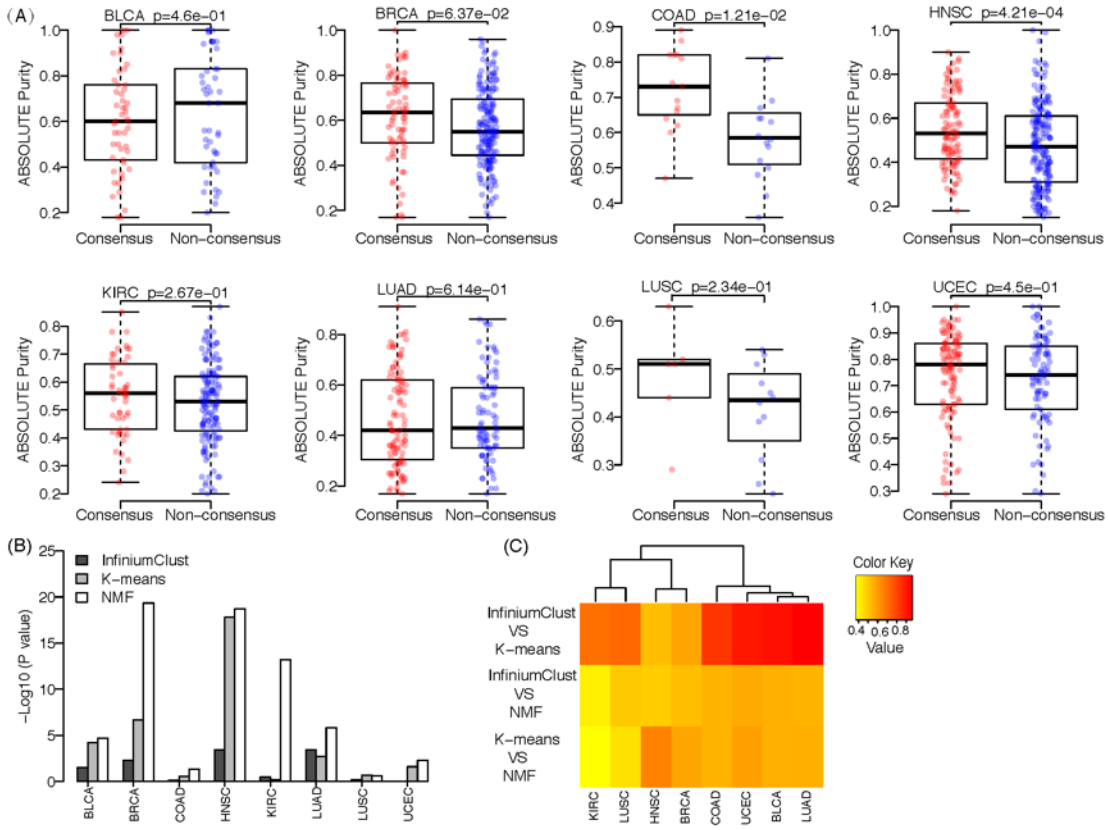


Figure A.3: Application of InfiniumClust to TCGA data by using ABSOLUTE purity. (A) ABSOLUTE Purity differences between consensus and non-consensus samples in eight cancer types of TCGA. (B) Testing ABSOLUTE purity differences among clusters from the three methods for 8 cancer types. P-values are from linear regression and F-test. (C) Overlap matrices of clusters in the three methods by using ABSOLUTE purity in 8 cancer types. Red: highly overlapped.

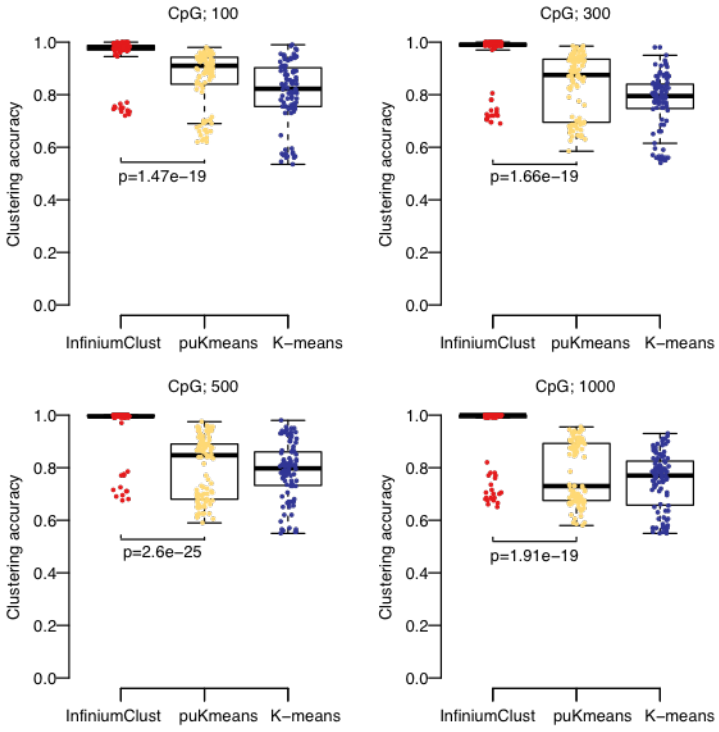


Figure A.4: Predicting accuracy in different number of CpG sites on InfiniumClust, puKmeans and K-means.

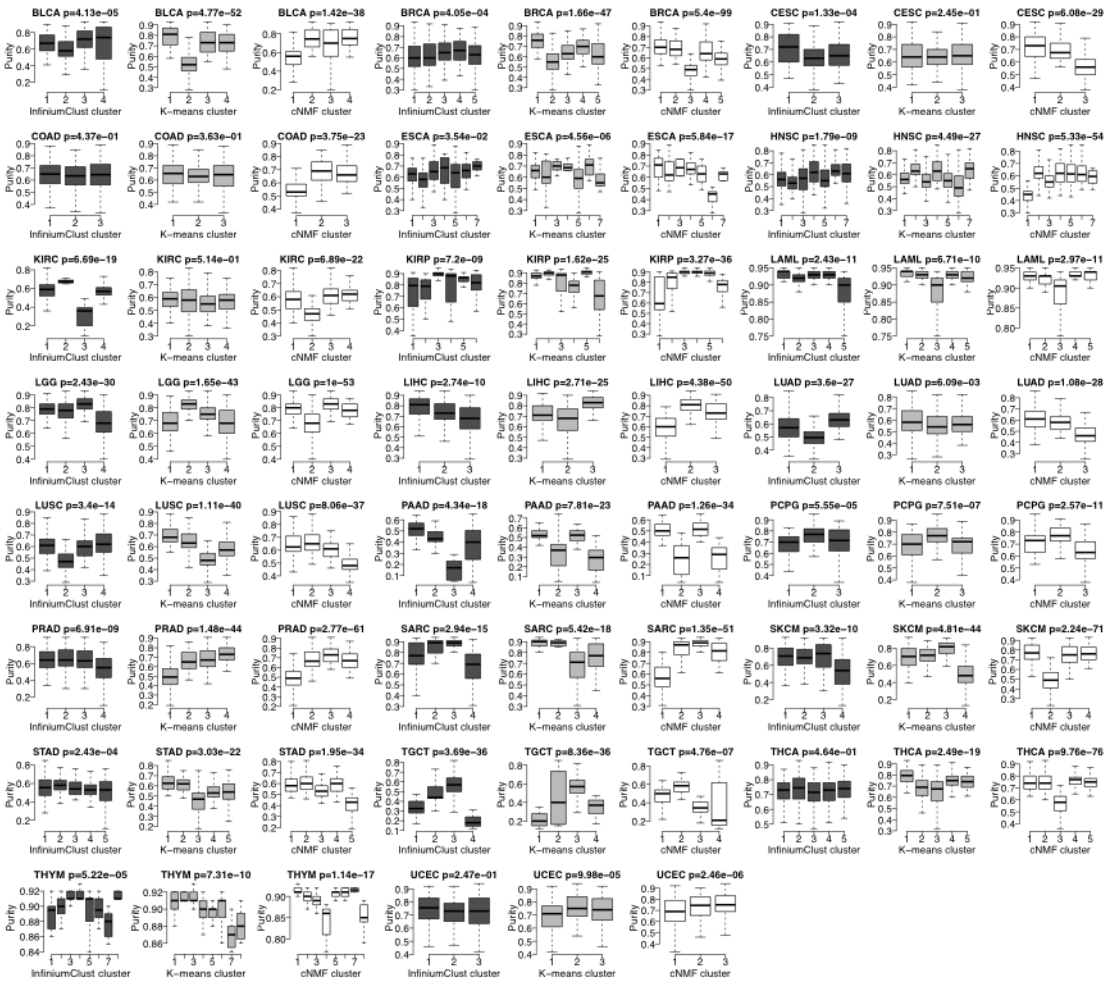


Figure A.5: InfiniumPurify Purity distributions of different clusters obtained from InfiniumClust, K-means and cNMF for 23 cancer types.

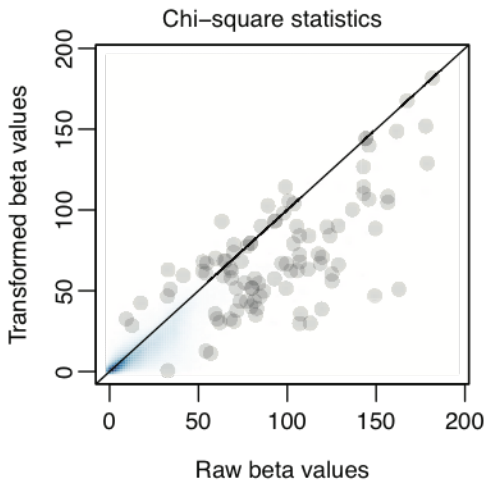
Cancer type	InfiniumClust-puKmeans	InfiniumClust-K-means	puKmeans-K-means
BLCA	0.720763723	0.603818616	0.503579952
BRCA	0.895442359	0.709115282	0.746648794
COAD	0.830564784	0.867109635	0.88372093
HNSC	0.649056604	0.518867925	0.547169811
KIRC	0.593846154	0.486153846	0.676923077
KIRP	0.688405797	0.789855072	0.605072464
LIHC	0.905263158	0.815789474	0.884210526
LUAD	0.793991416	0.579399142	0.759656652
LUSC	0.501392758	0.690807799	0.467966574
PRAD	0.524475524	0.622377622	0.727272727
THCA	0.549514563	0.805825243	0.623300971
UCEC	0.890660592	0.86332574	0.86332574

Table A.1: Overlap of clusters by using InfiniumClust, puKmeans and K-means for clustering.

