UNIVERSITY OF CALIFORNIA

Los Angeles

Multi-feature ensemble learning on cell-free DNA
for accurately detecting and locating cancer

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Bioinformatics

by

Mary Louisa Stackpole

2020

ABSTRACT OF THE DISSERTATION

Multi-feature ensemble learning on cell-free DNA
for accurately detecting and locating cancer

by

Mary Louisa Stackpole
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2020
Professor Xianghong Jasmine Zhou, Chair

Early cancer detection and localization using cell-free DNA (cfDNA) faces multiple challenges, including the low fraction of tumor DNA in cfDNA and the molecular heterogeneity of cancer. Many features have been used to detect cancer in cfDNA, such as fragment length profiles, copy number changes, and microbial composition, but methylation in particular has been found to detect cancer early. Additionally, the tissue specificity of methylation has aided noninvasive cancer typing efforts. Typically, cfDNA methylation profiling is done through whole genome bisulfite sequencing (WGBS) or targeted approaches, but these protocols are plagued by high cost or require prior knowledge of informative regions. Another procedure, reduced representation bisulfite sequencing (RRBS), strikes a balance between these two extremes, but is only applicable to intact genomic DNA, not naturally fragmented cfDNA. Herein, we develop an integrated cancer detection and typing system, CancerRadar, that addresses these challenges. First, we present a novel protocol, cell-free Methylation Sequencing (cfMethylSeq), which adapts the RRBS protocol to be applicable to cfDNA. We show cfMethylSeq yields more than 12-fold enrichment over WGBS in CpG islands while reliably and reproducibly quantifying methylation and capturing broad, genome-wide signals. Next, we develop a computational platform to extract information from cfMethylSeq data and diagnose the patient. The platform derives cfDNA methylation, cfDNA fragment sizes, copy number changes, and microbial composition from the raw cfMethylSeq data, and performs multi-feature ensemble learning.

We demonstrate the power of CancerRadar in detecting and locating cancer in a cohort of 275 colon, liver, lung, and stomach cancer patients and 204 non-cancer individuals. For cancer detection, we achieved a sensitivity of 89.1% at 97% specificity in the independent validation set. For cancer typing, we achieved an accuracy of 91.5% in the independent validation set. We further show that integrating multiple features significantly increases the detection power, especially for early-stage cancer. Our novel protocol and computational procedure have the potential to revolutionize cancer detection and methylation analyses in cfDNA, and the data generated will be hugely beneficial to the cfDNA research community.

The dissertation of Mary Louisa Stackpole is approved.

William Hsu

Wenyuan Li

Frank Alber

Xinshu Xiao

Xianghong Jasmine Zhou, Committee Chair

University of California, Los Angeles

2020

To my friends and family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The completion of this work was undoubtedly a group effort. Above all I must thank my advisor, Jasmine Zhou, for her continued support and guidance throughout my PhD. Her dedication to science, approach to problem solving, and innovative mindset have taught me a great deal. I appreciate all the time she has spent ensuring I succeed.

Additionally, Frank Alber, Grace Xiao, and William Hsu have offered many helpful comments throughout this process, and their time and attention is gratefully recognized. I would also like to thank SAOs Gene Gray and Mandy McWeeney, who helped me wade through the intricacies of transferring universities halfway through grad school.

Every single one of the past and present members of the Zhou lab was essential to the realization of this project. Xiaohui Ni has spent innumerable hours explaining experimental concepts to me. I thank him for his seemingly endless patience. Wenyuan Li played a similar role in the computational arena, and I am indebted to his attention to detail. Jim Liu's ability to quickly implement any random idea frequently left me in awe. In addition to their technical knowledge, Shuo Li and Qingjiao Li supplied much-needed non-scientific conversation in our shared lab spaces. Benny Zeng, Yonggang Zhou, Zuyang Yuan, and Shanshan He provided the experimental foundation without which this work would not have been possible.

My friends have played a critical role in my sanity over the past several years. Lindsey, Grace, and Cara have provided loads of vacations, conversations, and laughter. Special thanks to Liana for her oftentimes incomprehensibly optimistic view of grad school, as well as offering cooking lessons, movie nights, and even sometimes her cat, Ophie. And I am forever grateful to my oldest friend, Sara, for her lifelong friendship and her ability to somehow always pay attention during my ramblings.

My path here literally would not have been possible without my parents, who spent a week driving me across the country at the onset of grad school. Their curiosity in my projects over the years has forced me to improve my scientific communication, and their sacrifices, support, and encouragement have been crucial to my success.

And finally, to my husband John: thank you for embarking on this ridiculous adventure with me. The countless weekends we spent trekking around California, mornings we spent doing the crossword, and even the months we spent quarantined in our tiny apartment have offered so many bright spots along this lengthy and uncertain journey. In many ways this thesis is a direct result of your support, understanding, patience, and love.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1418060. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Chapters 2 and 3 are a version of a manuscript currently in preparation: **Mary L. Stackpole**, Weihua Zeng, Shuo Li, Chun-Chi Liu, Yonggang Zhou, Shanshan He, Angela Yeh, Ziye Wang, Fengzhu Sun, Qingjiao Li, Zuyang Yuan, Asli Yildirim, Ping Jung Chen, Paul Winograd, Shize Li, Zorawar Noor, Edward Garon, Samuel French, Clara E. Magyar, Sarah Dry, Clara Lajonchere, Daniel Geschwind, Gina Choi, Sammy Saab, Frank Alber, Wing Hung Wong, Steven M. Dubinett, Denise Aberle, Vatche Agopian, Steven-Huy Han, Xiaohui Ni, Wenyuan Li, Xianghong Jasmine Zhou. Multi-feature ensemble learning on cell-free DNA for accurately detecting and locating cancer. MLS, XN, YG, SH, and WZ designed and troubleshooted the experimental protocol. XN, YG, SH, and WZ performed wet-lab experiments. MLS analyzed and processed all data produced from the experiments. MLS and WL performed the cancer detection analyses. WL, SL, and MLS performed the cancer typing analyses. WL developed the read deconvolution framework used for Type 1 and 3 features. SL made the microbial and Type 3 features. All other features were made by MLS. MLS, WZ, WL, and XJZ wrote the paper with input from other authors. XJZ supervised the project.

VITA

2014	B.S. (Mathematics), University of Maryland, College Park, MD
2014	B.S. (Cellular Biology and Genetics), University of Maryland, College Park, MD
2014-2016	Graduate Student Researcher, Computational Biology PhD program, University of Southern California, Los Angeles, CA
2016	M.S. (Statistics), University of Southern California, Los Angeles, CA
2017-Present	PhD candidate, Bioinformatics IDP, University of California, Los Angeles, CA

PUBLICATIONS

Mary L. Stackpole, Weihua Zeng, Shuo Li, Chun-Chi Liu, Yonggang Zhou, Shanshan He, Angela Yeh, Ziye Wang, Fengzhu Sun, Qingjiao Li, Zuyang Yuan, Asli Yildirim, Ping Jung Chen, Paul Winograd, Shize Li, Zorawar Noor, Edward Garon, Samuel French, Clara E. Magyar, Sarah Dry, Clara Lajonchere, Daniel Geschwind, Gina Choi, Sammy Saab, Frank Alber, Wing Hung Wong, Steven M. Dubinett, Denise Aberle, Vatche Agopian, Steven-Huy Han, Xiaohui Ni, Wenyuan Li, Xianghong Jasmine Zhou. Multi-feature ensemble learning on cell-free DNA for accurately detecting and locating cancer. *In preparation.*

Wenyuan Li, Qingjiao Li, Shuli Kang, **Mary Same**, Yonggang Zhou, Carol Sun, Chun-Chi Liu, Lea Matsuoka, Linda Sher, Wing Hung Wong, et al. Cancerdetector: ultra-sensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. Nucleic acids research, 46(15):e89–e89, 2018

CHAPTER 1

Introduction

Cell-free DNA (cfDNA) is DNA found in the bloodstream that is not associated with cells. Since its presence was first described in 1948 [68], researchers have described numerous interesting characteristics, such as elevated levels in patients with cancer or other disorders [57] and the presence of fetal cfDNA in the plasma of pregnant women [67]. Curiously, cfDNA is fragmented in nature, exhibiting a characteristic length distribution with a peak around 160 bp. This has led to the theory that cfDNA is derived from nucleosome-bound fragments [42]. Although the biological mechanisms of cfDNA's release into the bloodstream remain unclear, mapping locations of naturally fragmented cfDNA have been used to infer gene expression [104], study nucleosome positioning [90], understand DNA damage [45, 106], and even reconstruct 3D genome organization [66]. Additionally, many researchers have taken advantage of the presence of cfDNA for use in different clinical scenarios, including prenatal diagnosis, cancer diagnosis, and organ transplantation monitoring [56]. For example, dramatic genome-wide hypomethylation, a trademark of cancer, can be detected in cfDNA and used to detect malignancies [12]. These liquid biopsies offer a noninvasive alternative to traditional biopsies [12].

Early detection of cancer before it metastasizes holds our greatest hope for increasing cancer survival. Understandably, cfDNA has drawn attention for solving this task due to its potential to noninvasively detect, pinpoint, and monitor cancer in blood [59, 56, 12, 25]. However, this effort is hindered by major challenges. First, the amount of tumor DNA present in the blood is low, especially in early stages of cancer [59]. Second, due to the diversity of cancer types, subtypes, and additional covariates such as age, environment, or comorbidities, genetic and epigenetic aberrations associated with cancer can vary wildly.

Third, due to the heterogeneity of cancer, any study that undertakes cfDNA-based cancer detection must employ a large sample size of both cancer and normal samples.

Due to the low amount of tumor DNA present in the blood, a successful cancer detection method should capture as many tumor-derived fragments as possible. Some current studies aim to do this through small, panel-based approaches that deeply sequence areas likely to contain tumor cfDNA fragments [64, 14, 117]. These studies require targets to be identified beforehand. Alternatively, genome-wide approaches have also been used, aiming to capture broad signals to make up for the lack of individual tumor-derived signals [99, 59, 46, 15]. However, whole-genome sequencing is cost prohibitive for clinical use.

The heterogeneous nature of cancer hints that a noninvasive test that can capture diverse attributes of cancer has the best change of success. Indeed, several cfDNA features have been shown to have diagnostic power, including cfDNA methylation [64, 59, 46, 25, 117, 89], fragment length [15, 42, 90], copy number variation (CNV) [12, 15], and microbiome composition [82]. However, these disparate features have never been comprehensively integrated into one model before. This is because library preparation methods that can profile methylation deeply, such as targeted panels, lack genome-wide information about fragment length or CNV. Similarly, genome-wide sequencing techniques often lack the depth needed to use methylation features effectively, if methylation is profiled at all. Whole genome bisulfite sequencing (WGBS), currently the only commercially available method for genome-wide methylation profiling in cfDNA, can obtain all these features, but it is prohibitively expensive for large scale studies. An alternative genome-wide methylation profiling technique available for genomic DNA is reduced representation bisulfite sequencing (RRBS), a method that employs restriction enzymes to cut intact DNA into small fragments in regions with a high CpG content, and subsequently size-selects these small fragments to enrich for CpG sites. However, due to the fragmented nature of cfDNA, this RRBS approach cannot be used.

To address all these challenges, this dissertation presents an integrated experimental and computational system, CancerRadar, for the accurate and affordable detection of cancer. This system includes (1) a cost-effective experimental approach to comprehensively profile diverse cfDNA features, and (2) an integrative computational learning framework for cancer detection which is scalable to many types of features and a high number of markers.

Our experimental approach, named cell-free DNA Methylation Sequencing (cfMethylSeq), adapts the RRBS technology to work on cfDNA. By successfully doing so, we dramatically lower the cost of genome-wide methylation profiling in cfDNA. At the same time, we are able to profile genome-wide features such as CNV, fragment length, and microbial composition for no additional costs.

We sequenced hundreds of cfDNA samples with our cfMethylSeq approach. Our computational system integrates the methylation and other features obtained from the cfMethylSeq data into a multi-modal predictive model that accurately and sensitively detects cancer in cfDNA. Our results demonstrated the strong complementary effect of the heterogeneous feature types, and our method is adaptable and scalable to increased feature and sample sizes.

In chapter 2, we describe the development of the cfMethylSeq method by adapting the RRBS technology to be applicable to cfDNA. RRBS involves a restriction enzyme digestion step followed by a size selection. The restriction enzyme step cuts at CCGG sites in the genome; places where these sites are close together produce smaller fragments than regions with low CCGG content. The subsequent size selection step then enriches for these CCGG-dense regions, resulting in a final library that preferentially covers CpG islands, gene promoters, and other methylation-informative regions [24]. cfDNA is naturally fragmented, with most fragments around 165 bp, so the size selection step key to enrichment in RRBS cannot be used. In brief, to adapt the RRBS procedure to apply to fragmented DNA, all original cfDNA fragments are blocked from ligating to adapters, and only those fragments that are cut twice by a restriction enzyme are able to ligate to adapters and get sequenced. We performed extensive benchmarking of our procedure, and showed that it (1) produced enrichment in genomic regions comparable to RRBS, and (2) reliably and reproducibly called methylation. Our procedure yields a 12.8 fold cost reduction compared to WGBS.

In chapter 3, we use solid tissue RRBS data and the hundreds of cfDNA samples sequenced with our novel cfMethylSeq protocol to develop and test our computational method to detect and type cancer. We employ the highly successful stacked classifier framework [1] and achieve 85.6% sensitivity at 99% specificity in leave-one-out cross validation, and 89.1% sensitivity at 97% specificity in the independent validation cohort. Existing methods for cancer detection and typing in cfDNA rely on targeted gene panels [14], WGBS data [59, 46, 12], targeted sequencing [65], or other experimental approaches [89, 25]. Our approach not only outperforms all these other methods, but does so inexpensively and without having to find targets a priori. The stacked classifier framework is an ensemble machine learning approach using many different features and classifier types. We used broad, genome wide features such as counts of fragments in genomic bins, enrichment in genomic areas, and overall methylation levels, in addition to small (<350 bp in length), learned methylation features.

In summary, we developed CancerRadar, an integrated computational and experimental framework for affordable, noninvasive cancer detection. Our experimental approach, cfMethylSeq, dramatically lowers the cost of methylome profiling in cfDNA, and does not require a priori selection of targets. Our computational framework makes use of not only the methylation information gleaned from the cfMethylSeq data, but also genome-wide attributes such as copy number changes. Importantly, the cfMethylSeq data we generated constitutes a wealth of information that can still be learned from as more samples are collected. Because we did not target a few specific genes [117] or use a biased methylation approach [89], our data can be used for countless other purposes. Additionally, our computational cancer detection framework can adapt and improve as more training samples are acquired.

CHAPTER 2

cfMethylSeq: a novel protocol for profiling methylation in cfDNA

2.1 Introduction

Cytosine DNA methylation is a stably inherited DNA modification that has the potential to alter chromatin structure and transcription of genes [63]. It has been implicated in various biological mechanisms such as cellular differentiation, gene regulation, suppression of transposable elements, and development [61, 84, 93, 79]. These epigenetic alterations have been increasingly observed as playing a role in cancer in the past few decades [35]. Methylation at CpG dinucleotides in particular has been observed to change in both a genome-wide manner and at the individual gene level in cancer [35]. Noninvasive liquid biopsies using cfDNA have garnered much attention in the recent past due to their potential to transform cancer diagnosis, screening, and treatment [2]. Methylation-based cfDNA cancer detection methods may be more promising than approaches based on mutations [77, 78] because pervasive methylation changes are one of the first hallmarks of cancer [12]. Furthermore, methylation has been shown to be tissue-specific, lending its use to cancer typing [96, 46, 73]. Therefore an experimental method that can accurately measure methylation in cfDNA is of great importance.

Measurement of cytosine methylation typically relies on bisulfite conversion. The chemical sodium bisulfite deaminates unmethylated cytosines to uracil, while methylated cytosines are left unchanged [31, 22, 71]. After bisulfite conversion and subsequent PCR amplification, this uracil is converted to a thymine basepair. Ultimately, methylated cytosines will remain as cytosines, while unmethylated cytosines will be read as thymine. The use of sodium bisulfite has been applied to several sequencing and probe-based techniques for measuring methylation in DNA. In techniques that yield basepair resolution, these bisulfite-converted sequenced reads can then be mapped to an in silico bisulfite converted genome using programs such as Bismark [51]. Any T mapped to a C in the reference genome is considered unmethylated, whereas any C that remained as a C is considered methylated. Other experimental techniques use the differences in sequences generated from bisulfite conversion of methylated and unmethylated DNA to design capture probes, where certain probes can only bind to the unmethylated form of a bisulfite-converted DNA sequence while others bind to the methylated form. There are over 28 million CpG sites in the human genome; most approaches profile the methylation of a small fraction of these sites [92]. Below we outline various methods for profiling methylation:

1. Whole genome bisulfite sequencing (WGBS)

   Whole genome bisulfite sequencing (WGBS) profiles methylation at nearly every CpG site in the genome at basepair resolution using bisulfite conversion [40]. It is considered the gold standard for methylation analyses [75]. Since there is no targeting of specific genomic regions, the whole genome is sequenced without bias. Consequently, this method is cost prohibitive for most studies that require deep sequencing of specific loci of interest. Since the first genome-wide WGBS study in humans [63], WGBS has been widely used but rarely on large scales [122]. Since large swaths of the genome contain no CpG sites, many reads sequenced with WGBS are wasted data for methylation analyses. WGBS in cfDNA also requires large amounts of input material, further limiting its use [122]. Nevertheless, WGBS on cfDNA has been used in numerous small scale studies [12, 42, 59, 46, 41].

2. Reduced representation bisulfite sequencing (RRBS)

   An alternative to WGBS is reduced representation bisulfite sequencing (RRBS) [71, 24]. RRBS typically involves a restriction enzyme digest step that cuts the genome at CCGG sites. A size selection step then selects fragments that are within a short size range (<350 bp), enriching for regions where CCGG sites are close together. Since these regions are mostly in CpG islands (CGIs) and gene promoters, RRBS is able to sequence the majority of methylation-informative genomic regions for relatively low cost, while only in theory covering around 4% of the genome. Traditional RRBS also provides representative, but lower, coverage of other genomic regions, such as CGI shores [24]. When applicable, traditional RRBS is useful for studying changes in methylation that occur as a result of diseases or normal biological processes, as these changes often occur in CGIs and gene promoters. Because the amount of the genome covered is small, this method can inexpensively sequence to a high depth [24]. Since this method requires a size-selection step for enrichment, it is inapplicable to fragmented DNA, like cfDNA [62]. Nevertheless, one study did apply a single cell RRBS protocol directly to cfDNA [25], with limited enrichment for the regions of interest.

3. Microarray

   The Infinium HumanMethylation 27 (Infinium 27K), Infinium HumanMethylation450 (Infinium 450k), and Infinium EPIC arrays are microarray-based technologies [85] covering around 27,000, 450,000, and 850,000 CpG sites in the human genome, respectively. Microarray technologies hybridize bisulfite-converted DNA to a chip containing beads. There are two beads for each profiled CpG site, one for the methylated version and one for the unmethylated version. After hybridization, single-base extension occurs with differentially labeled nucleotides for the methylated and unmethylated cases. After staining to differentiate the two types of incorporated nucleotides, the chip is scanned to determine the ratio of intensities between the unmethylated and methylated versions at each probe, yielding a methylation ratio at each profiled CpG site, also known as a β value [6]. These platforms have been widely used, due to their low cost and simplicity. A large number of 450k datasets are publicly available on the Gene Expression Omnibus (GEO) [5] as well as The Cancer Genome Atlas (TCGA) [114]. However, due to their limited number of profiled CpG sites, they are unable to be used for an exhaustive search of epigenetically modified loci across the genome [101]. Because these methods measure the average methylation at an individual CpG site across all reads covering the site of interest, they lack the read level information used in more recent cancer detection algorithms [59, 25]. Furthermore, 500 ng-1 μg input DNA is required [98], rendering microarrays inapplicable to cfDNA, although one study used microarrays on cfDNA by pooling samples [23]. Instead, studies involving cfDNA often use the large amount of publicly available 450k data to learn biomarkers, and then these biomarkers are used to detect cancer in cfDNA [59, 46] or develop targets for panel-based approaches [117].

4. Methylation capture sequencing (MC Seq)

   Methylation capture sequencing (MC Seq) uses target-specific bait sequences. This allows for specific survey of genomic loci of interest. This also has lower cost and processing time than WGBS, and overcomes the limitation of lower genomic coverage in the microarray-based methods. However, the baits targeting the regions of interest and the regions themselves must be identified [122].

   MC Seq uses DNA (or RNA) baits that contain complementary sequences of targeted regions to select these regions for sequencing. Since baits are specifically designed for regions of interest, bias due to CpG density is eliminated. Despite its benefits over WGBS, microarray, and other affinity-enrichment platforms, MC Seq has been used little in the field [36].

   MC Seq can be approached in two ways: convert-then-capture, or capture-then-convert. In convert-then-capture, the DNA is converted using bisulfite and then captured. In capture-then-convert, the DNA is first captured and then treated with bisulfite, so common methylation patterns do not have to be considered in the bait design step. However, the capture-then-convert approach requires large amounts of native, unamplified DNA as input [37]. Bisulfite conversion can also lead to substantial DNA loss, further hampering its use in the small amounts of material available post-capture, and leading to substantial PCR duplication [58]. Although targeted bisulfite sequencing has numerous advantages over WGBS, RRBS, and microarrays, such as less wasted data and the ability to tailor data to regions of interest, it has been underrepresented in the literature [54]. This is likely because of the up front initial investment required to identify and generate the capture probes.

5. Methylated DNA Immunoprecipitation Sequencing (MeDIP-seq)

   Methylated DNA Immunoprecipitation Sequencing (MeDIP-seq) uses methylation-specific antibodies to extract DNA fragments containing methylated cytosines [40, 112]. These fragments are then sequenced, with or without bisulfite conversion. MeDIP-seq allows for broader coverage than RRBS, but does not allow for base-pair resolution if bisulfite conversion has not been applied. Only fragments that contain methylated cytosines are sequenced, so unbiased methylation ratios cannot be obtained. Furthermore, enrichment of regions can vary due to CpG density [36]. A cfDNA version, cfMeDIP-seq, was developed to adapt the MeDIP-seq procedure to low input amounts [89].

Of the available methods for profiling DNA methylation, RRBS [71, 24] has distinct advantages. Compared to WGBS, RRBS is able to sequence the majority of promoters and other methylation-informative genomic regions for relatively low cost. In WGBS, around 35% of sequenced fragments contain no CpG sites and are therefore wasted data for methylation analyses. RRBS, in contrast, guarantees that every sequenced fragment contains a CpG site, dramatically lowering the cost associated with deep methylation profiling. RRBS also holds advantages over targeted sequencing techniques, such as methylation capture sequencing. While these methods can deeply sequence genomic regions of interest, they require informative targets to be selected beforehand, limiting their use for exploratory studies. However, because cfDNA is naturally fragmented, exhibiting a characteristic length around 165 bp [96], size selection does not yield enrichment when RRBS is applied to cfDNA [62].

Specifically, the traditional RRBS protocol enriches for CpG-rich genomic regions in intact genomic DNA by cutting the genome with a restriction enzyme (typically MspI, cut site 5'-C↓CGG-3'), and then selecting short fragments. Ultimately, the final library consists of genomic regions where CCGG sites are close together. These regions are mostly CGIs and gene promoters [24]. Most cfDNA fragments are around 165 bp [96, 90, 15]. This is thought to be related to nucleosome wrapping, and has been exploited to noninvasively infer gene expression or tissue composition [104, 90]. In regards to RRBS, however, this small fragment size of 165 bp falls within the size selection range that is key to enriching CpG-rich regions of the genome. This size selection step yields little enrichment when applied to fragmented DNA.

In this chapter, we describe the development and validation of the cfMethylSeq protocol, an adaptation of RRBS for cfDNA. First, initial simulations are described to assess the feasibility of performing RRBS on fragmented DNA. Once proven useful, an initial protocol is developed and troubleshooted. The finalized protocol first blocks all original cfDNA fragments from ligating to adapters, then the restriction enzyme digest is performed. Only those fragments that are cut twice by a restriction enzyme are able to ligate to adapters and get sequenced. Hundreds of cfDNA samples are then sequenced with the finalized protocol. These samples are analyzed in aggregate to confirm enrichment in the regions of interest, verify methylation calls are correct, and to ensure reproducibility. Our cfMethylSeq procedure yields greater enrichment, basepair resolution, and less sample loss compared to other cfDNA-based methylome approaches. This enables the methylomes of cfDNA samples to be inexpensively analyzed at basepair resolution, a much-needed resource in the noninvasive cancer detection field. Furthermore, because there is no need to identify targets beforehand, the data generated using the cfMethylSeq procedure can be used to not only discover new biomarkers in cancer and other diseases but also to study basic biological processes such as aging.

2.2 Results

2.2.1 Feasibility simulations

In silico studies were performed to determine if RRBS would be useful on cfDNA, i.e., if the characteristic enrichment found in methylation-informative regions in RRBS on genomic DNA still held in cfDNA. There are two reasons why this might not be the case. First, because cfDNA fragments are on average 165 bp in length, they will lead to naturally shorter

RRBS libraries compared to RRBS libraries on intact genomic DNA—regions in the genome where CCGG sites are over 165 bp apart will rarely be captured in cfDNA, but easily found in intact genomic DNA. Second, the genomic regions covered by cfDNA are not totally random [78], so enrichment in methylation informative regions could be less or more than anticipated. To test if the procedure would be useful, we performed in silico MspI digestion on high coverage whole-genome sequencing (WGS) cfDNA data from [90]. Typical RRBS applied to genomic DNA has around 40% of fragments in CGIs and 22% of fragments in gene promoters. In silico MspI digestion of high coverage WGS cfDNA data found 39.5% of final fragments in CGIs and 18.8% of final fragments in gene promoters, after mapping and without deduplication. For comparison, the original WGS libraries had about 3% of their fragments in CGIs and 1.4% in promoters. These analyses allowed us to conclude that, if the size selection step could be bypassed, RRBS would still provide similar enrichment in CGIs and promoter regions when applied to cfDNA.

2.2.2 The cfMethylSeq procedure

Once the application of RRBS to cfDNA was shown to have potential, we developed the cell-free DNA Methylation Sequencing (cfMethylSeq) technique. The procedure is outlined in figure 2.1a. In brief, we first block both ends of all cfDNA fragments by dephosphorylating their 5'-ends and adding ddNTP to their 3'-ends. These fragments cannot be ligated to adapters and will not be sequenced. After MspI digestion (cut site 5'-C↓CGG-3'), only digested cfDNA fragments with two or more CCGG sites will be able to ligate to adapters containing duplex unique molecular identifiers (UMIs) and get sequenced, resulting in a final library enriched in CpG sites. As shown in figures 2.1b and 2.1c, the cfMethylSeq libraries on cfDNA show characteristic insert fragment lengths at 68 bp, 135 bp, and 203 bp, similar to patterns seen in conventional RRBS libraries prepared from solid tissue. These peaks are a result of MspI digestion of Alu repeat elements. In contrast, the conventional RRBS libraries prepared from cfDNA show a strong peak/band with 160 bp insert—the characteristic size of full-length cfDNA fragments without MspI digestion. This indicates that, as expected, a large proportion of the undigested 165 bp cfDNA fragments were captured during traditional

RRBS, and little enrichment in the regions of interest will be observed.

2.2.3 Computational validation of final procedure

Hundreds of samples were sequenced with our cfMethylSeq procedure. Below we analyze their statistics in aggregate and compare to RRBS on solid tissue samples and WGBS on cfDNA.

2.2.3.1 Reads from cfMethylSeq libraries map to expected locations

Because RRBS involves digestion with the restriction enzyme MspI, all resulting sequenced fragments start and end at known locations in the genome where CCGG sites (MspI cut sites) are within a certain distance from each other. We can measure our on-target rate by calculating how many fragments in our library map exactly to such fragments. On average about 85% of all cfMethylSeq reads fell in expected RRBS locations, compared to about 92% for typical RRBS libraries on solid tissue. More broadly, we can also measure whether the fragments start and end at CCGG sites. The vast majority of our cfMethylSeq fragments had CCGG cut sites on both ends, with the next most common scenario being a CCGG site on only one end. Specifically, on average 85.7% of reads in 479 cfMethylSeq libraries have MspI sites on both ends, compared to 91.8% of reads in 251 conventional RRBS libraries on solid tissues and 0.006% of reads in 37 cfDNA WGBS libraries. Our slight reduction compared to RRBS is due to incomplete ddNTP labeling or no dephosphorylation during our procedure, allowing some fragment ends to ligate to adapters even though they were not the result of MspI cleavage. For comparison, in WGBS almost none of the fragments fall in these locations (figure 2.2a).

WGBS covers far more CpG sites than either RRBS or cfMethylSeq, but cfMethylSeq covers the vast majority (97.5%) of CpG sites that are also covered by RRBS (figure 2.2b). About 30% of the CpG sites covered by cfMethylSeq were not covered by RRBS; this is due to incomplete labeling. However, these CpG sites are not covered to the same depths as CpG sites shared between cfMethylSeq and RRBS. Figure 2.2c shows the coverage of

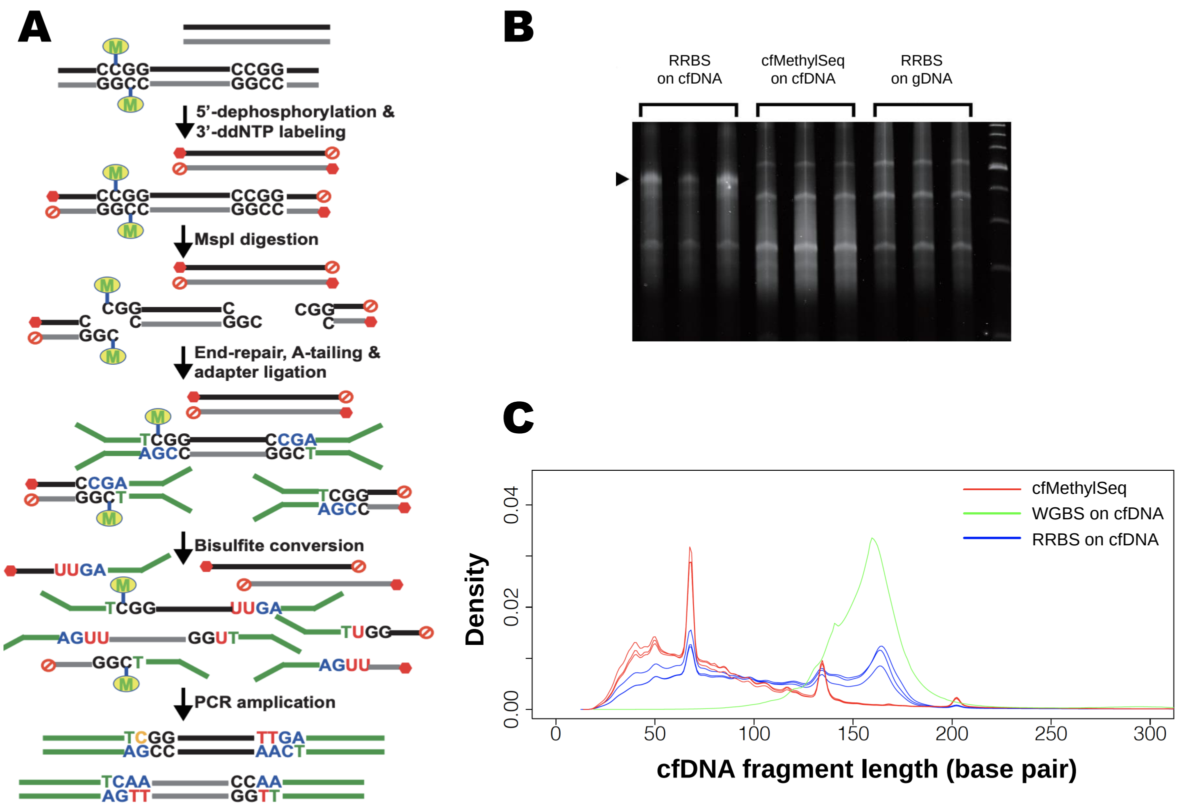

Figure 2.1: The cfMethylSeq procedure. (A) Diagram of the cfMethylSeq protocol (B) Typical TBE-UREA PAGE image of cfMethylSeq libraries made from cfDNA, compared with conventional RRBS with cfDNA or intact genomic DNA as input material. The non-specific ligation product from undigested cfDNA fragments with the conventional RRBS protocol is indicated by an arrow. (C) The fragment length profiles of libraries sequenced with the cfMethylSeq protocol on cfDNA (red), compared to WGBS on cfDNA (green) and conventional RRBS on cfDNA (blue).