Supplementary Material. The arcsine transformed beta values follow a single normal distribution from normal samples, and a mixture of normal distributions from tumor samples.

First, we demonstrate the importance of data transformation. It is known that a majority of the methylation levels are close to 0 or 1, so that the raw beta values are likely non-normal. However, after arcsine transformation, the normality of the data becomes better. We demonstrate this by comparing the goodness of fit test statistics for raw and transformed beta values from normal BRCA samples (96 samples), as shown in the figure below. Each dot represents a CpG site. The test statistics become much smaller after transformation, especially for sites with large test statistics (violation of normality). Thus the arcsine transformation helps the normality assumption and it is the reason why we apply it on raw data before clustering.

Next, we check the assumption that transformed beta values from normal samples follows a single normal distribution for a CpG site. Again, we performed goodness of fit test for normality on data from BRCA normal control samples. As a comparison, we also performed the test for data from BRCA tumor samples. The figure below shows the distribution of chi-square goodness of fit test p-values and statistics for

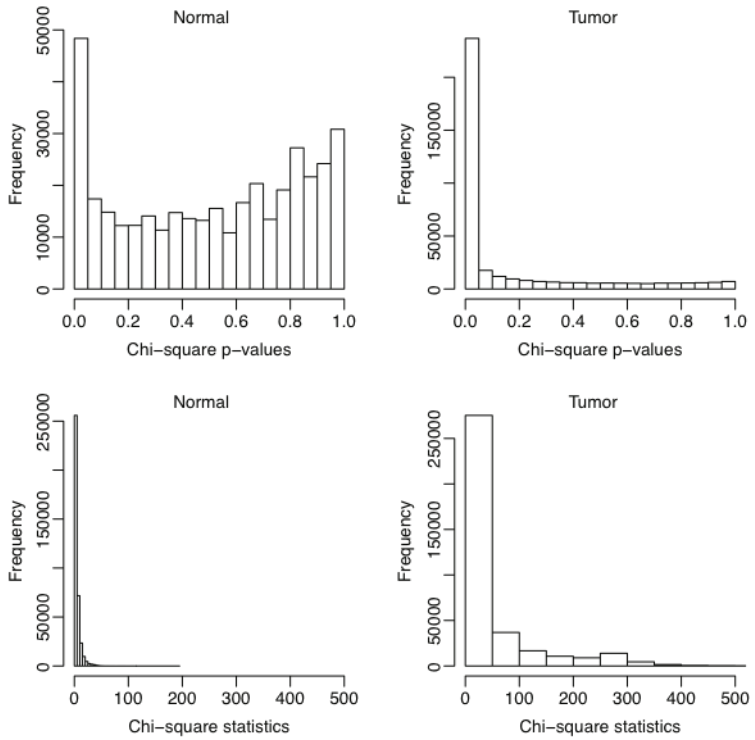


normal and tumor, for all CpG sites. A majority of the p-values from normal samples are uniformly distributed, showing that the normality assumption holds for most CpG sites. There is a small fraction of sites with small p-values, indicating violation of a single normal distribution assumption. These are likely to be the sites where different types of normal cells showing differential methylation. On the contrary, the p-values from tumor samples are highly skewed toward 0, suggesting that most of the tumor beta values are not from a single normal distribution. The histograms for test statistics lead to the same conclusion. Overall, the results show normal data approximately follow a single distribution, but tumor data are not.

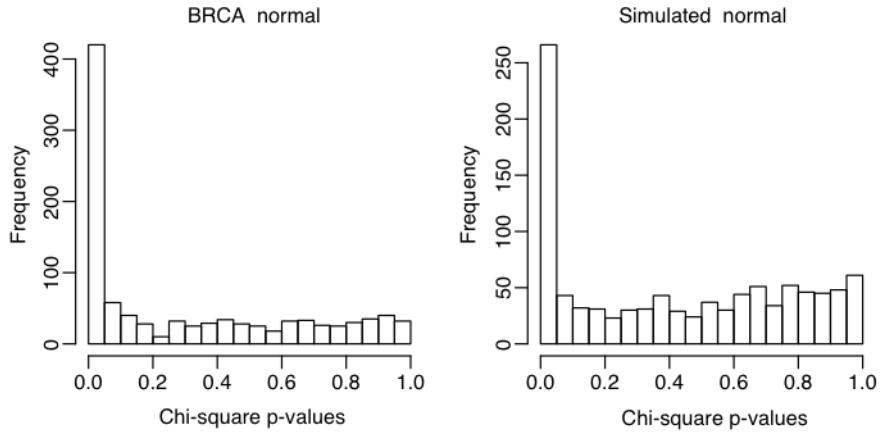
The assessment of normality in tumor data is more difficult since it is assumed to follow a mixture of normals. However, it is known that any distribution can be approximated by a mixture of normal distributions, thus we believe using a mixture of normal for tumor data will work reasonably well. Overall, these results suggest that the normality assumptions hold well in real data.

Moreover, We also performed goodness of fit test for normality on the 1000 CpG sites used in our clustering algorithm. We again compute the p-values and test statistics for these sites for both normal and tumor samples.

As shown in the figure below (left panel), the violation of normality from these CpG sites is more severe compared with the results from all CpG sites (figure provided



above): about 30% of the CpG sites violate the normality assumption. To assess how this violation affects the results, we performed additional simulation studies. We generated tumor and normal data from beta distribution and skipped the arcsine transformation. The normal assumption apparently does not hold. The right panel of the figure below shows the chi-square goodness of fit test for the most variable 1000 CpG sites, and it looks very similar to the real data. We applied InfiniumClust to the simulated data, and still obtained very high accuracy (0.98 on average). The result shows that our method is robust to the data distribution assumption.


BRCA normal
Frequency
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
1.0
Chi-square p-values
Simulated normal
Frequency
0
50
100
150
200
250
0.0
0.2
0.4
0.6
0.8
1.0
Chi-square p-values

Appendix B

Appendix for Chapter 3

Method		CIBERSORT proportion predict			RPC proportion predict			QP proportion predict		
		HCC	control	preg	HCC	control	preg	HCC	control	preg
Truth	HCC	18	9	0	17	10	0	14	13	0
	control	7	25	0	7	25	0	3	29	0
	preg	0	1	16	0	1	16	0	0	17

Table B.1: Classification confusion matrices based on real data using different reference-based methods. HCC: hepatocellular carcinoma patients. control: healthy, unpregnant control people. preg: healthy pregnant women. Cibersort: using tissue proportions solved from Cibersort for prediction. RPC: using tissue proportions solved from Robust Partial Correlations for prediction. QP: using tissue proportions solved from Quadratic Programming procedure for prediction. Cibersort predicted accuracy: 0.78. RPC predicted accuracy: 0.76. QP predicted accuracy: 0.79.

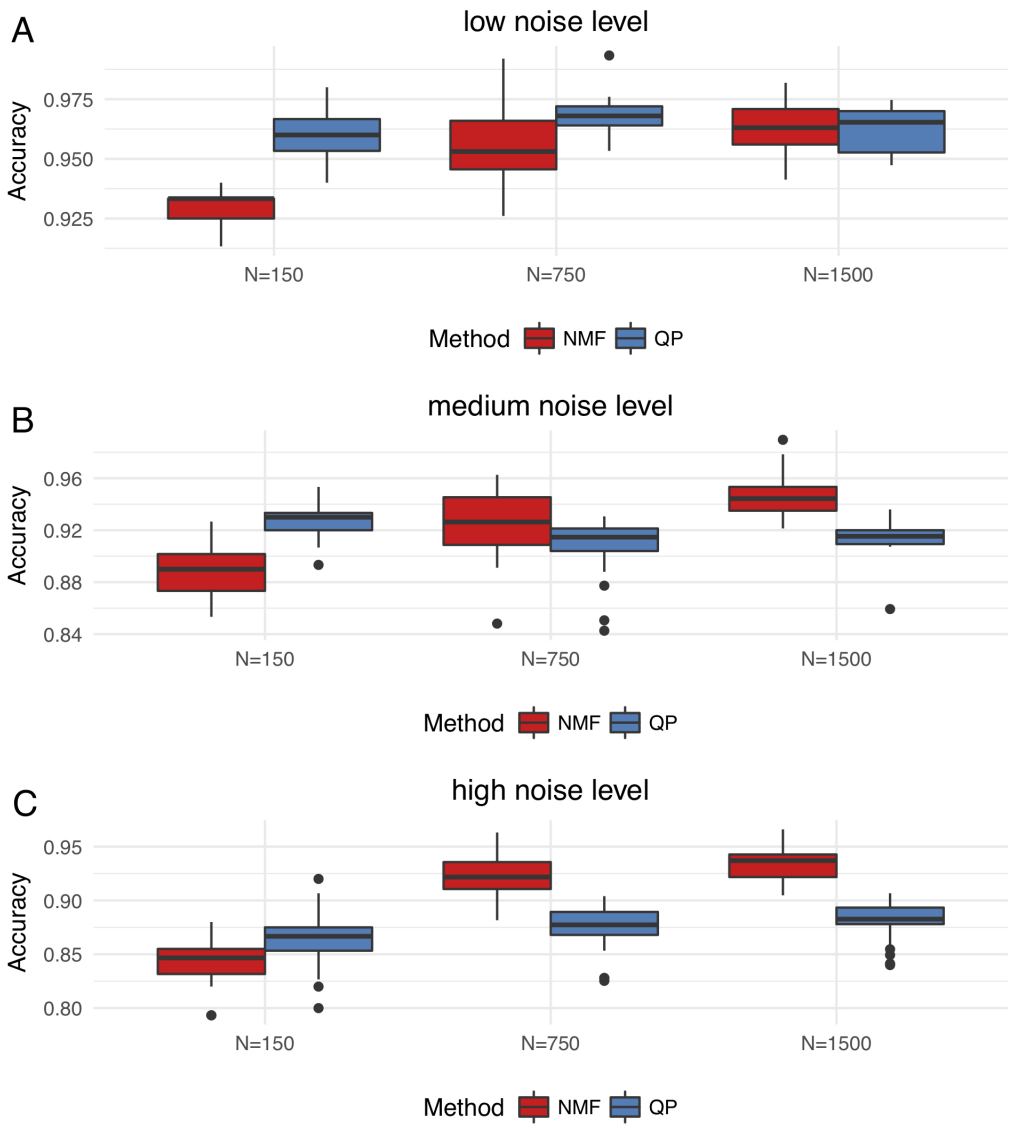


Figure B.1: Boxplot of classification accuracies for NMF and QP under different noise level and sample size. NMF: Non-negative matrix factorization (NMF) approach. QP: using tissue proportions solved from Quadratic Programming procedure for prediction. N represent the total sample size used in simulation. 3-fold cross validation was conducted. A, low noise level; B, medium noise level; C, high noise level.

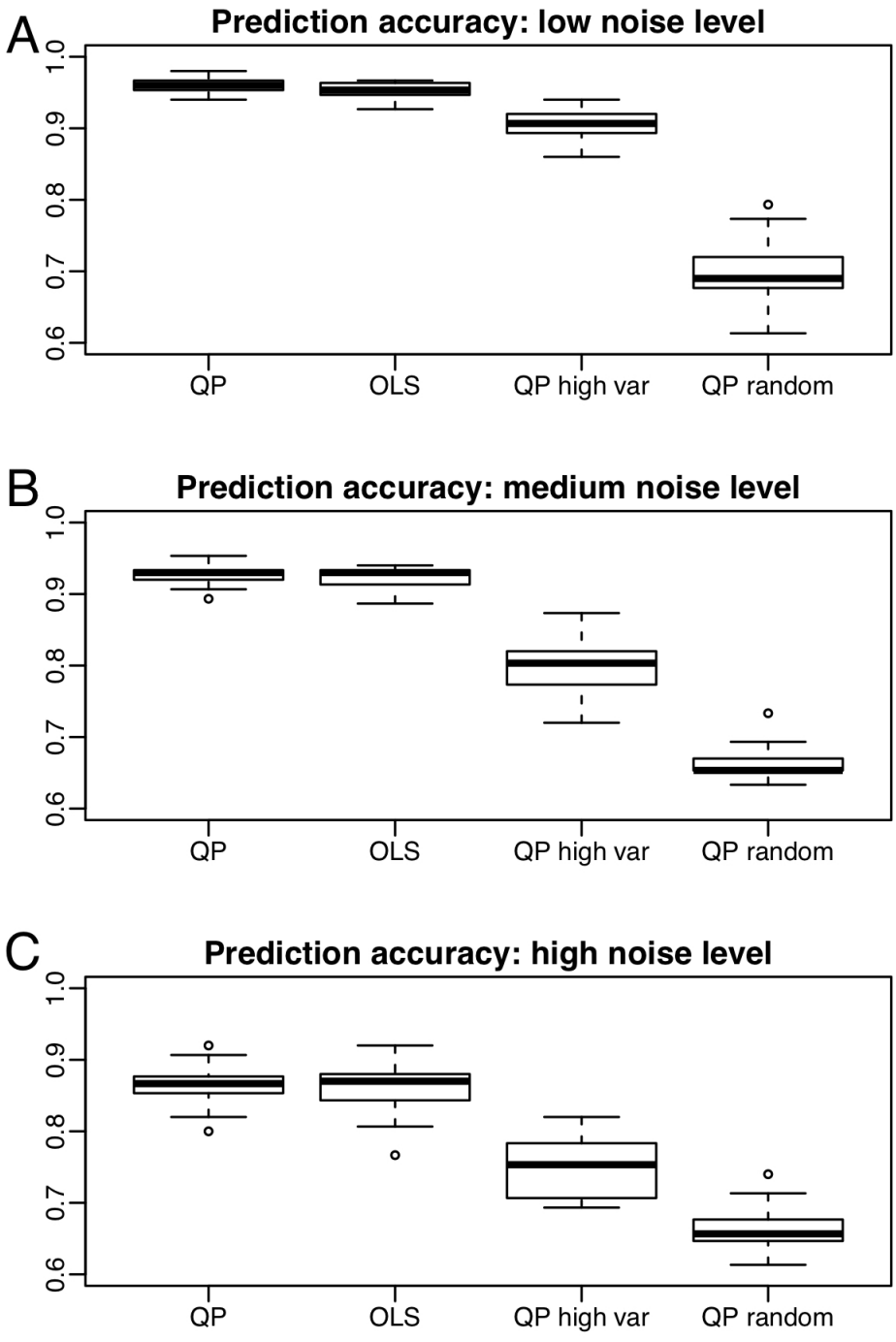


Figure B.2: Boxplot of classification accuracies for multiple methods in simulations. QP: using tissue proportions solved from Quadratic Programming procedure for prediction. OLS: using tissue proportions directly solved from Ordinary Least Square without constraint in QP. QP high var: using a high-noise reference as the reference in the QP step to solve tissue proportions, then use the solved proportions for prediction. QP random: using a randomly shuffled reference as the reference in the QP step to solve tissue proportions, then use the solved proportions for prediction. A, low noise level; B, medium noise level; C, high noise level.

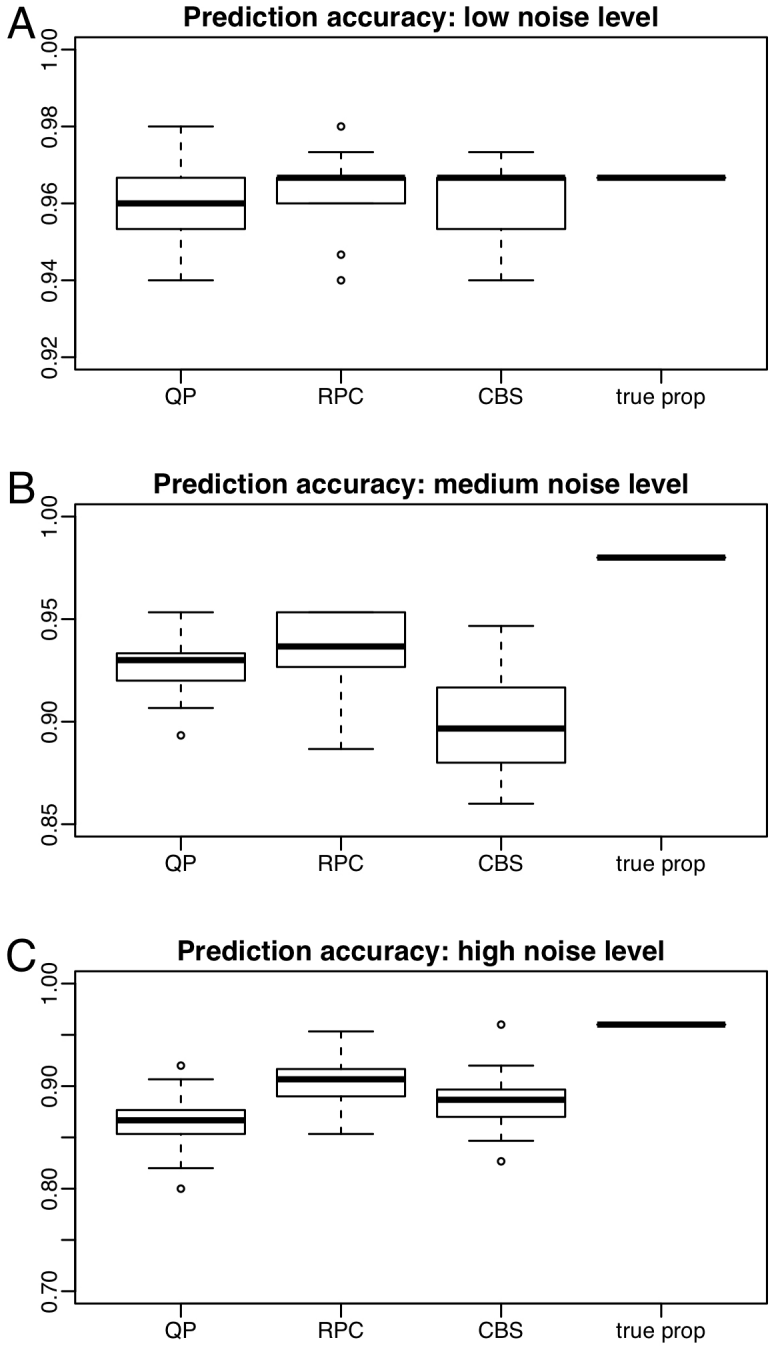


Figure B.3: Boxplot of classification accuracies for multiple reference-based methods in simulations. QP: using tissue proportions solved from Quadratic Programming procedure for prediction. RPC: using tissue proportions solved from Robust Partial Correlations for prediction. CBS: using tissue proportions solved from Cibersort for prediction. True prop: using true tissue proportions for prediction. A, low noise level; B, medium noise level; C, high noise level.

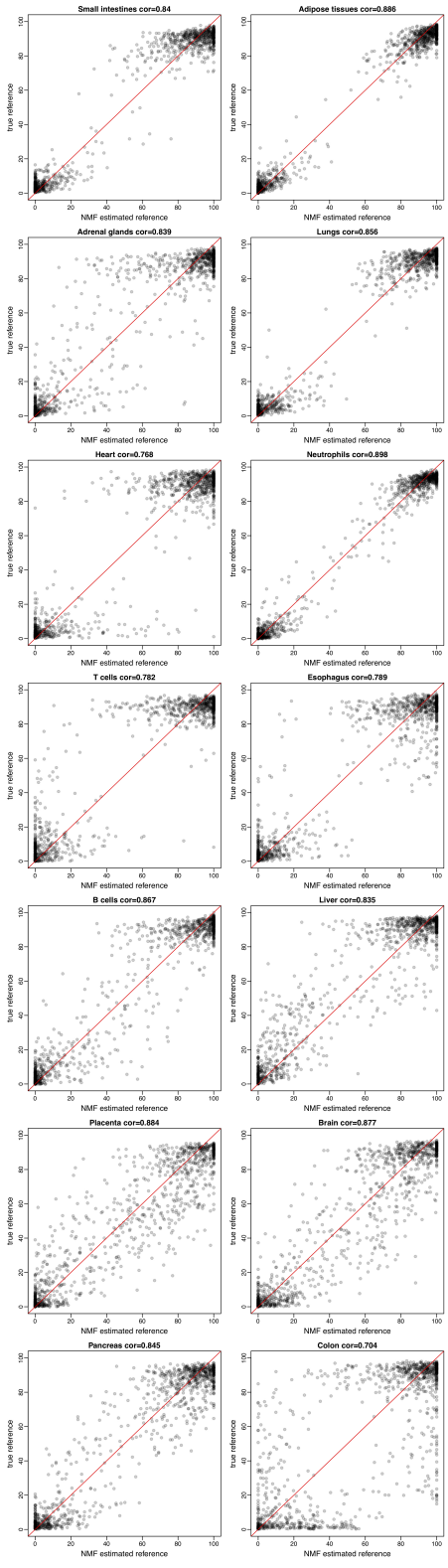


Figure B.4: Scatterplots of NMF estimated reference methylation levels versus true reference methylation levels in all 14 tissues in simulation. Spearmans correlation is shown in each panel.