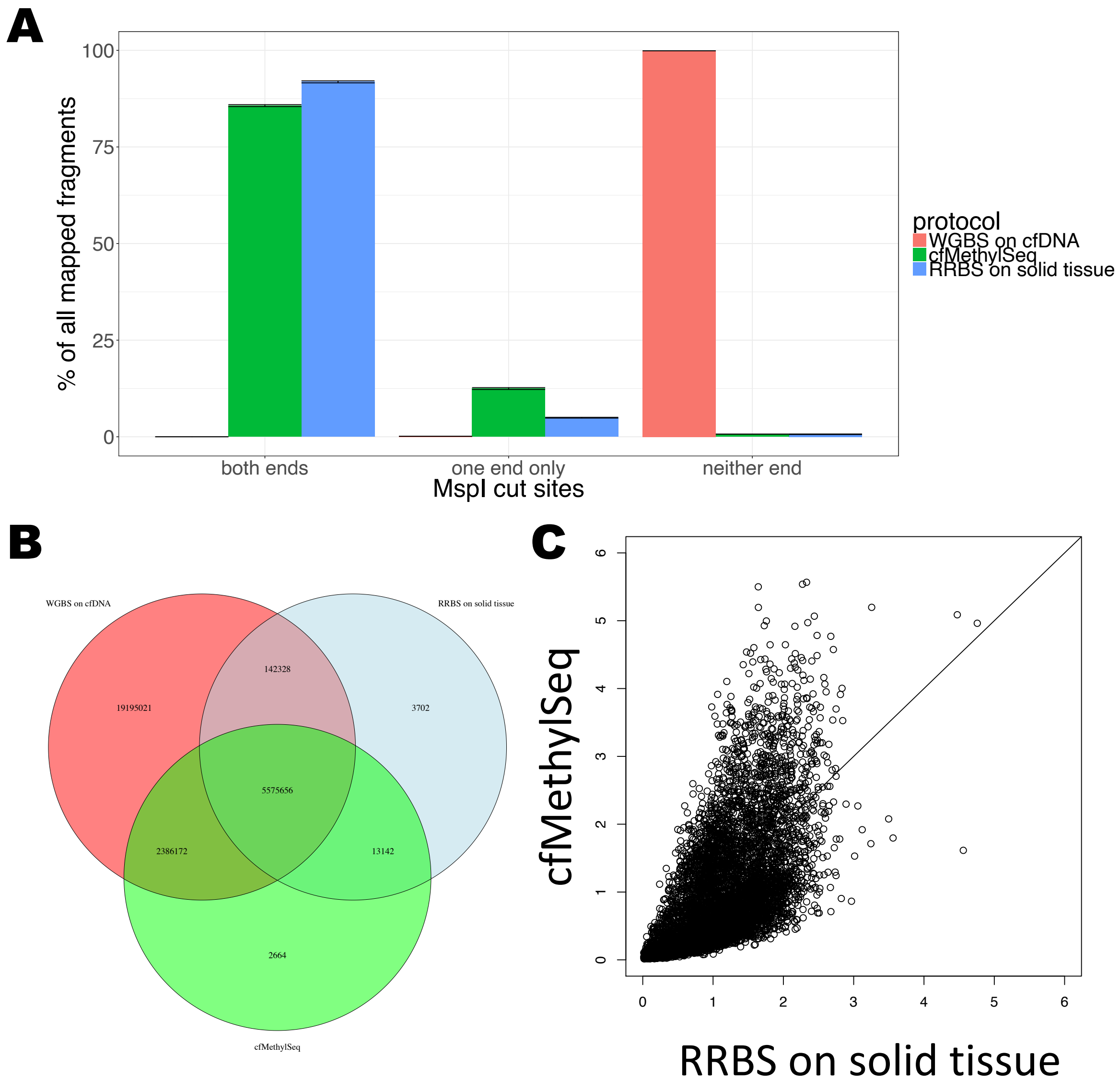

Figure 2.2: cfMethylSeq reads fall in expected locations. (A) The percentage of reads with MspI sites on both ends, on only one end, and on neither end from our cfMethylSeq protocol on cfDNA (green), RRBS on solid tissue (blue), and WGBS on cfDNA (red). (B) Venn diagram of CpG sites covered by RRBS on solid tissue, cfMethylSeq on cfDNA, and WGBS on cfDNA. A CpG site is considered covered if it sequenced at least once in over 90% of the samples profiled for each protocol. (C) Read coverage in 10000 randomly chosen CpG sites covered by both RRBS and cfMethylSeq samples. Each point in the scatter plot stands for one CpG sites, the x-coordinate is the normalized coverage in RRBS on solid tissue, and the y-coordinate is the normalized coverage in cfMethylSeq.

10000 random CpG sites covered by both cfMethylSeq and RRBS samples. The x-axis is the coverage in RRBS and the y-axis is the coverage in cfMethylSeq, after normalization. The coverage is highly consistent between RRBS and cfMethylSeq (Pearson correlation 0.84), meaning cfMethylSeq is producing libraries with profiles similar to RRBS.

2.2.3.2 Genomic enrichment

As a result of our high on target rates, CpG-dense regions are enriched. 34.11%, 12.38%, and 13.14% of cfMethylSeq reads fall into CGI, shore, and shelf regions, compared to 33.65%, 13.35%, and 14.04% for conventional RRBS libraries on solid tissue. In WGBS cfDNA libraries, only 2.66% of reads fall in CGIs while most (88.32%) fall in uninformative opensea regions (figure 2.3a). That is, cfMethylSeq offers 12.8 fold enrichment over WGBS in CGIs. Similarly, in gene regions, 15.44% of cfMethylSeq reads fall in gene promoter regions and 23.33% fall in exonic regions, compared to 17.12% and 23.12% in conventional RRBS libraries on solid tissue, while only 1.1% and 5.4% of WGBS reads fall in promoters and exons, respectively (figure 2.3b).

Due to the limited number of CpGs profiled in comparison to WGBS, cfMethylSeq reaches a much higher depth for a much lower number of mapped reads, similar to RRBS (figure 2.3c). In WGBS, hundreds of millions of mapped read pairs must be sequenced to reach even moderate depth, compared to around 50 million for RRBS and cfMethylSeq. This highlights the cost-effective nature of the cfMethylSeq and RRBS protocols.

2.2.3.3 Validity and reproducibility of cfMethylSeq

To ensure our cfMethylSeq procedure could accurately profile methylation levels, a solid tissue sample was sequenced with RRBS, and sonicated and sequenced with cfMethylSeq. Comparisons between traditional RRBS libraries on solid tissue and cfMethylSeq libraries on the same tissue's sheared gDNA show that cfMethylSeq can obtain similar methylation levels with correlation increasing as coverage increases (figure 2.4).

For CpG sites with more than 10x coverage in both approaches, the correlation between

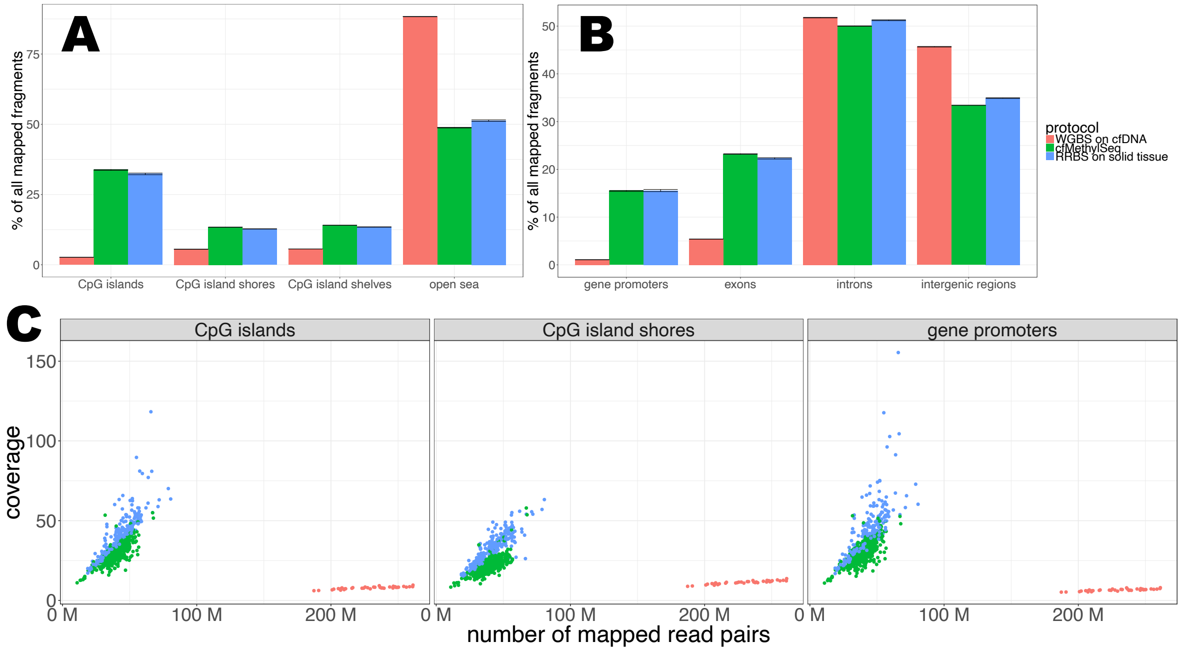

Figure 2.3: cfMethylSeq offers genomic enrichment and reduced cost. (A) The percentage of mapped fragments that fall in CGIs, CGI shores, CGI shelves, and opensea regions is shown for cfMethylSeq libraries, RRBS libraries and WGBS libraries on cfDNA. (B) The percentage of mapped fragments that fall in gene promoters, exons, introns, and intergenic regions is shown for cfMethylSeq libraries, RRBS libraries, and WGBS libraries on cfDNA. Regions are defined by UCSC table browser. (C) The number of mapped read pairs (x axis) required to obtain a certain depth of coverage (y axis) over CpG sites covered at least once in each procedure in cfMethylSeq (green), WGBS (red), and RRBS (blue).

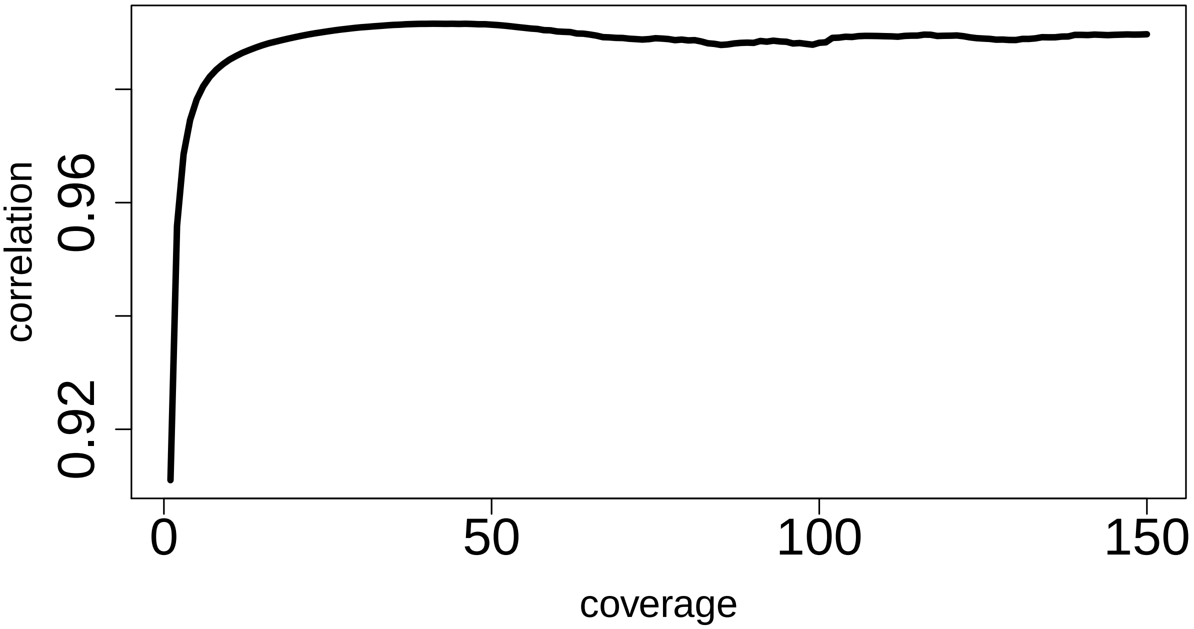

Figure 2.4: Methylation concordance between a solid tissue sample sequenced with RRBS, and sheared and sequenced with cfMethylSeq, increases with depth of coverage. The traditional RRBS sample and sheared cfMethylSeq sample are subsetted to the CpG sites that are covered by both samples at minimum depth of coverage specified on the x-axis. Then, the Pearson correlation (y-axis) of the methylation rate (β value) is taken in these CpG sites.

	sheared cfMethylSeq vs intact RRBS	technical cfMethylSeq replicates
Coverage correlation	0.8962661	0.9832131
Methylation correlation	0.9852304	0.9904725

Table 2.1: Concordance of coverage and methylation calls between (column 1) a gDNA sample sequenced with traditional RRBS and sheared and sequenced with cfMethylSeq and (column 2) replicate cfDNA samples both sequenced with cfMethylSeq. Pearson correlation measurements are calculated among CpG sites covered $> 10x$ in both samples.

methylation levels in cfMethylSeq on sheared gDNA and RRBS on intact gDNA is 0.9852 for the same sample (first column of table 2.1). Additionally, to test the reproducibility of cfMethylSeq, one cfDNA sample was sequenced twice using the procedure (second column of table 2.1). The coverage and methylation are highly consistent between the replicates.

Overall these experiments showed that cfMethylSeq can accurately profile methylation levels, is similar to RRBS in terms of genomic coverage, and can be performed multiple times and get consistent results.

2.2.4 Comparison to other methods

2.2.4.1 Traditional RRBS on cfDNA

Three cfDNA samples were sequenced with both cfMethylSeq and traditional RRBS. Table 2.2 shows the on target rate in these libraries. Typically for traditional RRBS on solid tissue, this number is over 90%. For cfMethylSeq, the value is typically around 85%. Traditional RRBS on cfDNA yielded on target rates of 28-35%; while this is much lower than the 85% observed in cfMethylSeq, it is still higher than expected. Initially we anticipated there would be almost no enrichment.

Fragment length profiles for these libraries are shown in figure 2.5. As expected, the RRBS libraries have a peak around 160 bp for the undigested cfDNA that was able to pass

sample	protocol	on target rate	% genome covered
1	RRBS	27.30740	45.191152
2	RRBS	35.83175	28.096360
3	RRBS	28.26273	43.354064
1	cfMethylSeq	87.72469	7.361843
2	cfMethylSeq	89.99379	6.459448
3	cfMethylSeq	86.02260	8.315694

Table 2.2: Comparison between cfMethylSeq and traditional RRBS on three cfDNA samples. The on target rate measures the percentage of mapped fragments that fell in characteristic RRBS locations. % genome covered measures the percent of basepairs in the genome that were covered by at least one mapped fragment

the size selection step. However, these libraries still have an RRBS peak at 68 bp; there is still some enrichment for RRBS regions.

Overall these libraries show that RRBS is not directly applicable to cfDNA, but there is more enrichment than theoretically anticipated. There is still a peak at 68 bp, only about 50% of the genome is covered rather than $> 90\%$, and the on target rate is close to 30%, not near 0 as originally predicted from simulations on WGS data.