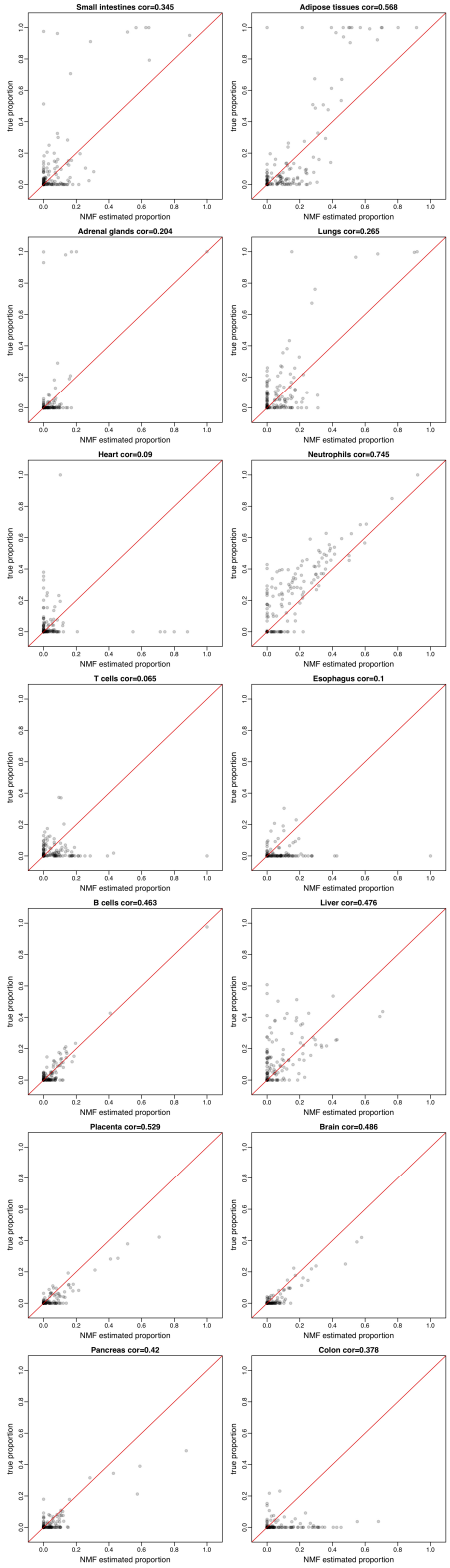


Figure B.5: Scatterplots of NMF estimated tissue proportions versus true tissue proportions in all 14 tissues in simulation. Spearmans correlation is shown in each panel.

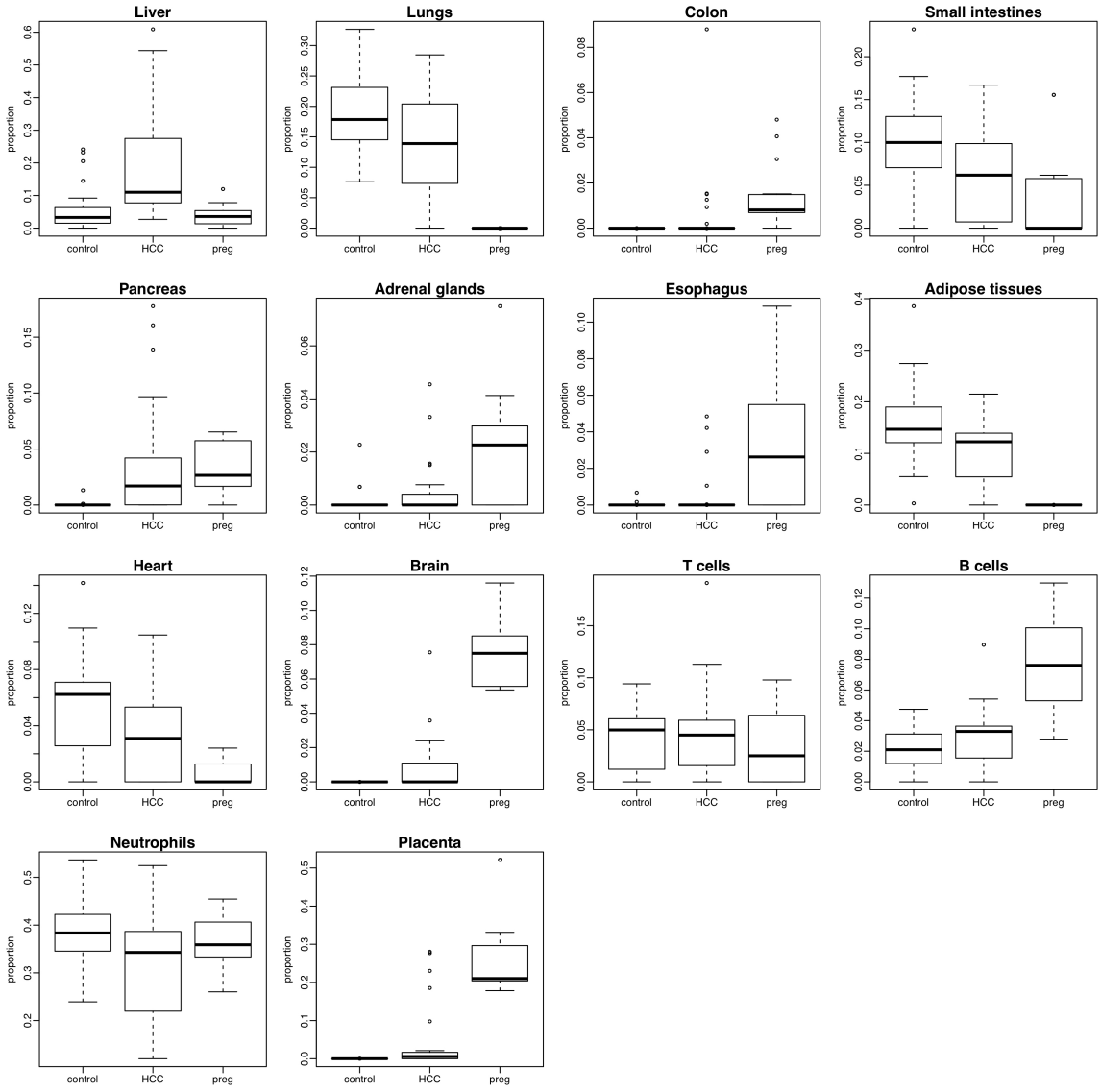


Figure B.6: Boxplot of real data solved tissue proportions for all 14 tissues, respectively, among 3 groups. One panel for each tissue.

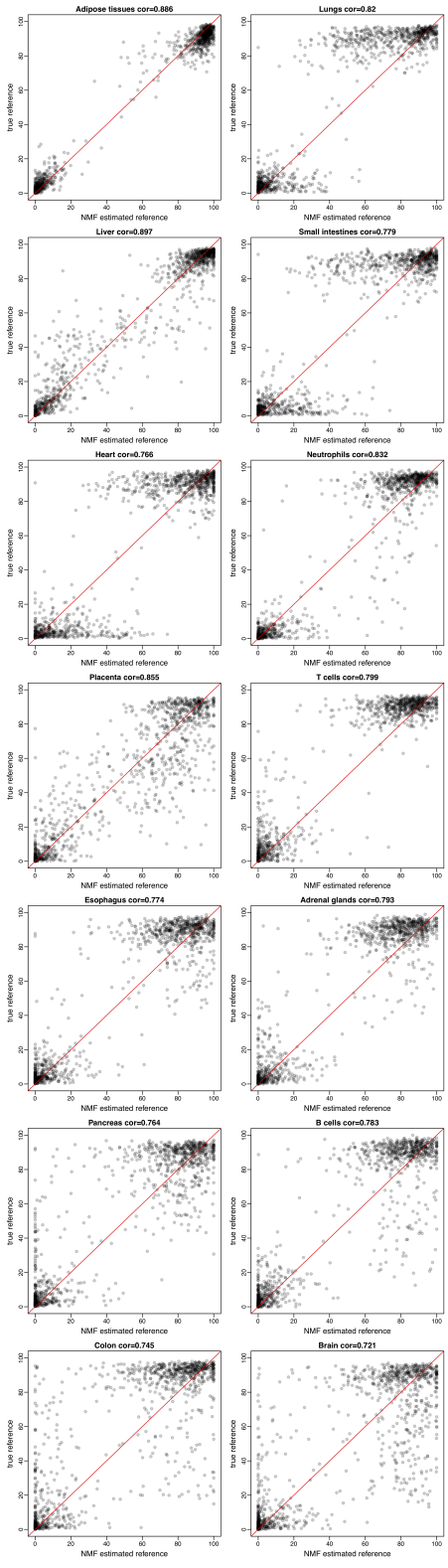


Figure B.7: Scatterplots of NMF estimated reference methylation levels versus true reference methylation levels in all 14 tissues in real data from Sun K et al. PNAS 2015. Spearmans correlation is shown in each panel.

Appendix C

Appendix for Chapter 4

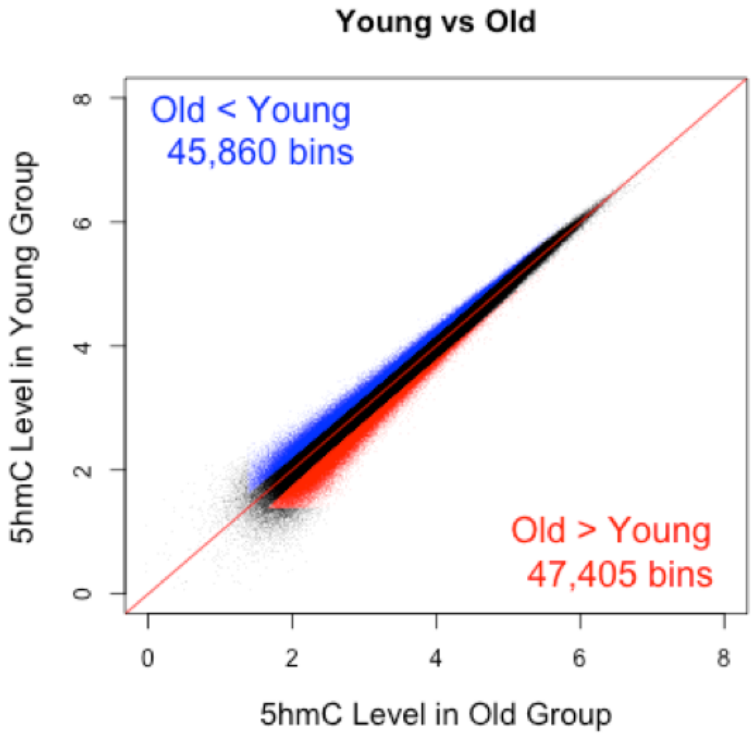


Figure C.1: Genome-scale patterns of 5hmC exhibited general increase in old samples. Global 5hmC normalized reads in cfDNA of young and old samples were counted in 5-kb binned human genome (hg19). Bins with more than 4 reads are highlighted in color. 45,860 bins (blue) showed more 5hmC reads in young than old group, whereas 47,405 bins (red) contained fewer 5hmC reads in young group.

Table C.1: Subjects information and mapping rates.

Patient ID	Age	Group	Total Reads #	Mapped Reads #	% of Alignment
NS0001	73	Old	20620940	16833720	81.63%
NS0002	70	Old	24405232	20472234	83.88%
NS0054	24	Young	27799251	22911152	82.42%
NS0003	24	Young	20694005	16933366	81.83%
NS0004	74	Old	17744068	14139450	79.69%
NS0005	27	Young	21343938	17627700	82.59%
NS0006	27	Young	24002467	19794803	82.47%
NS0007	23	Young	27176486	22001543	80.96%
NS0008	70	Old	16184975	12940173	79.95%
NS0009	24	Young	27427366	22348100	81.48%
NS0010	30	Young	27218638	22500079	82.66%
NS0011	30	Young	27074327	22618490	83.54%
NS0012	73	Old	25021058	20956187	83.75%
NS0013	71	Old	25340923	20793804	82.06%
NS0014	24	Young	24435506	19899891	81.44%
NS0015	23	Young	25601779	21286193	83.14%
NS0016	76	Old	21321136	17330107	81.28%
NS0017	69	Old	26895623	22107968	82.20%
NS0018	76	Old	25874331	21178665	81.85%
NS0019	68	Old	22978841	19111322	83.17%
885296	85	AD	21697210	17884224	82.43%
710785	74	AD	20847691	17477643	83.83%
710642	84	AD	21060631	17296572	82.13%
710969	67	AD	23241602	18595340	80.01%
885862	90	AD	20199690	16690537	82.63%
711904	63	AD	18749790	15504879	82.69%
713048	69	AD	25235132	20877621	82.73%
710978	81	AD	24167035	19494196	80.66%
886906	83	AD	21102684	17247851	81.73%
710286	80	AD	24351773	19818729	81.39%

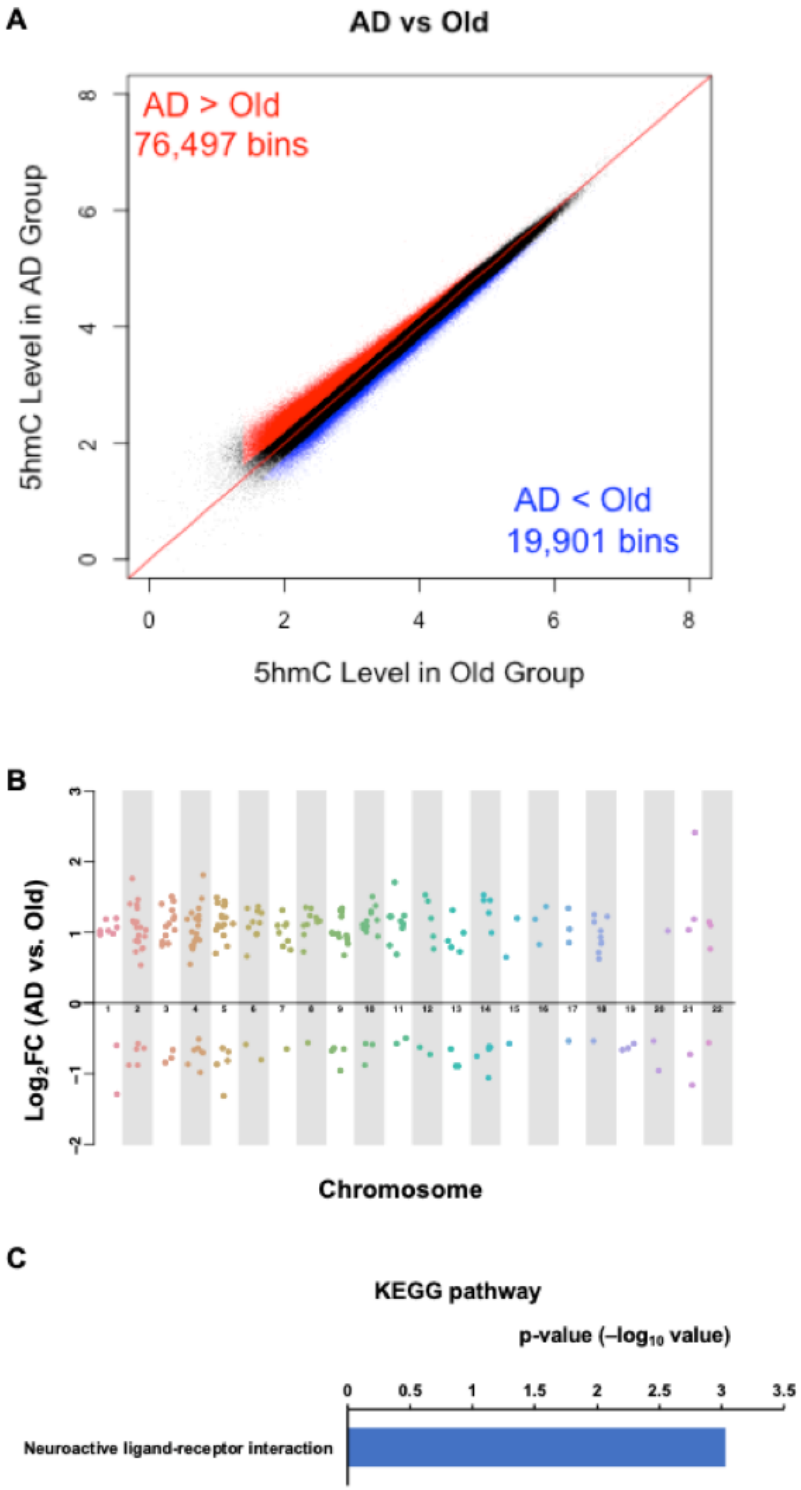


Figure C.2: Genome-scale patterns of 5hmC exhibited general increase in AD samples. (**A**) Global 5hmC normalized reads in cfDNA of AD and old samples were counted in 5-kb binned human genome (hg19). Bins with more than 4 reads are highlighted in color. 76,497 bins (red) showed more 5hmC reads in AD than old group, whereas 19,901 bins (blue) contained fewer 5hmC reads in AD group. (**B**) Chromosomal distribution of 236 disease-associated DhMRs indicates they are relatively universally located in autosome. (**C**) Pathway analyses showed that the nearest genes of disease-associated DhMRs are significantly involved in Neuroactive ligand-receptor interaction.

Bibliography

Ahn, J., Yuan, Y., Parmigiani, G., Suraokar, M. B., Diao, L., Wistuba, I. I. and Wang, W. (2013), 'Demix: deconvolution for mixed cancer transcriptomes using raw measured data', Bioinformatics **29**(15), 1865–1871.

AJ, J. A., Colburn, W. A., DeGruttola, V. G., DeMets, D. L., Downing, G. J., Hoth, D. F., Oates, J. A., Peck, C. C., Schooley, R. T., Spilker, B. A. et al. (2001), 'Biomarkers and surrogate endpoints: preferred definitions and conceptual framework', Clinical pharmacology and therapeutics **69**(3), 89–95.

Alisch, R. S., Barwick, B. G., Chopra, P., Myrick, L. K., Satten, G. A., Conneely, K. N. and Warren, S. T. (2012), 'Age-associated dna methylation in pediatric populations', Genome research **22**(4), 623–632.

Allen, M., Heinzmann, A., Noguchi, E., Abecasis, G., Broxholme, J., Ponting, C. P., Bhattacharyya, S., Tinsley, J., Zhang, Y., Holt, R. et al. (2003), 'Positional cloning of a novel gene influencing asthma from chromosome 2q14', Nature genetics **35**(3), 258.

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., Voet, T. et al. (2016), 'Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity', Nature methods **13**(3), 229.

Aran, D., Sirota, M. and Butte, A. J. (2015), 'Systematic pan-cancer analysis of tumour purity', Nature communications **6**, 8971.

Avraham, A., Cho, S. S., Uhlmann, R., Polak, M. L., Sandbank, J., Karni, T., Pappo, I., Halperin, R., Vaknin, Z., Sella, A. et al. (2014), 'Tissue specific dna methylation in normal human breast epithelium and in breast cancer', PloS one **9**(3), e91805.

Bao, L., Pu, M. and Messer, K. (2014), 'Abscn-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data', Bioinformatics **30**(8), 1056–1063.

Bekris, L. M., Yu, C.-E., Bird, T. D. and Tsuang, D. W. (2010), 'Genetics of alzheimer disease', Journal of geriatric psychiatry and neurology **23**(4), 213–227.