It is likely that the enrichment is due to the fill-in step in the library protocol. As observed in [45], cfDNA fragments often have jagged (5' or 3' protruding) ends, with estimates of around 87% of all cfDNA fragments having jagged ends. In order to be sequenced in a double-stranded library protocol, such as cfMethylSeq, these ends need to be filled in before adapters can ligate. During RRBS, MspI digestion will produce jagged fragment ends that need to be filled in containing only C and G. Consequently, during the end repair step, only C, G, and A (for A-tailing the fragment after end repair) are added. Therefore any fragment that needs to be filled in with a T nucleotide cannot be end repaired and cannot ligate to adapters. It is likely that these jagged ends often contain A, and cannot be end-repaired nor ligate to adapters. This could lead to a lot of original, undigested cfDNA fragments being unable to be

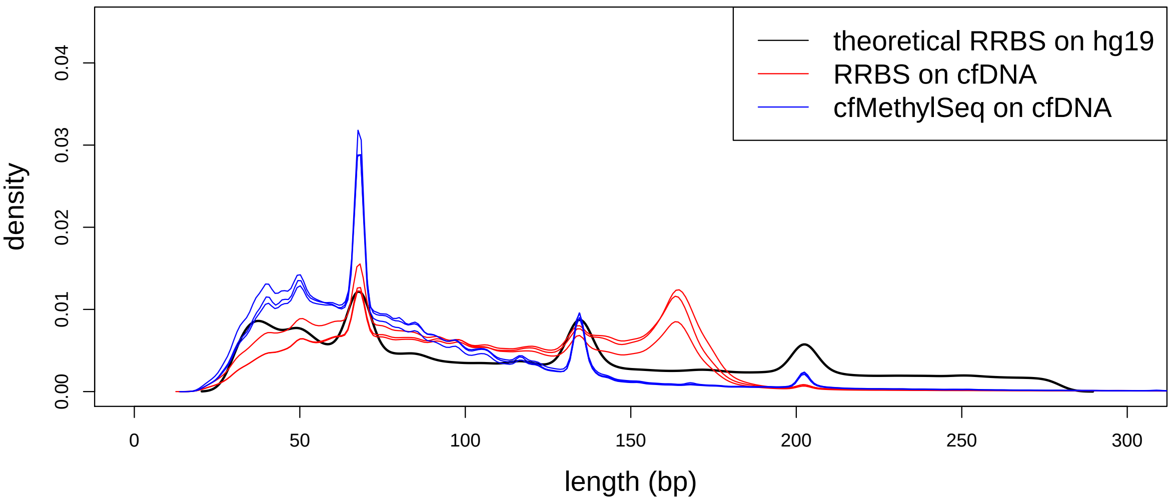

Figure 2.5: Length profiles in cfMethylSeq on three cfDNA samples (blue) and traditional RRBS on the same three cfDNA samples (blue). The black line represents the length profile from in silico RRBS on hg19.

sequenced, leading to the observed enrichment of MspI digested fragments not predicted by the initial simulations. Furthermore, the high damage observed in cfDNA [90] could lead to a natural dephosphorylation in cfDNA fragments, even though no dephosphorylation step was performed. This would prevent adapter ligation in these damaged fragments. Nevertheless, although some enrichment is observed, it is nowhere near the levels of cfMethylSeq and will not offer the cost reduction of our method.

2.2.4.2 cf-RRBS

While this dissertation was being prepared, van Paemel et al. [105] simultaneously developed a method very similar to our cfMethylSeq method for applying RRBS to cfDNA, named cf-RRBS. In our cfMethylSeq procedure, cfDNA fragments without the desired digestion sites are blocked from ligating to adapters on both ends. In cf-RRBS, fragments without the desired digestion sites can still ligate to adapters, but in this scenario a nick is formed between the adapter and the fragment. These fragments are subsequently removed with

sample	protocol	% adapter dimers	on target rate
1	cf-RRBS	65.93863	84.80486
2	cf-RRBS	54.66597	83.38934
3	cf-RRBS	45.30468	91.91102
1	cfMethylSeq	21.38745	89.38593
2	cfMethylSeq	7.52928	88.99743
3	cfMethylSeq	3.33045	94.12502

Table 2.3: Mapping statistics from 3 cfDNA samples sequenced with our cfMethylSeq method and cf-RRBS [105]

an exonuclease digestion step that removes nicked DNA [17]. This cf-RRBS technique has two drawbacks, leading to the under-utilization of the precious cfDNA. First, in addition to removing cfDNA fragments without the desired digestion sites, exonuclease digestion will also remove any DNA that contains nicks, a scenario that has been estimated to occur in 30% of cfDNA fragments [11]. In contrast, our cfMethylSeq procedure can still build the intact cfDNA strand into the library if only one strand contains the nick. Second, adapters ligate to all cfDNA fragments, including those without the desired digestion sites, leading to lower ligation efficiency in the fragments of interest and therefore lowering library yield. Indeed, their reported data showed a low library yield with high duplication rates (63% ± 13.89%) [105].

To compare our cfMethylSeq method to their cf-RRBS method, we sequenced 3 cfDNA samples with both methods. Results are shown in table 2.3. With the cf-RRBS procedure, 45 to 65% of the sequenced reads were adapter dimers and therefore wasted data. Adapter dimers are exacerbated with the cf-RRBS procedure because of the high input adapter amount needed to ligate to all fragments, even those without digestion sites. All of the samples sequenced with our method had higher on target rates compared to the same sample sequenced with the method from [105], illustrating the efficiency of our adapter ligation strategy for preserving the regions of interest.

2.2.4.3 cfMeDIP-seq

The cfMeDIP-seq procedure [89] captures all reads with at least one methylated C. This C does not necessarily have to be in a CpG context, however in mammalian DNA non-CpG context methylation is rare [55]. Initially, we hypothesized the cfMeDIP-seq procedure would miss a lot of informative tumor reads in cancer cfDNA because hypomethylated repeat regions are a hallmark of cancer [12]; if so this method would only be able to capture these reads if there was at least one methylated C in them. In our analyses on WGBS data on cfDNA, a methylated C occurs on roughly 55% of all reads, yielding only 2-fold enrichment over traditional WGBS (figure 2.6). However, over 75% of fragments that contain a CpG site contain a methylated site, meaning cfMeDIP-seq will be able to capture a large portion of the fragments with CpG sites.

While cfMeDIP-seq is able to sequence most fragments with a CpG site, the advantage of our cfMethylSeq procedure lies in the enrichment of CpG islands and gene promoters. Furthermore, cfMeDIP-seq does not contain a bisulfite conversion step. After sequencing, all that is known is that at least one CpG site on the read was methylated, but not which CpGs or how many. Unbiased methylation ratios cannot be obtained, and recent read-based algorithms cannot be used on this data [59, 25, 64]. In contrast, cfMethylSeq is able to capture all methylation states without bias at basepair resolution in CpG-dense regions.

2.2.5 UMIs are necessary in cfMethylSeq

PCR amplification is an essential library preparation step in both cfMethylSeq and RRBS. However, if multiple copies (PCR duplicates) of the same read end up getting sequenced, they must be removed, otherwise they will result in biased methylation and coverage measurements [80]. However, PCR deduplication is not recommended for RRBS because standard deduplication algorithms do not work (see Methods) [51]. While not ideal, this problem is typically ignorable in RRBS because high amounts of input DNA (e.g., from solid tissue) lead to lower PCR duplication rates. However, in cfMethylSeq, which uses low amounts of cfDNA as input, the duplication rate could possibly be very high. The addition of UMIs into

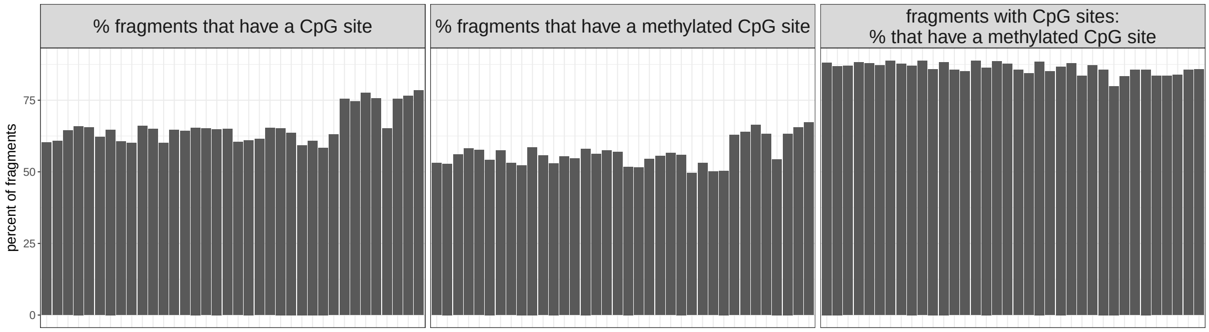

Figure 2.6: Enrichment of cfMeDIP-seq. cfDNA WGBS samples were analyzed to find the percentage of mapped fragments that contain a CpG site (A) and that contain a methylated CpG site (B). Of fragments that contain a CpG site, the percentage that contained a methylated CpG site is shown in (C).

our data enables us to truly measure duplication rates in RRBS and UMIs. As expected, we observe higher PCR duplication rates in cfMethylSeq (average 27%) compared to RRBS libraries on solid tissue (average 8%) (figure 2.7). Without the UMI, there would be no way to identify or remove these PCR duplicates in our cfMethylSeq data, which could have a large effect on methylation measurements and hinder downstream analyses.

2.3 Methods

2.3.1 Addition of UMIs

2.3.1.1 Calculations for UMI length

In standard library preparation protocols, each DNA fragment is copied several times during PCR. Sometimes multiple clones of the same molecule may end up getting sequenced, even though they represent only one initial molecule. Following bisulfite conversion, unmethylated DNA is T rich while methylated DNA is C rich. Due to differences in melting temperatures, unmethylated DNA amplifies easier than methylated DNA, leading to biases in the final library that could affect methylation measurements [110]. To get rid of this experimental artifact, PCR deduplication is performed after mapping the reads to the genome. Typical

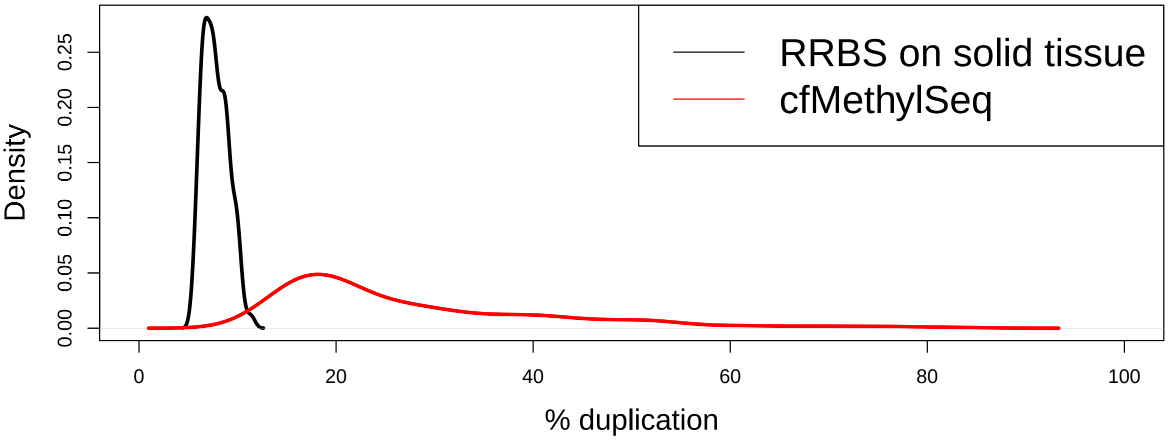

Figure 2.7: PCR duplication rates across cfMethylSeq (red) and solid tissue RRBS (black) samples. Rates were calculated using the UMIs present in our custom adapters.

programs use only the mapping position (chr, start, end, and strand) to identify PCR duplicates, and keep the highest quality fragment at each position [51]. This strategy is effective for moderately low-depth, untargeted library preparation strategies, such as WGBS, since it is unlikely that the same exact mapping location will correspond to more than one distinct molecule. For RRBS, this procedure is no longer practical, since the DNA molecules are physically cut with a restriction enzyme. For example, let the CCGG sites in the following fragments map to the same genomic location:

ACTCCGGNN...NNCCGGTCG

TACACTCCGGNN..NNCCGGT

After MspI digestion, both fragments become:

CGGNN..NNC

CGGNN..NNC

In the final RRBS library, there will be several distinct molecules that map to the same location and strand. A solution to this is the addition of unique molecular identifiers (UMIs), DNA sequences added during library preparation that uniquely tag individual molecules before they go through PCR [109]. In the final sequenced library, PCR duplicates will have the same mapping location and the same UMI, whereas distinct fragments may have the same mapping location but different UMIs.

The length of the UMI is an important parameter: longer UMIs can pose experimental challenges, but short UMIs lead to collisions; when distinct fragments map to the exact same location and share the same UMI. In general, the longer the UMI, the lower the chance of barcode collision as there will be more possible distinct UMIs.

Newman et al. [78] analyzed barcode collisions to determine if the 4 bp index barcodes used in their CAPP-seq method with error suppression (iDEs) were sufficient. A barcode collision happens when distinct fragments (i.e., not pcr duplicates) map to the exact same location (chr, start, end, strand) and share the same barcode. If this happens, one of these molecules will get marked as a PCR duplicate of the other, even though in reality it was a distinct molecule. With 4 bp barcodes, there were 256 (4^4) unique possible barcodes; the probability that any two molecules share the same barcode is $\frac{1}{256} = .0039$. They looked at cfDNA sequencing data and determined what fraction of distinct molecules in the overall data had redundant start/end coordinates. Less than 50% of the cfDNA data (and less than 10% of sheared genomic DNA) had redundant positions; so the majority of fragments were expected to be unaffected by barcode collisions. Nevertheless, they calculated what fraction of the data would be lost because of barcode collisions using the formula for determining the number of expected collisions in a hash table (n=number of molecules with the redundant positions, k=256; number of unique barcodes):

$$E(\texttt{barcode collisions}) = n - k + k(\frac{k-1}{k})^n$$

For example, if there are 3 ($n = 3$) molecules with redundant positions, and we have 256 possible barcodes, we expect $3 - 256 + 256(\frac{255}{256})^3 = 0.01170349$ collisions; in other words, $\frac{0.01170349}{3} = .0039 = .39\%$ of these fragments will have to be discarded because they match to the same position and have the same barcode; so we cannot tell if they are PCR duplicates or distinct fragments.

Even though Newman et al. [78] used the targeted CAPP-Seq method [77], the likelihood of collisions due to distinct molecules getting the same barcode is the same as for WGS. This is not true for RRBS.

In CAPP-seq, probes pull down fragments that map to certain locations (like a microarray). However, the actual fragment that gets sequenced is the original fragment, not just the part that attached to the probe. For example, let the CAPP-Seq probe attach to the sequence CAT in the following fragments:

ACGCATACG

ATTACGCATACGGGT

Even though these two fragments attached to the same CAPP-seq probe, one of them is longer than the other, so they will map to different locations. There is no chance of a barcode collision.

In RRBS, the situation is different since we are physically cutting the fragment. For example, let the CCGG sites in the following fragments map to the same genomic location:

ACTCCGGNN...NNCCGGTCG

TACACTCCGGNN..NNCCGGT

After MSPI, both fragments become:

CGGNN..NNC

CGGNN..NNC

These fragments now can have a barcode collision, even though they were originally distinct fragments that mapped to different locations.

In [78], the number of distinct molecules mapping to each distinct genomic location had to be determined to calculate the expected number of barcode collisions. For RRBS, the situation is simpler. Because MspI will cut the genome at predefined locations, we just need to know how many total genomes are in the starting material. Every genome will be cut at the same places, so we will end up with n fragments at each site if there are n genomes present. For example, if there are 1000 genomes in the starting material, we will end up with 1000 fragments mapped to the exact location that are truly distinct. We need to have enough distinct barcodes (i.e., barcodes that are long enough) so that the probability that any two of these fragments being assigned the same barcode is very low.

The human genome is $3.59x10^{-12}$g. Therefore, in 500 ng (=10^{-9} g) of DNA input, we have:

$$\frac{500x10^{-9}g}{3.59x10^{-12}g} = 139,275$$

copies of the genome. In 20 ng input, we have 5571 copies of the genome.

The percent lost can then be calculated as follows:

Say we have 139,275 genomes. Then, at any one location, we have 139,275 distinct molecules that are indistinguishable (theoretically). If we have barcodes of length t, then we have $k = 4^t$ possible barcodes (e.g., $t = 4$ leads to $k = 4^4 = 256$ distinct barcodes).

The expected number of these 139,275 (n) fragments that will be assigned the same barcode, given the number of distinct barcodes ($k = 256$), is:

$$E(\texttt{barcode collisions}) = n - k + k(\frac{k-1}{k})^n$$
$$= 139275 - 256 + 256(\frac{255}{256})^{139275}$$
$$= 139019$$

This means we will lose $\frac{139019}{139275} = 99.8\%$ of the fragments at this location (we will count as PCR duplicates even though they are distinct).

Using this formula, the percent lost for 500 ng human (139,275 genomic equivalents), 20 ng human (5571 genomic equivalents), and 500 ng mouse (170,241 genomic equivalents) for different barcode lengths is shown in figure 2.8. In order to avoid too much loss, UMIs longer than 12 bp would be ideal.