Bell, G. I. and Polonsky, K. S. (2001), 'Diabetes mellitus and genetically programmed defects in β-cell function', Nature **414**(6865), 788.

Bell, J. (2004), 'Predicting disease using genomics', Nature **429**(6990), 453.

Bell, K. F. and Hardingham, G. E. (2011), 'The influence of synaptic activity on neuronal health', Current opinion in neurobiology **21**(2), 299–305.

Berger, S. L., Kouzarides, T., Shiekhattar, R. and Shilatifard, A. (2009), 'An operational definition of epigenetics', Genes & development **23**(7), 781–783.

Bird, A. (2002), 'Dna methylation patterns and epigenetic memory', Genes & development **16**(1), 6–21.

Bird, A. P. and Wolffe, A. P. (1999), 'Methylation-induced repressionbelts, braces, and chromatin', Cell **99**(5), 451–454.

Blennow, K. (n.d.), 'Leon mj de, zetterberg h (2006) alzheimers disease', Lancet **368**, 387–403.

Bloushtain-Qimron, N., Yao, J., Snyder, E. L., Shipitsin, M., Campbell, L. L., Mani, S. A., Hu, M., Chen, H., Ustyansky, V., Antosiewicz, J. E. et al. (2008), 'Cell type-specific dna methylation patterns in the human breast', Proceedings of the National Academy of Sciences **105**(37), 14076–14081.

Bradley-Whitman, M. and Lovell, M. (2013), 'Epigenetic changes in the progression of alzheimer's disease', Mechanisms of ageing and development **134**(10), 486–495.

Brindle, J. T., Antti, H., Holmes, E., Tranter, G., Nicholson, J. K., Bethell, H. W., Clarke, S., Schofield, P. M., McKilligin, E., Mosedale, D. E. et al. (2002), 'Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1 h-nmr-based metabonomics', Nature medicine **8**(12), 1439.

Brunet, J.-P., Tamayo, P., Golub, T. R. and Mesirov, J. P. (2004), 'Metagenes and molecular pattern discovery using matrix factorization', Proceedings of the national academy of sciences **101**(12), 4164–4169.

Cardenas, A., Allard, C., Doyon, M., Houseman, E. A., Bakulski, K. M., Perron, P., Bouchard, L. and Hivert, M.-F. (2016), 'Validation of a dna methylation reference panel for the estimation of nucleated cells types in cord blood', Epigenetics **11**(11), 773–779.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A. et al. (2012), 'Absolute quantification of somatic dna alterations in human cancer', Nature biotechnology **30**(5), 413.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R. and Maayan, A. (2013), 'Enrichr: interactive and collaborative html5 gene list enrichment analysis tool', BMC bioinformatics **14**(1), 128.

Chen, H., Dzitoyeva, S. and Manev, H. (2012), 'Effect of aging on 5-

hydroxymethylcytosine in the mouse hippocampus', Restorative neurology and neuroscience **30**(3), 237–245.

Chen, Y., Damayanti, N., Irudayaraj, J., Dunn, K. and Zhou, F. C. (2014), 'Diversity of two forms of dna methylation in the brain', Frontiers in genetics **5**, 46.

Cheng, Y., Sun, M., Chen, L., Li, Y., Lin, L., Yao, B., Li, Z., Wang, Z., Chen, J., Miao, Z. et al. (2018), 'Ten-eleven translocation proteins modulate the response to environmental stress in mice', Cell reports **25**(11), 3194–3203.

Christensen, B. C., Houseman, E. A., Marsit, C. J., Zheng, S., Wrensch, M. R., Wiemels, J. L., Nelson, H. H., Karagas, M. R., Padbury, J. F., Bueno, R. et al. (2009), 'Aging and environmental exposures alter tissue-specific dna methylation dependent upon cpg island context', PLoS genetics **5**(8), e1000602.

Chung, C. H., Bernard, P. S. and Perou, C. M. (2002), 'Molecular portraits and the family tree of cancer', Nature genetics **32**, 533.

Cichocki, A., Zdunek, R. and Amari, S.-i. (2006), New algorithms for non-negative matrix factorization in applications to blind source separation, in 'Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on', Vol. 5, IEEE, pp. V–V.

Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C. et al. (2018), 'scnmt-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells', Nature communications **9**(1), 781.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M. and Jacobsen, S. E. (2008), 'Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning', Nature **452**(7184), 215.

Coppieters, N., Dieriks, B. V., Lill, C., Faull, R. L., Curtis, M. A. and Dragunow, M. (2014), 'Global changes in dna methylation and hydroxymethylation in alzheimer's disease human brain', Neurobiology of aging **35**(6), 1334–1344.

Crowley, E., Di Nicolantonio, F., Loupakis, F. and Bardelli, A. (2013), 'Liquid biopsy: monitoring cancer-genetics in the blood', Nature reviews Clinical oncology **10**(8), 472.

Das, P. M. and Singal, R. (2004), 'Dna methylation and cancer', Journal of clinical oncology **22**(22), 4632–4642.

Dupont, C., Armant, D. R. and Brenner, C. A. (2009), Epigenetics: definition, mechanisms and clinical perspective, in 'Seminars in reproductive medicine', Vol. 27, © Thieme Medical Publishers, pp. 351–357.

Feng, H., Conneely, K. N. and Wu, H. (2014), 'A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data', Nucleic acids research **42**(8), e69–e69.

Feng, H., Jin, P. and Wu, H. (2018), 'Disease prediction by cell-free dna methylation', Briefings in bioinformatics .

Feyzi, E., Saldeen, T., Larsson, E., Lindahl, U. and Salmivirta, M. (1998), 'Age-dependent modulation of heparan sulfate structure and function', Journal of Biological Chemistry **273**(22), 13395–13398.

Ficz, G., Branco, M. R., Seisenberger, S., Santos, F., Krueger, F., Hore, T. A., Marques, C. J., Andrews, S. and Reik, W. (2011), 'Dynamic regulation of 5-hydroxymethylcytosine in mouse es cells and during differentiation', Nature **473**(7347), 398.

Fraga, M. F. and Esteller, M. (2007), 'Epigenetics and aging: the targets and the marks', Trends in Genetics **23**(8), 413–418.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L. and Paul, C. L. (1992), '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.

Gao, H.-T., Li, T.-H., Chen, K., Li, W.-G. and Bi, X. (2005), 'Overlapping spectra resolution using non-negative matrix factorization', Talanta **66**(1), 65–73.

Ghosh, S., Yates, A. J., Frühwald, M. C., Miecznikowski, J. C., Plass, C. and Smiraglia, D. (2010), 'Tissue specific dna methylation of cpg islands in normal human adult somatic tissues distinguishes neural from non-neural tissues', Epigenetics **5**(6), 527–538.

Globisch, D., Münzel, M., Müller, M., Michalakis, S., Wagner, M., Koch, S., Brückl, T., Biel, M. and Carell, T. (2010), 'Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates', PloS one **5**(12), e15367.

Gloyn, A. L., Weedon, M. N., Owen, K. R., Turner, M. J., Knight, B. A., Hitman, G., Walker, M., Levy, J. C., Sampson, M., Halford, S. et al. (2003), 'Large-scale association studies of variants in genes encoding the pancreatic β-cell katp channel subunits kir6. 2 (kcnj11) and sur1 (abcc8) confirm that the kcnj11 e23k variant is associated with type 2 diabetes', Diabetes **52**(2), 568–572.

Guo, S., Diep, D., Plongthongkum, N., Fung, H.-L., Zhang, K. and Zhang, K. (2017), 'Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma dna', Nature genetics **49**(4), 635.

Hackett, J. A. and Surani, M. A. (2013), 'Dna methylation dynamics during the mammalian life cycle', Phil. Trans. R. Soc. B **368**(1609), 20110328.

Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D. et al. (2011), 'Increased methylation variation in epigenetic domains across cancer types', Nature genetics **43**(8), 768.

Hao, X., Luo, H., Krawczyk, M., Wei, W., Wang, W., Wang, J., Flagg, K., Hou, J., Zhang, H., Yi, S. et al. (2017), 'Dna methylation markers for diagnosis and prognosis of common cancers', Proceedings of the National Academy of Sciences **114**(28), 7414–7419.

Harris, R. A., Wang, T., Coarfa, C., Nagarajan, R. P., Hong, C., Downey, S. L., Johnson, B. E., Fouse, S. D., Delaney, A., Zhao, Y. et al. (2010), 'Comparison of sequencing-based methods to profile dna methylation and identification of monoallelic epigenetic modifications', Nature biotechnology **28**(10), 1097.

Hatt, L., Aagaard, M. M., Graakjaer, J., Bach, C., Sommer, S., Agerholm, I. E., Kølvraa, S. and Bojesen, A. (2015), 'Microarray-based analysis of methylation status of cpgs in placental dna and maternal blood dna–potential new epigenetic biomarkers for cell free fetal dna-based diagnosis', PloS one **10**(7), e0128918.

He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L. et al. (2011), 'Tet-mediated formation of 5-carboxylcytosine and its excision by tdg in mammalian dna', Science **333**(6047), 1303–1307.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. and Glass, C. K. (2010), 'Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities', Molecular cell **38**(4), 576–589.

Hendrich, B. and Bird, A. (1998), 'Identification and characterization of a family of mammalian methyl-cpg binding proteins', Molecular and cellular biology **18**(11), 6538–6547.

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D., Niu, B., McLellan, M. D., Uzunangelov, V. et al. (2014), 'Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin', Cell **158**(4), 929–944.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K. and Kelsey, K. T. (2012), 'Dna methylation arrays as surrogate measures of cell mixture distribution', BMC bioinformatics **13**(1), 86.

Houseman, E. A., Christensen, B. C., Yeh, R.-F., Marsit, C. J., Karagas, M. R., Wrensch, M., Nelson, H. H., Wiemels, J., Zheng, S., Wiencke, J. K. et al. (2008), 'Model-based clustering of dna methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions', BMC bioinformatics **9**(1), 365.

Houseman, E. A., Kile, M. L., Christiani, D. C., Ince, T. A., Kelsey, K. T. and Marsit, C. J. (2016), 'Reference-free deconvolution of dna methylation data and mediation by cell composition effects', BMC bioinformatics **17**(1), 259.

Irier, H. A. and Jin, P. (2012), 'Dynamics of dna methylation in aging and alzheimer's disease', DNA and cell biology **31**(S1), S–42.

Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. A., He, C. and Zhang, Y. (2011), 'Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine', Science **333**(6047), 1300–1303.

Jensen, T. J., Kim, S. K., Zhu, Z., Chin, C., Gebhard, C., Lu, T., Deciu, C., van den Boom, D. and Ehrich, M. (2015), 'Whole genome bisulfite sequencing of cell-free dna and its cellular contributors uncovers placenta hypomethylated domains', Genome biology **16**(1), 78.

Johnson, W. E., Li, C. and Rabinovic, A. (2007), 'Adjusting batch effects in microarray expression data using empirical bayes methods', Biostatistics **8**(1), 118–127.

Jung, M. and Pfeifer, G. P. (2015), 'Aging and dna methylation', BMC biology **13**(1), 7.

Kang, S., Li, Q., Chen, Q., Zhou, Y., Park, S., Lee, G., Grimes, B., Krysan, K., Yu, M., Wang, W. et al. (2017), 'Cancerlocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free dna', Genome biology **18**(1), 53.

Khare, T., Pai, S., Koncevicius, K., Pal, M., Kriukiene, E., Liutkeviciute, Z., Irimia, M., Jia, P., Ptak, C., Xia, M. et al. (2012), '5-hmc in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary', Nature structural & molecular biology **19**(10), 1037.

Kriaucionis, S. and Heintz, N. (2009), 'The nuclear dna base 5-hydroxymethylcytosine is present in purkinje neurons and the brain', Science **324**(5929), 929–930.

Kuan, P. F., Wang, S., Zhou, X. and Chu, H. (2010), 'A statistical framework for illumina dna methylation arrays', Bioinformatics **26**(22), 2849–2855.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A. et al. (2016), 'Enrichr: a comprehensive gene set enrichment analysis web server 2016 update', Nucleic acids research **44**(W1), W90–W97.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J. et al. (2015), 'Integrative analysis of 111 reference human epigenomes', Nature **518**(7539), 317.

Legendre, C., Gooden, G. C., Johnson, K., Martinez, R. A., Liang, W. S. and Salhia, B. (2015), 'Whole-genome bisulfite sequencing of cell-free dna identifies signature associated with metastatic breast cancer', Clinical epigenetics **7**(1), 100.

Lehmann-Werman, R., Neiman, D., Zemmour, H., Moss, J., Magenheim, J., Vaknin-Dembinsky, A., Rubertsson, S., Nellgård, B., Blennow, K., Zetterberg, H. et al. (2016), 'Identification of tissue-specific cell death using methylation patterns of circulating dna', Proceedings of the National Academy of Sciences **113**(13), E1826–E1834.

Leong, D., Rai, R., Nguyen, B., Lee, A. and Yip, D. (2014), 'Advances in adjuvant systemic therapy for non-small-cell lung cancer', World journal of clinical oncology **5**(4), 633.

Li, B. and Yu, Q. (2008), 'Classification of functional data: A segmentation approach', Computational Statistics & Data Analysis **52**(10), 4790–4800.

Li, E., Beard, C. and Jaenisch, R. (1993), 'Role for dna methylation in genomic imprinting', Nature **366**(6453), 362.

Li, W., Zhang, X., Lu, X., You, L., Song, Y., Luo, Z., Zhang, J., Nie, J., Zheng, W., Xu, D. et al. (2017), '5-hydroxymethylcytosine signatures in circulating cell-free dna as diagnostic biomarkers for human cancers', Cell research **27**(10), 1243.

Li, X., Yao, B., Chen, L., Kang, Y., Li, Y., Cheng, Y., Li, L., Lin, L., Wang, Z., Wang, M. et al. (2017), 'Ten-eleven translocation 2 interacts with forkhead box o3 and regulates adult neurogenesis', Nature communications **8**, 15903.

Li, Y. E., Xiao, M., Shi, B., Yang, Y.-C. T., Wang, D., Wang, F., Marcia, M. and Lu, Z. J. (2017), 'Identification of high-confidence rna regulatory elements by combinatorial classification of rna–protein binding sites', Genome biology **18**(1), 169.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H. and Ecker, J. R. (2008), 'Highly integrated single-base resolution maps of the epigenome in arabidopsis', Cell **133**(3), 523–536.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M. et al. (2009), 'Human dna methylomes at base resolution show widespread epigenomic differences', nature **462**(7271), 315.

Lokk, K., Modhukur, V., Rajashekar, B., Märtens, K., Mägi, R., Kolde, R., Koltšina, M., Nilsson, T. K., Vilo, J., Salumets, A. et al. (2014), 'Dna methylome profiling of human tissues identifies global and tissue-specific methylation patterns', Genome biology **15**(4), 3248.

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. and Kroemer, G. (2013), 'The hallmarks of aging', Cell **153**(6), 1194–1217.

Lutsik, P., Slawski, M., Gasparoni, G., Vedeneev, N., Hein, M. and Walter, J. (2017), 'Medecom: discovery and quantification of latent components of heterogeneous methylomes', Genome biology **18**(1), 55.

MacArthur, D., Manolio, T., Dimmock, D., Rehm, H., Shendure, J., Abecasis, G., Adams, D., Altman, R., Antonarakis, S., Ashley, E. et al. (2014), 'Guidelines for investigating causality of sequence variants in human disease', Nature **508**(7497), 469.

Maiti, A. and Drohat, A. C. (2011), 'Thymine dna glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine potential implications for active demethylation of cpg sites', Journal of Biological Chemistry **286**(41), 35334–35338.

Mastroeni, D., Grover, A., Delvaux, E., Whiteside, C., Coleman, P. D. and Rogers, J. (2010), 'Epigenetic changes in alzheimer's disease: decrements in dna methylation', Neurobiology of aging **31**(12), 2025–2037.

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B. et al. (2008), 'Genome-scale dna methylation maps of pluripotent and differentiated cells', Nature **454**(7205), 766.

Meissner, C. and Ritz-Timme, S. (2010), 'Molecular pathology and age estimation', Forensic science international **203**(1-3), 34–43.

Meng, X.-L. and Rubin, D. B. (1993), 'Maximum likelihood estimation via the ecm algorithm: A general framework', Biometrika **80**(2), 267–278.

Morrison, J. H. and Baxter, M. G. (2012), 'The ageing cortical synapse: hallmarks and implications for cognitive decline', Nature Reviews Neuroscience **13**(4), 240.

Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P., Tong, F. et al. (2006), 'A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes', Cancer cell **10**(6), 515–527.

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M. and Alizadeh, A. A. (2015), 'Robust enumeration of cell subsets from tissue expression profiles', Nature methods **12**(5), 453.

Nicholson, D. A., Yoshida, R., Berry, R. W., Gallagher, M. and Geinisman, Y. (2004), 'Reduction in size of perforated postsynaptic densities in hippocampal axospinous synapses and age-related spatial learning impairments', Journal of Neuroscience **24**(35), 7648–7653.

Nordlund, J., Bäcklin, C. L., Zachariadis, V., Cavelier, L., Dahlberg, J., Öfverholm, I., Barbany, G., Nordgren, A., Övernäs, E., Abrahamsson, J. et al. (2015), 'Dna methylation-based subtype prediction for pediatric acute lymphoblastic leukemia', Clinical epigenetics **7**(1), 11.

Ogino, S., Fuchs, C. S. and Giovannucci, E. (2012), 'How many molecular subtypes? implications of the unique tumor principle in personalized medicine', Expert review of molecular diagnostics **12**(6), 621–628.

Onuchic, V., Hartmaier, R. J., Boone, D. N., Samuels, M. L., Patel, R. Y., White, W. M., Garovic, V. D., Oesterreich, S., Roth, M. E., Lee, A. V. et al. (2016), 'Epigenomic deconvolution of breast tumors reveals metabolic coupling between constituent cell types', Cell reports **17**(8), 2075–2086.

Park, Y. and Wu, H. (2016), 'Differential methylation analysis for bs-seq data under general experimental design', Bioinformatics **32**(10), 1446–1453.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z. et al. (2009), 'Supervised risk predictor of breast cancer based on intrinsic subtypes', Journal of clinical oncology **27**(8), 1160–1167.

Pastor, W. A., Pape, U. J., Huang, Y., Henderson, H. R., Lister, R., Ko, M., McLoughlin, E. M., Brudno, Y., Mahapatra, S., Kapranov, P. et al. (2011), 'Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells', Nature **473**(7347), 394.

Paynter, N. P., Chasman, D. I., Buring, J. E., Shiffman, D., Cook, N. R. and Ridker, P. M. (2009), 'Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21. 3', Annals of internal medicine **150**(2), 65–72.

Petricoin III, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C. et al. (2002), 'Use of proteomic patterns in serum to identify ovarian cancer', The lancet **359**(9306), 572–577.

Qi, L., Chen, L., Li, Y., Qin, Y., Pan, R., Zhao, W., Gu, Y., Wang, H., Wang, R., Chen, X. et al. (2015), 'Critical limitations of prognostic signatures based on risk scores summarized from gene expression levels: a case study for resected stage i non-small-cell lung cancer', Briefings in bioinformatics **17**(2), 233–242.

Rauch, T. A. and Pfeifer, G. P. (2010), 'Dna methylation profiling using the methylated-cpg island recovery assay (mira)', Methods **52**(3), 213–217.

Reik, W. (2007), 'Stability and flexibility of epigenetic gene regulation in mammalian development', Nature **447**(7143), 425.

Ross, M. E., Zhou, X., Song, G., Shurtleff, S. A., Girtman, K., Williams, W. K., Liu, H.-C., Mahfouz, R., Raimondi, S. C., Lenny, N. et al. (2003), 'Classification of pediatric acute lymphoblastic leukemia by gene expression profiling', Blood **102**(8), 2951–2959.

Schirmer, E. C., Florens, L., Guan, T., Yates, J. R. and Gerace, L. (2003), 'Nuclear membrane proteins with potential disease links found by subtractive proteomics', Science **301**(5638), 1380–1382.

Schultz, M. D., He, Y., Whitaker, J. W., Hariharan, M., Mukamel, E. A., Leung, D., Rajagopal, N., Nery, J. R., Urich, M. A., Chen, H. et al. (2015), 'Human body epigenome maps reveal noncanonical dna methylation variation', Nature **523**(7559), 212.