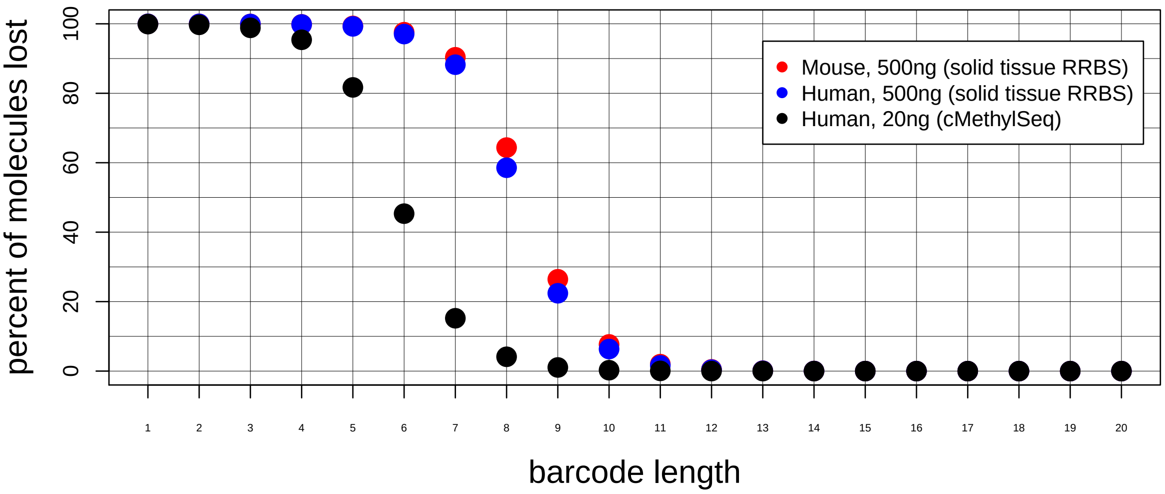

Figure 2.8: Data loss due to barcode collisions. Three different scenarios are shown for input material: 500ng mouse DNA (red), 500 ng human DNA (blue), and 20 ng human DNA (black). 500 ng input human DNA is the scenario expected for solid tissue RRBS, while 20 ng input human DNA is the scenario expected for cfMethylSeq. These input sources and amounts determine the number of genomic equivalents available. From this number, we can predict barcode collision rates: we calculate the percentage of distinct molecules (y-axis) that would be assigned the same barcode of length (x-axis) and incorrectly labeled as a PCR duplicate and removed.

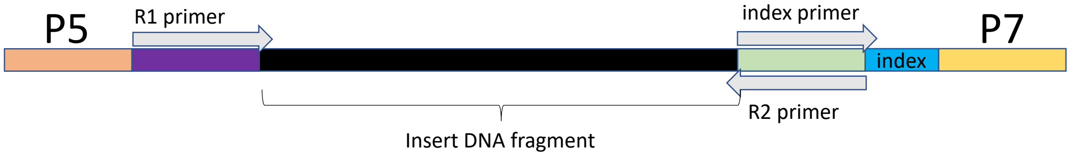

Figure 2.9: Format of a sequencing read. All components not shown in black are necessary for PCR amplification and/or sequencing on Illumina machines. The DNA insert of interest is shown in black.

2.3.1.2 Overview of adding UMIs to sequencing adapters

The overall structure of the read at the end of library preparation is shown in figure 2.9. No matter how the DNA is modified to add molecular barcodes, it still needs the overall structure to be similar to figure 2.9. Otherwise it cannot be sequenced using an Illumina machine.

Most commercial UMI adapters have the UMI in the index position, and increase the length of the index to include both the traditional index and the inserted UMI. This is not ideal, as most sequencing companies will charge prohibitive fees for sequencing longer index lengths. Instead, if the UMI is between the DNA insert and the R1 or R2 primers, no special sequencing parameters need to be used.

2.3.1.3 Final UMI addition procedure

The final barcoding strategy, based off of [48], is illustrated in figure 2.10. Two oligos (purple and green in figure 2.10) were purchased from IDT and ligated to each other. One oligo contains a random sequence of 8 nucleotides followed by a fixed sequence. After ligation, the other oligo is filled in (dNTP, ddH_2O, and Klenow exo-). In this way, the 8bp random sequence is copied to the other strand. The fixed sequence following the 8bp random sequence contains a restriction enzyme cut site (HpyCH4III, cut site 5'-ACN↓GT-3'). Following restriction enzyme digestion, this adapter now has a T-overhang with a phosphodiester bond and can ligate to DNA that has an A overhang. The reaction to ligate the adapters to the DNA involves T4 ligase buffer, T4 DNA ligase, HC, ATP, and DTT.

2.3.2 Development of final processing and analysis pipeline

The DNA libraries of both cfMethylSeq (for cfDNA) and RRBS (for tissue genomic DNA) were sequenced with 150 bp paired-end reads using HiSeqX (Illumina) by Genewiz, Inc. (South Plainfield, NJ, USA). The detailed processing pipeline is as follows:

2.3.2.1 Sequencing data UMI reformatting

Our custom adapters contain UMIs and fixed sequences in the beginnings of both R1 and R2. These sequences need to be removed before mapping. A custom script was used to remove UMIs from the beginnings of R1 and R2 and write them into the read name for deduplication later on. After library preparation with our custom adapters, the format of both R1 and R2 should be:

```
[8 bp UMI][TGACT][start of read]
```

However, due to sequencing errors or errors in adapter generation, a small percentage (<4%) of reads do not follow this format. These reads can have a UMI longer or shorter than 8bp, or be missing the fixed sequence (TGACT) entirely. Additionally, about 4% of reads in the expected format have a sequencing or other error in the fixed sequence. The initial processing script therefore first checks for the presence of TGACT in bases 9-13 of a sequenced read. If there is an exact match in both R1 and R2, the first 8 bp and fixed sequence of both R1 and R2 are written into the read names of both R1 and R2 in the format `[original read name]:R1:[UMI for R1]:R2:[UMI for R2]:F1:[fixed sequence for R1]:F2:[fixed sequence for R2]`, and bases 1-13 are removed from each read. For example, if original R1 and R2 for a read pair were as follows:

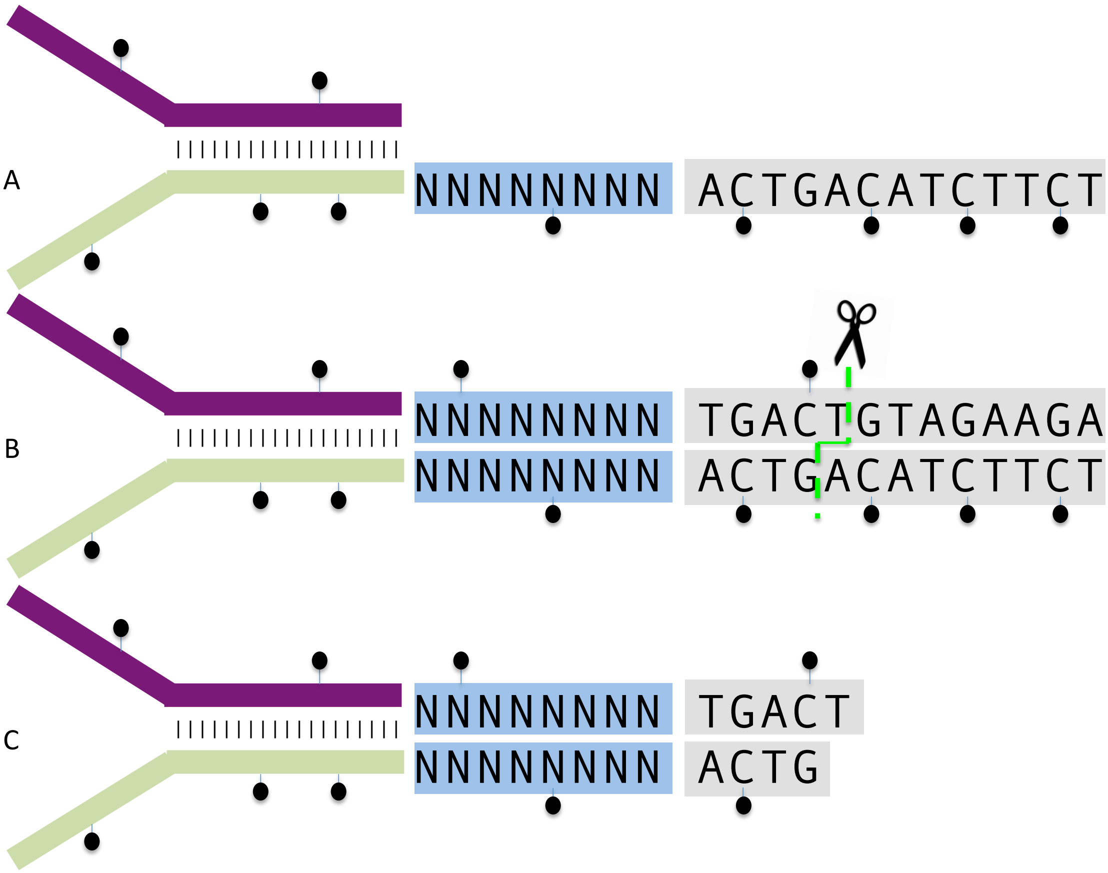

Figure 2.10: Depiction of incorporating UMIs into our sequencing adapters. (A) Two oligos are ligated to each other. These contain the standard P5 and P7 sequencing primers needed for sequencing on Illumina machines. The green oligo contains an 8 bp random sequence followed by a fixed sequence. (B) The top oligo (purple) is filled in by copying the lower strand. This incorporates the 8bp random sequence into both strands. The fixed sequence after the random barcode is then digested with a restriction enzyme. (C) After digestion, the final adapter is left with a T overhang and is ready to ligate to A-tailed library inserts.

```
@A00454:77:HH7VMDSXX:1:1101:25265:1407_1:N:0:CTTGTA+TCTTTC

ACTCCACGTGACTCGGTTTATTTTATTGGAATTGGTTAGATAGTGGGTATAGTTTATAGAGGGTGAGTTGAAGTAGGGT

@A00454:77:HH7VMDSXX:1:1101:25265:1407_2:N:0:CTTGTA+TCTTTC

AGCTATGTTGACTCAATAAAACACCAACCCACCCTACTTCAACTCACCCTCTATAAACTATACCCACTATCTAACCAAT
```

After UMI reformatting the reads would be represented as:

```
@A00454:77:HH7VMDSXX:1:1101:25265:1407_1:N:0:CTTGTA+TCTTTC:R1:ACTCCACG:R2:AGCTATGT:F1:TGACT:F2:TGACT

CGGTTTATTTTATTGGAATTGGTTAGATAGTGGGTATAGTTTATAGAGGGTGAGTTGAAGTAGGGT

@A00454:77:HH7VMDSXX:1:1101:25265:1407_2:N:0:CTTGTA+TCTTTC:R1:ACTCCACG:R2:AGCTATGT:F1:TGACT:F2:TGACT

CAATAAAACACCAACCCACCCTACTTCAACTCACCCTCTATAAACTATACCCACTATCTAACCAAT
```

If there is no exact match for the fixed sequence at bases 9-13, a close match is tolerated if the Levenshtein distance between bases 9-13 and TGACT is 1. If this condition is satisfied, the first 8 bp and fixed sequenced are written into the read name and bases 1-13 are removed, as above. If still no conditions are satisfied, UMIs of different lengths allowing for Levenshtein distance of 1 in the shifted fixed sequence are allowed. UMI lengths are checked in the following order: 7 bp, 9 bp, 1-6 bp, then 10-12 bp. These would have TGACT (or 1 mismatch) in bases 8-12, 10-14, etc. If the final UMI is shorter than 8 bp, Ns are padded to the beginning of the UMI to make the final UMI 8 bp. For example, if the read pair were as follows:

```
@A00454:77:HH7VMDSXX:1:1101:25265:1407_1:N:0:CTTGTA+TCTTTC

CCACGTGACTCGGTTTATTTTATTGGAATTGGTTAGATAGTGGGTATAGTTTATAGAGGGTGAGTTGAAGTAGGGT

@A00454:77:HH7VMDSXX:1:1101:25265:1407_2:N:0:CTTGTA+TCTTTC

TATGTTGACTCAATAAAACACCAACCCACCCTACTTCAACTCACCCTCTATAAACTATACCCACTATCTAACCAAT
```

After UMI reformatting the reads would be represented as:

```
@A00454:77:HH7VMDSXX:1:1101:25265:1407_1:N:0:CTTGTA+TCTTTC:R1:NNNCCACG:R2:NNNTATGT:F1:TGACT:F2:TGACT

CGGTTTATTTTATTGGAATTGGTTAGATAGTGGGTATAGTTTATAGAGGGTGAGTTGAAGTAGGGT

@A00454:77:HH7VMDSXX:1:1101:25265:1407_2:N:0:CTTGTA+TCTTTC:R1:NNNCCACG:R2:NNNTATGT:F1:TGACT:F2:TGACT

CAATAAAACACCAACCCACCCTACTTCAACTCACCCTCTATAAACTATACCCACTATCTAACCAAT
```

If the final UMI is longer than 8 bp, the 8 bp closest to the fixed sequence are used as the UMI. For example, if the read pair were as follows:

```
@A00454:77:HH7VMDSXX:1:1101:25265:1407_1:N:0:CTTGTA+TCTTTC

ACTACTCCACGTGACTCGGTTTATTTTATTGGAATTGGTTAGATAGTGGGTATAGTTTATAGAGGGTGAGTTGAAGTAGGGT

@A00454:77:HH7VMDSXX:1:1101:25265:1407_2:N:0:CTTGTA+TCTTTC

GATAGCTATGTTGACTCAATAAAACACCAACCCACCCTACTTCAACTCACCCTCTATAAACTATACCCACTATCTAACCAAT
```

After UMI reformatting the reads would be represented as:

```
@A00454:77:HH7VMDSXX:1:1101:25265:1407_1:N:0:CTTGTA+TCTTTC:R1:ACTCCACG:R2:AGCTATGT:F1:TGACT:F2:TGACT

CGGTTTATTTTATTGGAATTGGTTAGATAGTGGGTATAGTTTATAGAGGGTGAGTTGAAGTAGGGT

@A00454:77:HH7VMDSXX:1:1101:25265:1407_2:N:0:CTTGTA+TCTTTC:R1:ACTCCACG:R2:AGCTATGT:F1:TGACT:F2:TGACT

CAATAAAACACCAACCCACCCTACTTCAACTCACCCTCTATAAACTATACCCACTATCTAACCAAT
```

If still no conditions are satisfied, bases 1-8 are written into the read name and bases 9-13 are removed. R1 and R2 are processed separately, i.e., R1 can have a 9 bp UMI and R2 could be missing the fixed sequence, but the string added to the read name of both R1 and R2 will be the same. On average ≥96% of reads have a UMI of 8 bp with an exact fixed sequence (92%) or with 1 mismatch in the fixed sequence (4%).

2.3.2.2 Sequencing data trimming

Trim galore v0.4.4 [8] was used to trim the default Illumina adapters from the sequencing reads after UMI reformatting. If a read was adapter trimmed, an additional 13 bp were removed from its 3' end to remove the 8 bp UMI and 5 bp fixed sequence from the 3' adapter. During library preparation, MspI digestion is performed followed by end repair using unmethylated cytosines. These two bp need to be removed from the beginning of R2 and potentially from the ends of R1 before methylation calling [7]. Therefore an additional 2 bp were trimmed off the 3' ends of R1 and 2 bp were trimmed off of the 5' ends of R2 to remove this spurious methylation call. The trim galore command used was:

```
--three_prime_clip_R1 15 --three_prime_clip_R2 13 --clip_R2 2 --length 15 --phred33
```

2.3.2.3 Alignment, deduplication, and methylation calling

UMI-reformatted, trimmed sequences were aligned to the hg19 reference genome (GRCh37 Genome Reference Consortium Human Reference 37 (GCA_000001405.1) [13]) using Bismark v0.18.2 [51] with default parameters. Umi-Grinder v0.0.1 [50] was used to remove PCR duplicates based on the UMIs written into the read names during the UMI reformatting step and the mapping location. The original Umi-Grinder program was modified to count N's as matches rather than the default mismatches. Due to the length of the UMI and the rate of sequencing errors observed in the fixed sequence, 4 bp were allowed to mismatch in the 16 bp UMI (8 bp from R1 + 8 bp from R2) in order for a duplicate to be counted. For the mitochondrial chromosome, this requirement was lowered to 0 mismatches due to memory constraints. In our experience, the number of mismatches allowed made little difference; around 0.5% more mapped reads were counted as duplicates when the mismatch threshold was increased from 0 to 4 mismatches.

Methylation calls from the deduplicated bam files were extracted using bismark methylation extractor using the `-p` option. Bed files were created for the deduplicated bam files using `bamtobed` from the bedtools suite v2.26.0 [83]. In the bed file produced by `bamtobed`, each mapped, deduplicated fragment is represented by two lines: one for R1 and one for R2. These two lines were combined to create a fragment level bed file. To combine R1 and R2 into one fragment, the mapping locations and strands of R1 and R2 are compared. The mapping location of the final fragment is the lowest of the start positions for R1 and R2 and the highest of the end positions for R1 and R2. The final fragment maps to the same strand as R1.

For example, if there were two lines for fragment

```
GWNJ-0850:R1:ACCATGAG:R2:ACAGAATC:F1:TGACT:F2:TGACT
```

in the original bed file:

```
chr1 14889 15010 GWNJ-0850:R1:ACCATGAG:R2:ACAGAATC:F1:TGACT:F2:TGACT/2 32 +

chr1 14968 15090 GWNJ-0850:R1:ACCATGAG:R2:ACAGAATC:F1:TGACT:F2:TGACT/1 32 -
```

These would be represented as the following one line in the collapsed bed file:

```
chr1 14889 15090 GWNJ-0850:R1:ACCATGAG:R2:ACAGAATC:F1:TGACT:F2:TGACT 32 -
```

Similarly, methylation calls were gathered at the fragment level from Bismark's output files. Bismark provides two output files for CpG methylation calls, one for fragments mapping to the positive strand and another for fragments mapping to the negative strand. There is a line for every CpG site called on each read. For example, the fragment above had 8 methylation calls. It is represented in Bismark's CpG file for the negative strand (CpT_OB*) as:

```
GWNJ-0850:R1:ACCATGAG:R2:ACAGAATC:F1:TGACT:F2:TGACT + 14948 Z
GWNJ-0850:R1:ACCATGAG:R2:ACAGAATC:F1:TGACT:F2:TGACT + 14955 Z
GWNJ-0850:R1:ACCATGAG:R2:ACAGAATC:F1:TGACT:F2:TGACT + 14976 Z
GWNJ-0850:R1:ACCATGAG:R2:ACAGAATC:F1:TGACT:F2:TGACT + 15005 Z
GWNJ-0850:R1:ACCATGAG:R2:ACAGAATC:F1:TGACT:F2:TGACT - 15046 z
GWNJ-0850:R1:ACCATGAG:R2:ACAGAATC:F1:TGACT:F2:TGACT + 15029 Z
GWNJ-0850:R1:ACCATGAG:R2:ACAGAATC:F1:TGACT:F2:TGACT + 15090 Z
GWNJ-0850:R1:ACCATGAG:R2:ACAGAATC:F1:TGACT:F2:TGACT + 15086 Z
```

The second and fourth columns are + and Z, respectively, if the CpG was methylated and -, z if the CpG was unmethylated. These lines are collapsed into one line with three columns; the read name, the CpG locations (sorted), and the methylation string (in the same order as the CpG locations), with 1 representing a methylated CpG and 0 an unmethylated CpG:

```
GWNJ-0850:R1:ACCATGAG:R2:ACAGAATC:F1:TGACT:F2:TGACT 14948,14955,14976,15005,15029,15046,15086,15090 11111011
```

The fragment level bed and collapsed CpG file are then combined using unix join on the read name to create an annotated, fragment level bed file which contains the methylation calls on an individual sequenced fragment. For example, the output line for this example read would be:

```
chr1 14889 15090 GWNJ-0850:R1:ACCATGAG:R2:ACAGAATC:F1:TGACT:F2:TGACT 201 - 8 14948,14955,14976,15005,15029,15046,15086,15090 11111011
```

The columns of the annotated bed file are chromosome, fragment start, fragment end, fragment name, fragment length, strand, number of CpGs on the fragment, mapping locations of the CpGs, and the methylation string. Note that standard bed files are 0-based and the last bp is not included ("0 start, half-open"), e.g. chr1:4-5=T [103]. These custom bed files are 0-based but the last bp is included ("0 start, fully-closed") (e.g. chr1:4-5=TC). To change to a standard bed format, 1 bp needs to be added to the third column. The coordinates of the CpG sites in the eighth column are with respect to the mapped strand, i.e. in a negative mapped fragment they will need to be shifted down 1 bp to match the coordinates of a positive mapped fragment mapping to a similar location. This annotated, fragment level bed file was used for all downstream analyses.

2.3.3 cfMethylSeq enrichment analysis

2.3.3.1 Feasibility simulations

WGS cfDNA samples from [90] were in silico digested with MspI to evaluate the feasibility of performing RRBS on cfDNA. Specifically, paired end WGS cfDNA reads were merged into one fragment as described above so that the overall fragment length could be obtained. These reads were stored in a bed file, with one fragment per line. Then, any cfDNA fragment that covered two or more CCGG sites was retained. Only the portion of the original fragment between CCGG sites was kept. Any resulting fragment that was over 350 bp in length was removed. For these simulations, promoter and CGI regions were taken from [24] and lifted over to the hg19 genome [49]. MspI digestion sites were defined from CCGG matches in the human reference genome hg19 (GRCh37 Genome Reference Consortium Human Reference 37 (GCA 000001405.1))

2.3.3.2 Definition of genomic regions

CGI regions were downloaded directly from UCSC table browser [47], without masking. CGI shores were defined as 2000 bp flanking regions of CGIs, CGI shelves as the 2000 bp flanking regions beyond CGI shores, and CGI seas as every other region of the genome that was not a CGI, shore, or shelf. Gene promoters were defined from Gencode release v19 [29] and expanded to 1000 bp upstream and downstream, following procedures in [24]. Exons were extracted from the Gencode release. Repeat regions were downloaded from UCSC table browser [47].

2.3.3.3 Coverage calculation in genomic regions

A fragment was considered as covering a genomic region if any portion of the fragment overlapped with that genomic region. Specifically, a bed file representing the mapped fragments of a sample (after collapsing R1 and R2 into one fragment) was intersected with a bed file representing the region of interest using bedtools [83] using the `-u` option and the genomic region bed file as file `B`. Any fragments surviving this intersection were considered mapped to the genomic region of interest.

2.3.3.4 Methylation and coverage correlation calculations

Coverage and methylation comparisons between pairs of samples were performed at CpG sites only. To compare coverage and methylation at a certain depth of coverage, both samples were first subsetted to only CpG sites covered by both samples at the specified depth. For example, to compare methylation between a solid tissue sample's sheared genomic DNA's cfMethylSeq data and traditional RRBS data at 10x coverage, CpG sites that were covered 10 or more times in both samples were used. The methylation values and coverage (read depth) at these CpG sites were used to compute the correlation. For methylation, two vectors, one from each sample, were used to compute the correlation. Each vector's length was the number of CpG sites covered by both samples, and the entries were methylation ratios. For coverage correlation, the entries in the vector were read depths. Pearson correlation was used in all cases.

2.4 Discussion

Here, we outlined a novel experimental method, cfMethylSeq, to profile methylation in cfDNA at dramatically reduced cost. Our cfMethylSeq procedure yields 12.8 fold enrichment over WGBS in terms of CGI coverage, reliably profiles methylation measurements, and is reproducible. This protocol, as well as the dataset we have built using it, has the potential to greatly expand methylation analysis and biomarker discovery in cfDNA.

Our novel cfMethylSeq procedure, specifically designed to capture cfDNA fragments with two MspI cut sites, is broadly generalizable to capturing restriction enzyme digests on fragmented DNA. This could be beneficial for selecting cut fragments without having to go through a size selection step, which can be inefficient [39]. Currently, reliable methylation profiling in cfDNA is only commercially available through WGBS or targeted methylation sequencing. Our procedure strikes a balance between these two extremes: much less data is wasted compared to WGBS, and sequencing can be done at much higher coverage, inexpensively. In addition, cfMethylSeq still covers a large portion of the genome, allowing for de novo target identification, as well as further use of the data for other analyses. Our cfMethylSeq procedure also includes the development of an in-line UMI for methylation sequencing, something that is not commercially available [95].

Because our cfMethylSeq procedure does not require selection of targets a priori, our procedure can be used for biomarker discovery. This is not possible if a targeted panel [117, 64] is used. Methylation has been increasingly implicated in biological processes and diseases. Our procedure potentially allows for noninvasive monitoring and biomarker discovery in a cost-effective, large-scale manner. cfMethylSeq also allows for genome-wide attributes, such as copy number changes, to be measured for no additional costs.

Recently, two other approaches attempted to reduce methylome profiling costs in cfDNA: cf-RRBS and cfMeDIP-seq. The cf-RRBS method follows a similar procedure as our cfMethylSeq protocol, but their single-end blocking approach does not reach as high of an enrichment rate as our dual-end blocking approach. The cfMeDIP-seq method uses immunoprecipitation to pull down cfDNA fragments with at least one methylated CpG site, but this scenario occurs on ∼50% of all cfDNA fragments, therefore cfMeDIP-seq does not effectively enrich CpG-rich fragments. In addition, cfMeDIP-seq does not yield methylation measurements at basepair resolution, prohibiting highly sensitive read-based approaches [59, 25]. For approaches using targeted methylation [64, 117], panels need to be established a priori, and data cannot be used for marker discovery. None of these existing approaches have incorporated UMIs to facilitate single-molecular counting.

CHAPTER 3

Application of cfMethylSeq for cancer detection and typing

3.1 Introduction

Detecting malignancies before their metastasis is key to the fight against cancer. Recently, cfDNA has drawn much attention for accomplishing this task. Because cfDNA can be accessed through a blood draw, cancer status could potentially be monitored noninvasively through so-called liquid biopsies [16]. Despite its promise, major challenges of cfDNA-based cancer detection include: (1) the fraction of tumor DNA in the blood of early-stage cancer patients can be very low, (2) the molecular heterogeneity of cancer (e.g. diverse subtypes, stages, and etiologies), and (3) sample sizes that are too small to reflect the heterogeneous patient population (e.g. age, gender, and ethnicity).

Due to the low amount of tumor DNA present in the blood, a successful cancer detection method should capture as many tumor-derived fragments as possible. Some current studies aim to do this through small, panel-based approaches that deeply sequence areas likely to contain tumor cfDNA fragments [64, 14, 117]. These studies often result in false negatives because only a small proportion of tumor-derived fragments are observed. Alternatively, genome-wide approaches have also been used, aiming to capture broad signals to make up for the lack of individual tumor-derived signals [99, 59, 46, 15]. However, whole-genome sequencing is cost prohibitive for clinical use.

The heterogeneous nature of cancer hints that a noninvasive test able to capture diverse attributes of cancer has the best chance of success. Indeed, several cfDNA features have been shown to have diagnostic power, including cfDNA methylation [64, 59, 46, 25, 117, 89], fragment length [15, 42, 90], copy number variation (CNV) [12, 15], and microbiome composition [82]. Below we outline the use of these features:

1. Fragment length and copy number

   Broad copy number changes are commonly observed in several types of cancer [43]. These changes can be detected in cfDNA with very shallow sequencing [12, 33, 15]. In noninvasive prenatal testing, already a clinical reality, a similar technique is used to detect chromosomal abnormalities in the fetus, such as down syndrome. The situation is more complicated in cancer, however, because copy number changes can occur on variable chromosomes and can be much smaller than the whole chromosome [43].

   Fragment size has increasingly been recognized as a unique feature in cfDNA [44]. cfDNA fragments display characteristic fragment lengths around 165 bp, with multiples of 165 bp also being observed. This length corresponds to the length of DNA that would wrap around a nucleosome (147 bp) plus the linker DNA, leading to the theory that cfDNA fragments reflect apoptotic fragmentation [10]. Because nucleosome positioning is related to gene regulation and is cell-type specific, cfDNA fragment length profiles have been used to infer gene expression and tissue of origin [104, 90] and even reconstruct Hi-C maps [66]. Overall, the presence of shorter or longer cfDNA fragments has been used to successfully detect cancer [42, 15, 97], and fragment length analysis has yielded important insights into cfDNA's release into the bloodstream [44].

2. Microbial profiles

   Unique microbial signatures have been found in tissue and blood for most major types of cancer [76]. Although the presence of these signatures is poorly understood, possibly coming from live microorganisms, host cells, or lysed bacteria in the tumor microenvironment, Poore et al. [82] demonstrated the applicability of microbiome profiles for cancer detection in cfDNA. To use these methods, the sample of interest is sequenced, and reads that do not map to the human genome are then attempted to map to various microbial genomes. Consequently, sample contamination is a concern. However, an advantage is that this type of analysis can be performed retrospectively on current cfDNA sequencing data originally generated for other purposes.

3. Methylation

   Targeted tumor suppressor gene hypermethylation and broad, genome wide hypomethylation have been observed in cancer [43]. cfDNA methylation profiling in a genome-wide manner has been able to detect large scale methylation changes [12], and has been used to infer tissue of origin [42]. Because methylation changes occur early in tumorigenesis, methylation based liquid biopsies hold great promise [59]. Recent methylation-based cfDNA detection methods typically use either targeted methylation panels, where desired targets are found in solid tissue data and then sequenced deeply in the cfDNA [117, 64], or WGBS with sophisticated algorithms to compensate for low coverage [46, 59, 25]. Read level methylation analyses, such as α value [59] and methylation haplotype [25], have enhanced the power of these methods.

Although these methods all show promise for detecting cancer in cfDNA, to our knowledge they have not all been integrated into one model. Ensemble machine learning is a technique that boosts accuracy by combining multiple learning algorithms. These methods have gained popularity in bioinformatics because of their ability to deal with high dimensional features and small sample size [118]. An ensemble model reduces the chance of overfitting by combining multiple classifiers trained on the same data, using the training data in a more efficient way.

Ensemble methods improve classification by combining a group of layer 1 (base) classifiers in some way. This is because different layer 1 features may capture different aspects of the training data; when diverse and accurate layer 1 features are combined, the overall accuracy often increases over any layer 1 feature used individually. A straightforward example would be a facial recognition algorithm, where layer 1 classifiers detect different components such as ears, eyes, etc. As individual classifiers they will not perform well at detecting a face, but together they will have high accuracy. A drawback of these methods, however, is decreased interpretability.

These integrative methods will offer the highest gains when the layer 1 features are diverse. In this way, if one classifier makes a misclassification on a test sample, another classifier using complementary information about the same sample could be able to classify it correctly. When combined, these features compensate for each other.

Simple ensemble learning methods combine layer 1 classifiers through rudimentary metrics, such as voting or averaging. For example, each base classifier could vote cancer or noncancer for a test sample, and the majority vote wins. In more sophisticated stacked classifiers, the layer 1 model outputs are used to train a second meta-classifier [116]. The output of this layer 2 classifier is then the final prediction.

Some cfDNA-based cancer detection methods have combined multiple feature types to come to a cancer/noncancer classification. DELFI [15] uses both copy number and fragmentation profiles in their stochastic gradient boosting model. Chan et al. [12] used hypomethylation and copy number to classify a sample as cancer or noncancer. In the former case, all features were input into one model. In the latter case, individual classifiers were trained for each feature, and both "AND" and "OR" algorithms were tried, where the the sample was classified as cancer if it was classified as cancer by both models, or by either model, respectively.

However, different requirements of library preparation and sequencing depth have so far prevented the comprehensive integration of diverse cfDNA features. While CNV, fragment length, and microbiome composition can be obtained from shallow whole-genome sequencing, DNA methylation requires its own sequencing modality, e.g. WGBS, which is expensive, despite the fact that methylation-informative cytosines are primarily found in CGIs, occupying less than 5% of genome. The RRBS method employs restriction enzymes to cut intact DNA into small fragments in regions with a high CpG content, and subsequently size-selects these small fragments to enrich for CpG sites, therefore presenting a cost-effective approach for genome-wide methylation profiling. However, as described in chapter 2, the conventional RRBS approach is not applicable to naturally fragmented cfDNA.

In order to address the necessarily large sample size, a successful cancer detection test should be able to continually adapt as sample size increases. This is not possible with fixed panel-based approaches, where additional testing samples are only measured on the predefined sets of markers. These new samples cannot be used to discover or validate new markers in later research. Ideally, these precious patient samples would be profiled in a non-targeted manner, such that novel information can be gained for later test refinement.

To address all these challenges, this chapter presents an integrated experimental and computational system, named CancerRadar, for the accurate and affordable detection of cancer. CancerRadar includes (1) using our cost-effective experimental approach presented in chapter 2, cfMethylSeq, to comprehensively profile diverse cfDNA features, and (2) an integrative computational learning framework for cancer detection, which is scalable to many types of features and high numbers of markers.

We applied CancerRadar to the cfMethylSeq profiles of 479 individuals, comprised of 275 patients with colorectal, liver, lung, or stomach cancer and 204 non-cancer individuals (including patients with various diseases besides cancer), where $> 55\%$ of patients are from early stages (stages I and II), and for liver cancer 74% of patients are from stage I. Exploiting the vast amount of public data (TCGA, GEO, Epigenome Roadmap) and our newly generated data on solid tumors, we demonstrate the performance of CancerRadar in two tasks: (1) Detecting cancer: CancerRadar achieved a sensitivity of 89.1% at the specificity of 97% (i.e., one false positive), ranging from 75.1% to 96.2% sensitivity among the four cancer types, with an overall AUC of 0.986 (with standard deviation 0.011) in the independent validation set. (2) Locating cancer: the prediction of the tumor tissue of origin (TOO) yielded an accuracy of 91.5% (with standard deviation 4.7%) in the independent validation set. Our results demonstrated the strong complementary effect of the multiple feature types. Encouragingly, our data show that as training sample sizes increase, the detection power of CancerRadar continues to increase. Furthermore, we show that more markers are required to achieve the highest power as sample sizes grow, testifying to the importance of an expandable test. Since cfMethylSeq profiles cfDNA methylation in a non-targeted, genome-wide manner, current and additional samples can be used to expand and refine our test.