Schwarzenbach, H., Hoon, D. S. and Pantel, K. (2011), 'Cell-free nucleic acids as biomarkers in cancer patients', Nature Reviews Cancer **11**(6), 426.

Serre, D., Lee, B. H. and Ting, A. H. (2009), 'Mbd-isolated genome sequencing provides a high-throughput and comprehensive survey of dna methylation in the human genome', Nucleic acids research **38**(2), 391–399.

Sgouros, G. (1993), 'Bone marrow dosimetry for radioimmunotherapy: theoretical considerations.', Journal of nuclear medicine: official publication, Society of Nuclear Medicine **34**(4), 689–694.

Smith, Z. D., Gu, H., Bock, C., Gnirke, A. and Meissner, A. (2009), 'High-throughput bisulfite sequencing in mammalian genomes', Methods **48**(3), 226–232.

Smith, Z. D. and Meissner, A. (2013), 'Dna methylation: roles in mammalian development', Nature Reviews Genetics **14**(3), 204.

Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. and Shendure, J. (2016), 'Cell-free dna comprises an in vivo nucleosome footprint that informs its tissues-of-origin', Cell **164**(1), 57–68.

Song, C.-X., Szulwach, K. E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.-H., Zhang, W., Jian, X. et al. (2011), 'Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine', Nature biotechnology **29**(1), 68.

Song, C.-X., Yin, S., Ma, L., Wheeler, A., Chen, Y., Zhang, Y., Liu, B., Xiong, J., Zhang, W., Hu, J. et al. (2017), '5-hydroxymethylcytosine signatures in cell-free dna provide information about tumor types and stages', Cell research **27**(10), 1231.

Spruijt, C. G., Gnerlich, F., Smits, A. H., Pfaffeneder, T., Jansen, P. W., Bauer, C., Münzel, M., Wagner, M., Müller, M., Khan, F. et al. (2013), 'Dynamic readers for 5-(hydroxy) methylcytosine and its oxidized derivatives', Cell **152**(5), 1146–1159.

Stefansson, O. A., Moran, S., Gomez, A., Sayols, S., Arribas-Jorba, C., Sandoval, J., Hilmarsdottir, H., Olafsdottir, E., Tryggvadottir, L., Jonasson, J. G. et al. (2015),

'A dna methylation-based definition of biologically distinct breast cancer subtypes', Molecular oncology **9**(3), 555–568.

Sun, K., Jiang, P., Chan, K. A., Wong, J., Cheng, Y. K., Liang, R. H., Chan, W.-k., Ma, E. S., Chan, S. L., Cheng, S. H. et al. (2015), 'Plasma dna tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments', Proceedings of the National Academy of Sciences **112**(40), E5503–E5512.

Szulwach, K. E., Li, X., Li, Y., Song, C.-X., Wu, H., Dai, Q., Irier, H., Upadhyay, A. K., Gearing, M., Levey, A. I. et al. (2011), '5-hmc–mediated epigenetic dynamics during postnatal neurodevelopment and aging', Nature neuroscience **14**(12), 1607.

Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L. M., Liu, D. R., Aravind, L. et al. (2009), 'Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1', Science **324**(5929), 930–935.

Taiwo, O., Wilson, G. A., Morris, T., Seisenberger, S., Reik, W., Pearce, D., Beck, S. and Butcher, L. M. (2012), 'Methylome analysis using medip-seq with low dna concentrations', Nature protocols **7**(4), 617.

Tanić, M. and Beck, S. (2017), 'Epigenome-wide association studies for cancer biomarker discovery in circulating cell-free dna: technical advances and challenges', Current opinion in genetics & development **42**, 48–55.

Teschendorff, A. E., Breeze, C. E., Zheng, S. C. and Beck, S. (2017), 'A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies', BMC bioinformatics **18**(1), 105.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002), 'Diagnosis of multiple

cancer types by shrunken centroids of gene expression', Proceedings of the National Academy of Sciences **99**(10), 6567–6572.

Ulz, P., Thallinger, G. G., Auer, M., Graf, R., Kashofer, K., Jahn, S. W., Abete, L., Pristauz, G., Petru, E., Geigl, J. B. et al. (2016), 'Inferring expressed genes by whole-genome sequencing of plasma dna', Nature genetics **48**(10), 1273.

Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J. et al. (2002), 'A gene-expression signature as a predictor of survival in breast cancer', New England Journal of Medicine **347**(25), 1999–2009.

Van Horssen, J., Wesseling, P., Van Den Heuvel, L. P., De Waal, R. M. and Verbeek, M. M. (2003), 'Heparan sulphate proteoglycans in alzheimer's disease and amyloid-related disorders', The Lancet Neurology **2**(8), 482–492.

Varley, K. E., Gertz, J., Bowling, K. M., Parker, S. L., Reddy, T. E., Pauli-Behn, F., Cross, M. K., Williams, B. A., Stamatoyannopoulos, J. A., Crawford, G. E. et al. (2013), 'Dynamic dna methylation across diverse human cell lines and tissues', Genome research **23**(3), 555–567.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P. et al. (2010), 'Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1', Cancer cell **17**(1), 98–110.

Virmani, A. K., Tsou, J. A., Siegmund, K. D., Shen, L. Y., Long, T. I., Laird, P. W., Gazdar, A. F. and Laird-Offringa, I. A. (2002), 'Hierarchical clustering of lung cancer cell lines using dna methylation markers', Cancer Epidemiology and Prevention Biomarkers **11**(3), 291–297.

Waddington, C. H. (1939), An introduction to modern genetics, George Allen And Unwin Ltd Museum Street; London.

Wang, F., Zhang, N., Wang, J., Wu, H. and Zheng, X. (2016), 'Tumor purity and differential methylation in cancer epigenomics', Briefings in functional genomics **15**(6), 408–419.

Wang, Y., Wang, X., Lee, T.-H., Mansoor, S. and Paterson, A. H. (2013), 'Gene body methylation shows distinct patterns associated with different gene origins and duplication modes and has a heterogeneous relationship with gene expression in oryza sativa (rice)', New Phytologist **198**(1), 274–283.

Warton, K. and Samimi, G. (2015), 'Methylation of cell-free circulating dna in the diagnosis of cancer', Frontiers in molecular biosciences **2**, 13.

Waseem Bihaqi, S., Schumacher, A., Maloney, B., K Lahiri, D. and H Zawia, N. (2012), 'Do epigenetic pathways initiate late onset alzheimer disease (load): towards a new paradigm', Current Alzheimer Research **9**(5), 574–588.

Weidner, C. I., Lin, Q., Koch, C. M., Eisele, L., Beier, F., Ziegler, P., Bauerschlag, D. O., Jöckel, K.-H., Erbel, R., Mühleisen, T. W. et al. (2014), 'Aging of blood can be tracked by dna methylation changes at just three cpg sites', Genome biology **15**(2), R24.

Weigelt, B., Glas, A. M., Wessels, L. F., Witteveen, A. T., Peterse, J. L. and van't Veer, L. J. (2003), 'Gene expression profiles of primary breast tumors maintained in distant metastases', Proceedings of the National Academy of Sciences **100**(26), 15901–15905.

Weissinger, E. M., Schiffer, E., Hertenstein, B., Ferrara, J. L., Holler, E., Stadler, M., Kolb, H.-J., Zander, A., Zürbig, P., Kellmann, M. et al. (2007), 'Proteomic

patterns predict acute graft-versus-host disease after allogeneic hematopoietic stem cell transplantation', Blood **109**(12), 5511–5519.

Wenk, G. L. et al. (2003), 'Neuropathologic changes in alzheimer's disease', Journal of Clinical Psychiatry **64**, 7–10.

Wiwie, C., Baumbach, J. and Röttger, R. (2015), 'Comparing the performance of biomedical clustering methods', Nature methods **12**(11), 1033.

Wong, Y. F., Selvanayagam, Z. E., Wei, N., Porter, J., Vittal, R., Hu, R., Lin, Y., Liao, J., Shih, J. W., Cheung, T. H. et al. (2003), 'Expression genomics of cervical cancer: molecular classification and prediction of radiotherapy response by dna microarray', Clinical cancer research **9**(15), 5486–5492.

Wray, N. R., Goddard, M. E. and Visscher, P. M. (2007), 'Prediction of individual genetic risk to disease from genome-wide association studies', Genome research **17**(10), 1520–1528.

Wu, H., D'Alessio, A. C., Ito, S., Wang, Z., Cui, K., Zhao, K., Sun, Y. E. and Zhang, Y. (2011), 'Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells', Genes & development **25**(7), 679–684.

Wu, H., Wang, C. and Wu, Z. (2012), 'A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data', Biostatistics **14**(2), 232–243.

Xu, R.-h., Wei, W., Krawczyk, M., Wang, W., Luo, H., Flagg, K., Yi, S., Shi, W., Quan, Q., Li, K. et al. (2017), 'Circulating tumour dna methylation markers for diagnosis and prognosis of hepatocellular carcinoma', Nature materials **16**(11), 1155.

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

Zhang, N., Wu, H.-J., Zhang, W., Wang, J., Wu, H. and Zheng, X. (2015), 'Predicting tumor purity from methylation microarray data', Bioinformatics **31**(21), 3401–3405.

Zheng, X., Zhang, N., Wu, H.-J. and Wu, H. (2017), 'Estimating and accounting for tumor purity in the analysis of dna methylation data from cancer studies', Genome biology **18**(1), 17.

Zheng, X., Zhao, Q., Wu, H.-J., Li, W., Wang, H., Meyer, C. A., Qin, Q. A., Xu, H., Zang, C., Jiang, P. et al. (2014), 'Methylpurify: tumor purity deconvolution and differential methylation detection from single tumor dna methylomes', Genome biology **15**(7), 419.

Zhuang, J., Jones, A., Lee, S.-H., Ng, E., Fiegl, H., Zikan, M., Cibula, D., Sargent, A., Salvesen, H. B., Jacobs, I. J. et al. (2012), 'The dynamics and prognostic potential of dna methylation changes at stem cell gene loci in women's cancer', PLoS genetics **8**(2), e1002517.

Zilliox, M. J. and Irizarry, R. A. (2007), 'A gene expression bar code for microarray data', Nature methods **4**(11), 911.