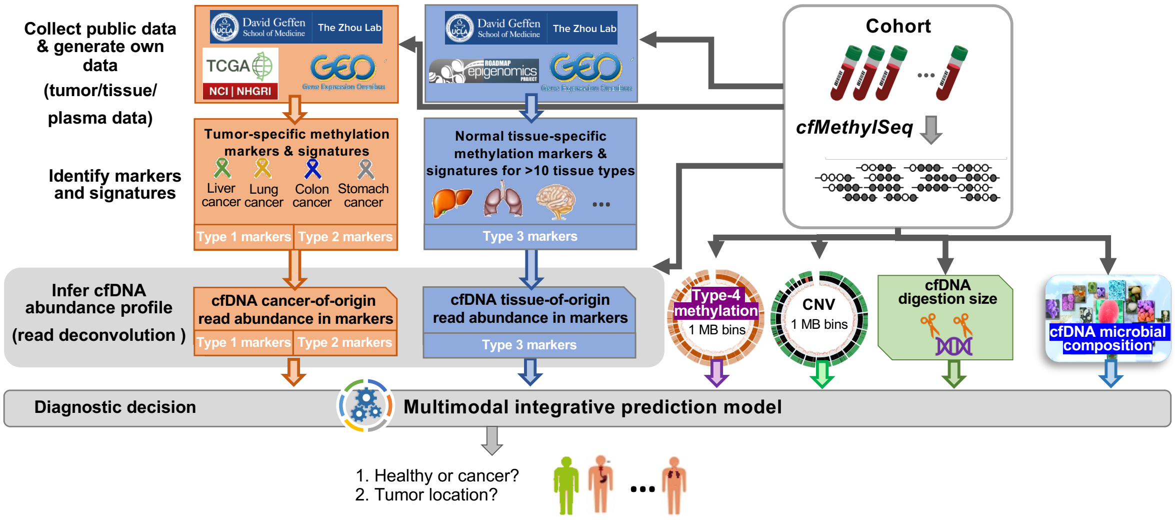
Collect public data & generate own data (tumor/tissue/ plasma data)
David Geffen School of Medicine
The Zhou Lab
TCGA
NCI | NHGRI
GEO
Gene Expression Omnibus
Roadmap epigenomics project
Cohort
cfMethylSeq
Identify markers and signatures
Tumor-specific methylation markers & signatures
Liver cancer
Lung cancer
Colon cancer
Stomach cancer
Type 1 markers
Type 2 markers
Normal tissue-specific methylation markers & signatures for >10 tissue types
Type 3 markers
Infer cfDNA abundance profile (read deconvolution)
cfDNA cancer-of-origin read abundance in markers
cfDNA tissue-of-origin read abundance in markers
Type-4 methylation
1 MB bins
CNV
1 MB bins
cfDNA digestion size
cfDNA microbial composition
Diagnostic decision
Multimodal integrative prediction model
1. Healthy or cancer?
2. Tumor location?

Figure 3.1: Flowchart of using multi-feature data for cancer detection and TOO prediction

3.2 Results

3.2.1 Multi-feature profiling using cfMethylSeq

479 plasma samples were sequenced with the cfMethylSeq protocol described in chapter 2 (see figure 2.1a). From the cfMethylSeq data we extract four types of cfDNA features for cancer detection: DNA methylation, copy number variation, digestion size, and microbiome signatures (figure 3.1), as detailed below:

3.2.1.1 DNA Methylation

Using cfMethylSeq data from a set of non-cancer individuals, our own RRBS data on solid tumors and normal tissues, and publicly available solid tumor and normal tissue sequencing and 450k data from the Cancer Genome Atlas (TCGA) [114], epigenome roadmap [52], and the Gene Expression Omnibus (GEO) [5], we identified four types of DNA methylation markers:

1. **β-value-based tumor markers (Type 1):** These are regions where the average methylation rate (i.e. β-value, $\beta=\frac{\#methylated\ CpGs\ over\ all\ mapped\ reads\ in\ region}{\#CpGs\ over\ all\ mapped\ reads\ in\ region}$) significantly differs between solid tumors and their adjacent normal tissues, as well as between solid tumors and the cfDNA of people without cancer. These regions are found separately for each cancer type, then the union is taken. Using our RRBS data of 101 solid tumor and adjacent normal pairs and cfMethylSeq data of 41 non-cancer cfDNA samples, we identified 41,494 Type 1 markers in total, including 22,416 hypomethylated and 19,077 hypermethylated markers in five cancer types (colon adenocarcinoma (COAD), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and stomach adenocarcinoma (STAD)). In the same way, using 450K array data from TCGA of 298 solid tissues [114] and our cfMethylSeq data of 41 non-cancer cfDNA samples, we also obtained 20,179 Type 1 markers in total, including 6,668 hypomethylated and 13,511 hypermethylated in tumor.

2. **α-value-based tumor markers (Type 2):** In contrast to Type 1 tumor markers that compare population-based average methylation rates (β-values) in a region, Type 2 markers compare the methylation rates of individual sequencing reads, so called α-values [59] (i.e, $\alpha=\frac{\#methylated\ CpGs\ in\ a\ mapped\ read}{\#CpGs\ in\ a\ mapped\ read}$). Specifically, these are regions in which the majority of non-cancer cfDNA samples show consistent α-values (i.e. nearly all reads are hyper or hypomethylated), while the majority of reads of at least one tumor tissue sample show clear opposite α-values. Using our RRBS data of solid tissues and cfMethylSeq data of cfDNA samples, we identified 41,493 Type 2 markers, including 33,871 hypomethylated and 8,220 hypermethylated markers in tumor. Since this process uses the methylation information of individual sequencing reads, Type 2 markers cannot be identified with array-based methylation data.

3. **β-value-based tissue markers (Type 3):** Since organs containing tumors undergo increased cell death and therefore yield an elevated quantity of cfDNA [119, 4, 28], tissue-specific cfDNA deconvolution can aid our inference in both detecting cancer and predicting its tissue-of-origin. The Type 3 markers are genomic regions where

the average β-value differentiates not only between one tissue type and all other tissue types (i.e., one-vs-rest for tissue-type-specific markers), but also between pairs of tissue types (i.e., one-vs-one for tissue-type-pair comparison markers). Using RRBS data of 17 tissue types from our own data and public sources [5], we identified 17,672 tissue markers. Separately, using TCGA Illumina 450K array data [114], we identified 9,818 array-based tissue markers for 11 tissue types.

4. **β-value-based 1 MB bins (Type 4):** Equal-sized 1MB bins are used as markers. In these markers, we aim to capture the hypomethylation of repeated DNA sequences, a hallmark of cancer [12, 87]. The β-value averaged across all repeat regions in a 1 MB bin is used as the feature value.

Deconvoluting tumor methylation signals: Given the cfMethylSeq data from a cfDNA sample, we either employed a probabilistic mixture model (for Type 1 and 3 markers) or directly compared the read-level α-value (for Type 2 markers) to deconvolute the tumor-derived or tissue-specific reads falling onto each marker region. For Type 1 and 3, reads falling in each marker region were assigned probabilities of coming from the tumor (Type 1) or tissue (Type 3) backgrounds. This read-based deconvolution exploits the pervasiveness of DNA methylation for signal enhancement. Here, we improved our previous read-level deconvolution algorithm [59] by (1) expanding 2-class to k-class ($k \geq 2$) for deconvoluting reads specific to k types of tissues; and (2) adding a new unknown class to collect reads that do not fit into any known classes (details in Methods). For Type 1 and 3 markers, we construct a profile vector where the length of the vector is the number of markers and the value in each entry is the normalized count of tumor-derived (tissue-derived) reads. In contrast, the Type 2 markers are designed such that the α value is overwhelmingly consistent in normal cfDNA samples; any observed read with the opposite α value signal is considered a tumor read.

3.2.1.2 cfDNA digestion sizes

cfDNA fragment size is recognized as a sensitive marker for cancer detection [15, 74, 42, 32]. Although our cfDNA fragments are digested by MspI during the cfMethylSeq procedure, the lengths of original cfDNA fragments impact the lengths of digested fragments, and hence we observed a high correlation between these two length entities. Here, we calculate the average length of all cfDNA fragments that fall in a 1MB bin, with subsequent normalization using non-cancer reference samples. A profile vector is constructed covering 2,734 bins (after removing chrX, chrY, chrM and bins with no mapped reads) across the genome.

3.2.1.3 Copy number variation (CNV)

The fraction of reads falling in each 1MB region was used as a feature, leading to a feature profile of length 2,734 (after removing chrX, chrY, chrM, and bins with no mapped reads).

3.2.1.4 Microbial composition (Microbial)

The profile of microbial abundances is of length 1,620 (including genomes from 1,017 bacteria, 1 archaea, 453 fungi, and 149 viruses). cfMethylSeq reads that were not able to map to the human genome were used to do this analysis. The abundance of a microbe is calculated as the count of these reads that uniquely map to its microbial genome, divided by the total number of the sequencing reads in the sample and the size of the microbial genome.

3.2.2 Multimodal predictive model integrating heterogeneous and multiscale signal types

To maximize the diagnostic power of these heterogeneous features, we developed a multi-view stacked model with two-layer learning. In Layer 1, a predictive model is learned using each individual feature type. In Layer 2, the predictions of Layer 1 models are “stacked” into an ensemble model (figure 3.2). Ensembles of multiple Layer 1 models ward off against overfitting and add a form of regularization. Even if some Layer 1 models have weak pre-

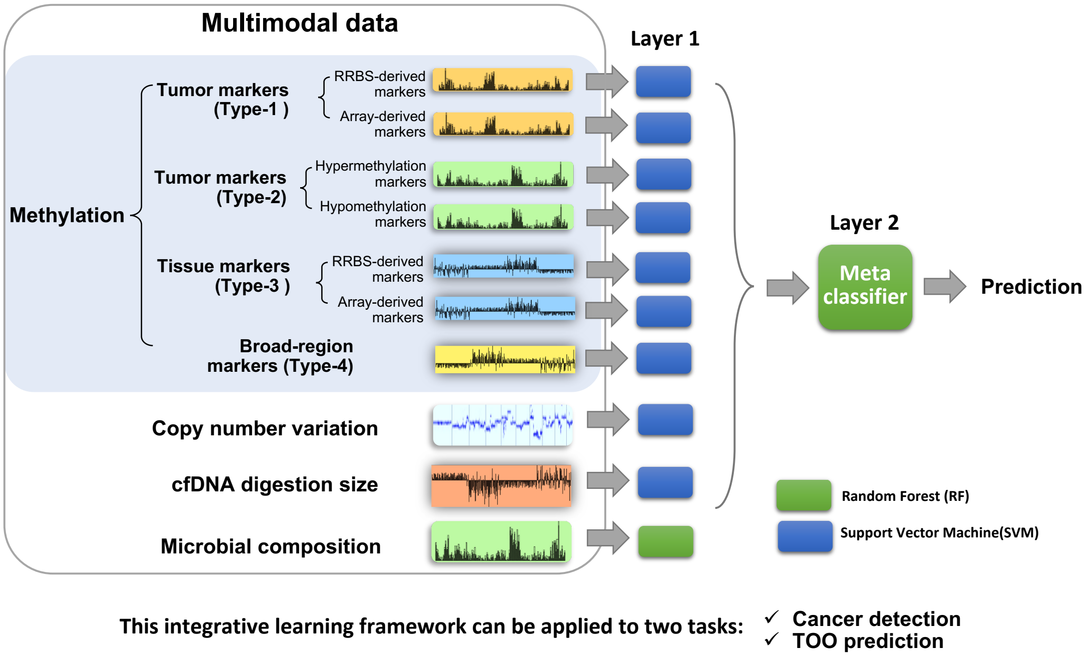

Figure 3.2: Flowchart of the integrative learning framework for two tasks: (1) cancer detection, and (2) TOO prediction. Task (1) used binary classifiers, and Task (2) used multiclass classifiers.

diction power, they may still contribute to improving accuracy by providing complementary information. We developed two stacked models, one for cancer detection and another for cancer typing. The architecture of the stacked classifier is designed to fully utilize the training samples and also avoid overfitting (see figure 3.21 in the Methods section for an in depth flowchart).

3.2.2.1 Model performance in cancer detection

The total 479 cfDNA samples sequenced with the cfMethylSeq procedure (from 42, 126, 67, 40 patients of liver, lung, colon, and stomach cancer, respectively, as well as from 204 non-cancer individuals) were split into four sets (figure 3.3): 41 non-cancer samples for marker discovery, 30 non-cancer samples for data standardization and age adjustment, 75% and 25% of the remaining 408 samples for leave-one-out cross validation (LOOCV) and independent validation (IV), respectively (Methods). For a robust performance evaluation, we repeated

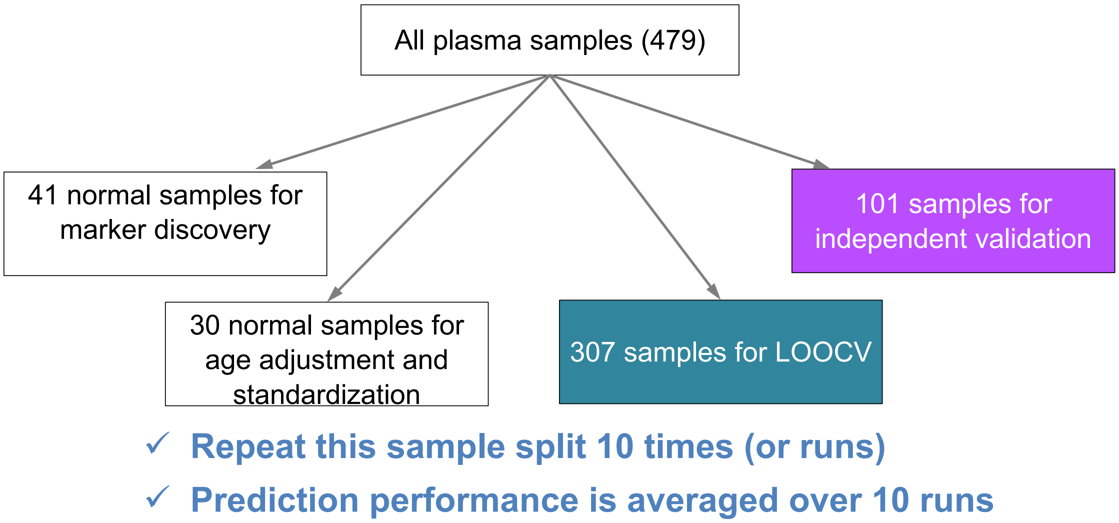
All plasma samples (479)
41 normal samples for marker discovery
30 normal samples for age adjustment and standardization
307 samples for LOOCV
101 samples for independent validation
✓ Repeat this sample split 10 times (or runs)
✓ Prediction performance is averaged over 10 runs

Figure 3.3: Overview of how the plasma samples in the cohort are used. All plasma samples are randomly split into four sets: marker discovery, age adjustment and standardization, LOOCV, and independent validation. This sample split is repeated 10 times (i.e., runs) and the prediction performance is averaged over 10 runs. The LOOCV set uses 1 sample in the LOOCV set as the test sample and the rest of the samples in this set as training samples. The independent validation set uses all samples in the LOOCV set as the training samples and all samples in the independent validation set as the test samples.

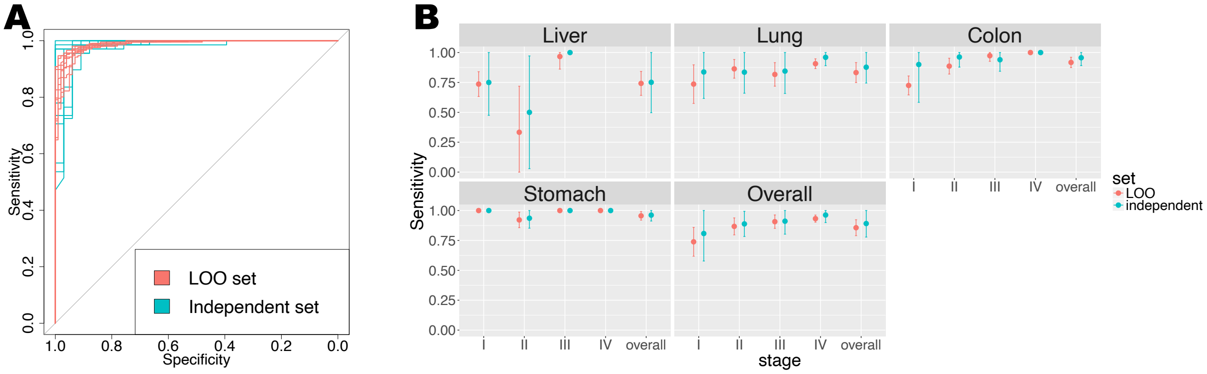

Figure 3.4: Cancer detection performance. (A) ROC curves of our method in both LOOCV and independent validation. (B) Sensitivity breakdown in each cancer stage and cancer type. Sensitivity is shown at a false positive rate of 1 (99% specificity in LOOCV, 97% specificity in independent validation). Results are not shown for stage IV liver cancer due to the small number of samples.

this split scheme 10 times and reported their average prediction performance. Our integrative prediction model achieved the average AUROC 0.989 (with standard deviation 0.003), yielding an average sensitivity of 85.6% (with standard deviation 6.7%) at the specificity of 99%, across 10 LOOCV runs (figure 3.4). This result is comparable with that of the independent validation cohort, i.e., the average AUROC 0.986 (with standard deviation 0.011) with the average sensitivity 89.1% (with standard deviation 11.3%) at specificity 97% (one falsely classified sample), over 10 runs (figure 3.4a). For non-metastatic (stages I-III) samples, our model achieved average AUROC 0.988 (with standard deviation 0.003), with the average sensitivity of 83.7% (with standard deviation 7.9%) at specificity 99%. In the independent validation cohort, non-metastatic samples achieved an average AUROC 0.984 (with standard deviation 0.013) with average sensitivity 87.4% (with standard deviation 12.7%) at specificity 97% over 10 runs. The stage-specific performance for individual cancer types is shown in figure 3.4b.

Stage	LOOCV	Independent	cancerSEEK
I	25.0	8.0	5
II	3.0	1.0	19
III	4.6	1.4	20
IV	3.0	1.0	0
overall	35.6	11.4	44

Table 3.1: Number of liver cancer samples by stage breakdown in our samples vs cancerSEEK's. Numbers are averaged over 10 runs for our LOOCV and independent sets.

Comparison to cancerSEEK cancerSEEK [14] is a published cfDNA cancer detection algorithm which uses proteins and targeted mutations to screen for eight common cancer types and identify tissue of origin. The multianalyte test uses a targeted panel to profile mutations in common cancer genes and measure levels of eight proteins. Their cohort consisted of 1005 cancer samples (stages I-III) and 812 healthy samples. They achieved a sensitivity of 62% at >99% specificity (805 out of 812 healthy correct; 99.138% specificity).

While we achieved average AUROCs of 0.989 in the LOOCV set, and 0.986 in the independent set, cancerSEEK achieved an average AUROC of 0.91 (figure 3.5a). To compare our results at the same specificity (99%) it is only possible to use our LOOCV set due to the limited number of normal samples in our independent set. Our LOOCV set achieved higher sensitivity at 99% specificity for stages I, II, III, and overall (cancerSEEK did not use any stage IV samples) (figure 3.5b). In terms of individual cancer types, our LOOCV set achieved superior sensitivities for all cancer types we profiled, except for liver cancer (figure 3.5c). However, the majority of our liver cancer samples are stage I, whereas very few liver cancer samples profiled in cancerSEEK were stage I (table 3.1). cancerSEEK achieved such a high sensitivity for liver cancer because only one stage III liver cancer sample was misclassified.

Comparison to GRAIL GRAIL developed a targeted methylation panel for their pan-cancer classifier [64]. Targets were identified using cfDNA WGBS, 450k, and solid tissue

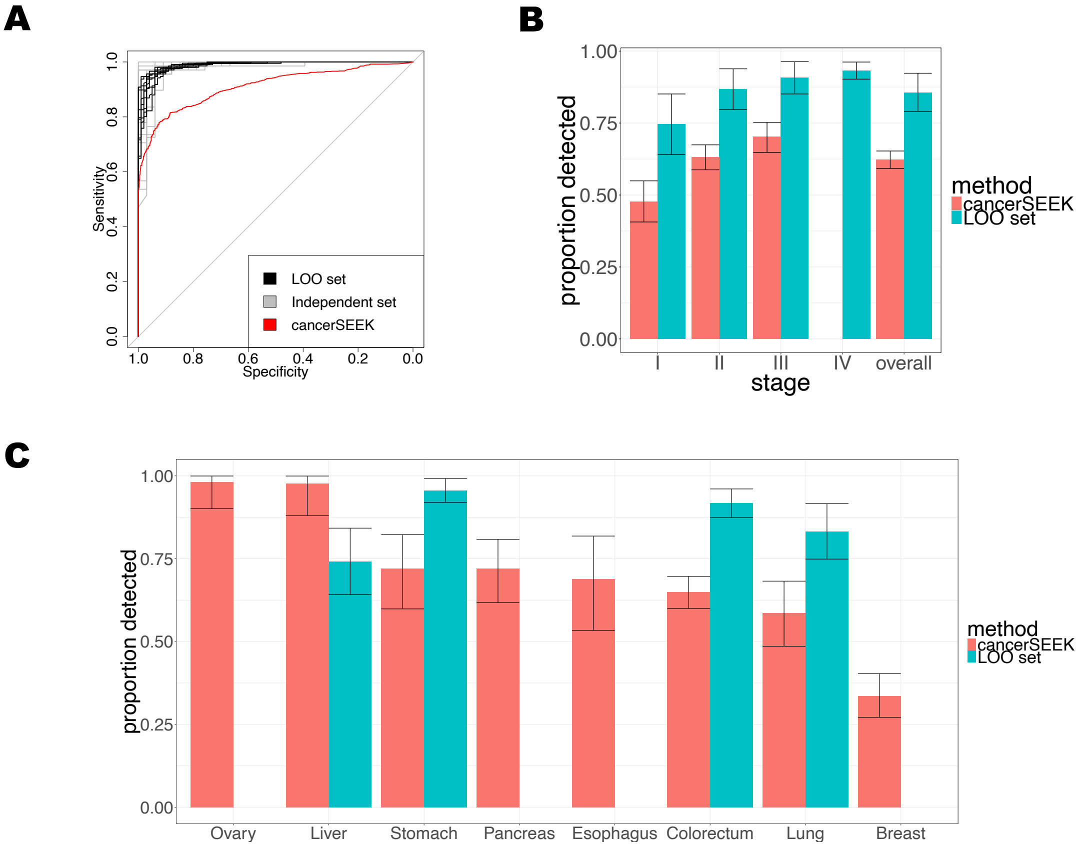

Figure 3.5: Comparison to cancerSEEK. (A) ROC curves comparing our independent (gray) and LOOCV (black) performance to that of cancerSEEK (red) (B) comparison of sensitivities for each cancer stage in our LOOCV set (blue) and cancerSEEK (red) (C) comparison of sensitivities for each cancer type in our LOOCV set (blue) and cancerSEEK (red); (B) and (C) are shown at 99% specificity for our LOOCV results

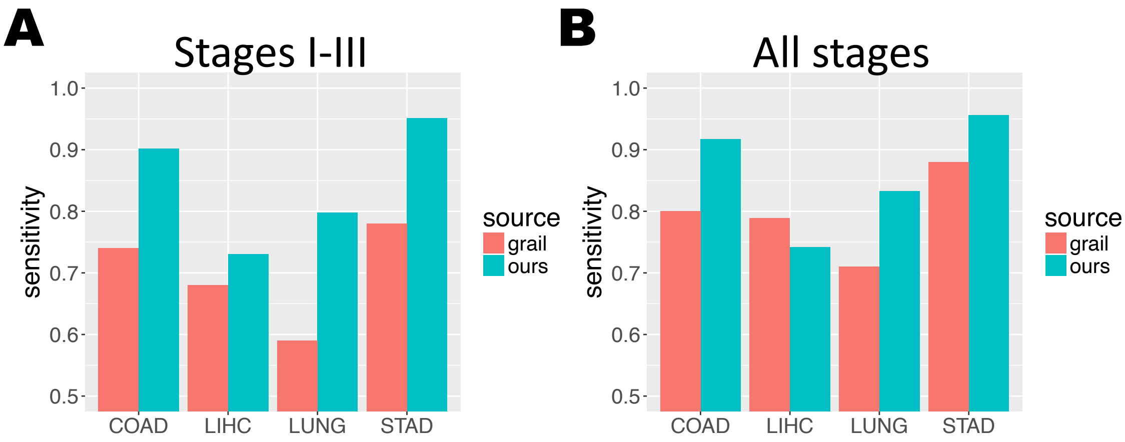

Figure 3.6: Comparison to GRAIL. (A) stages I-III (B) all stages. Results are shown on our LOOCV set at 99% specificity.

WGBS data. Their final panel covers 103,456 distinct regions (17.7 MB) including 1.1 million CpG sites, and probes only targeted fully methylated or fully unmethylated markers. Comparatively, our cfMethylSeq procedure theoretically covers 1.1 million regions (about 130 MB) and 5.7 million CpG sites.

cfDNA samples were profiled using their targeted panel. For each cfDNA sample, all reads falling into regions of interest were assigned probabilities of coming from different tissues, based on methylation patterns. Then, a logistic regression model was used to determine cancer/noncancer status.

Our results compared to Grail's results are shown in figure 3.6 . In non-metastatic stages (I-III), we achieve higher sensitivity for all cancer types profiled by our method (COAD, LIHC, LUNG (LUAD and LUSC combined), and STAD) (figure 3.6a) . For overall stages (figure 3.6b), we outperform Grail for all cancer types except for LIHC. Similar to the comparison for cancerSEEK [15], this is due to the overwhelming number of stage I LIHC samples in our sample set.

3.2.2.2 Model performance in cancer TOO prediction

We used the same strategy as pancancer detection to evaluate the performance of TOO prediction on 275 cfDNA cancer samples. Only cancer cfDNA samples correctly identified as having cancer during pancancer detection are used. Among 10 runs, for the 4 organ sites (Colon, Liver, Lung, Stomach) we achieved an average accuracy of 90.4% (standard deviation 0.5%) for LOOCV and 91.5% (standard deviation 5.0%) for independent validation (figure 3.7a). For early-stage (stage I and II) cancer patients alone, we achieved an accuracy of 86.5% (standard deviation 2.0%) in LOOCV and 89.1% (standard deviation 7.3%) in independent validation (figure 3.7b). Specifically, the prediction accuracies of colon/liver/lung/stomach identification are 81.8%/97.1%/96.7%/78.9% for all-stage and 70.1%/96.8%/96.2%/79.4% for early-stage cancer patients, respectively (figures 3.7a,b).

3.2.2.3 Enhanced predictive power by integrating multi-type features

We further evaluate how different features contribute to cancer detection performance. As shown in figure 3.8a, we observed the (1) reinforcing effect: 60.5% of all samples can be correctly predicted by all feature types, and the (2) complementary effect: methylation, CNV, cfDNA digestion size, and microbial abundances have uniquely correctly predicted 1.2%, 0.3%, 1.3%, and 1.5% of all samples in 10 runs, respectively. The ranking of features in terms of their Layer 1 AUROCs is methylation, digestion size, CNV, and microbial signatures with AUROCs of 0.972, 0.965, 0.946, and 0.928 (figure 3.8c). At the specificity of 99%, the ranking of sensitivity of 75.2%, 73.3%, 55.7%, and 14% was achieved for digestion size, methylation, CNV, and microbial compositions. If we leave out individual feature types, the average sensitivity will decrease by 14.4%, 4.4%, 5.1%, respectively for methylation, cfDNA digestion size, and microbial features. Among methylation markers, Type 1 achieved the highest AUROC of 0.959. The difference in sensitivity for cancer detection, between using all features and the best single feature is largest for stage 1 at 17.3%, indicating the necessity of integrating multiple features for early cancer detection.

For cancer TOO prediction, as shown in figure 3.7c, methylation features are the domi-

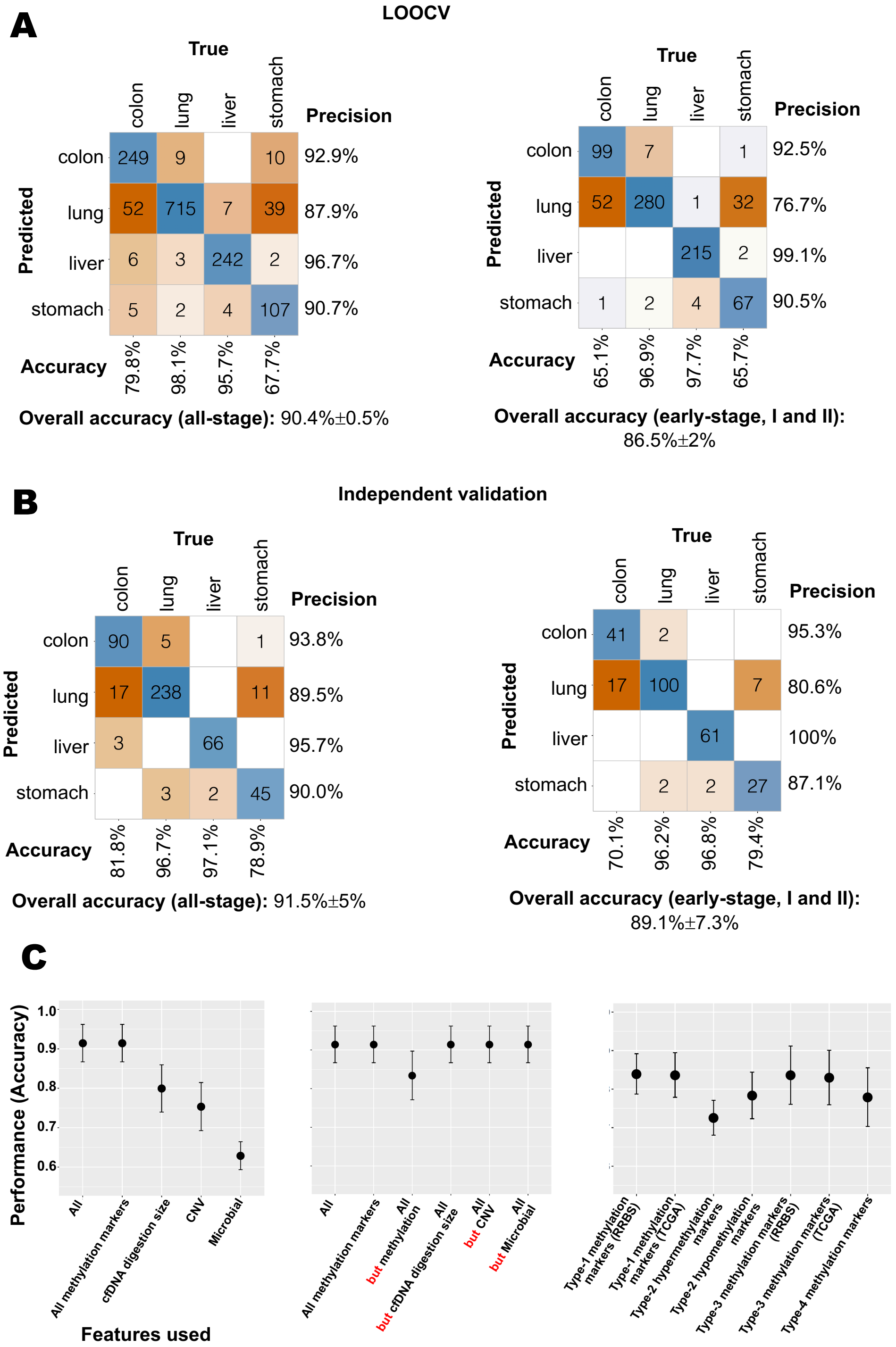

Figure 3.7: Cancer typing performance. Confusion matrices for all-stage and early-stage (i.e., stage I/II) cancer samples in (a) LOOCV and (b) independent validation. (c) Accuracy when using all feature types, each individual feature type, and all-but-one feature type.

nant contributor. Specifically, using the methylation features can achieve the same accuracy as integrating all feature types. That is, if we leave out the individual feature types (figure 3.7c), the average accuracy decreases by 8.8% for the methylation features, but does not decrease for cfDNA digestion size, CNV, and microbial features. Among the four methylation marker types, we observed (1) the reinforcing effect: 41.9% of all samples can be correctly predicted by all methylation marker types, and (2) the complementary effect: Type 1/2/3/4 methylation marker types have uniquely correctly predicted the locations of 1.4%, 5.0%, 3.0%, and 1.7% of all samples in 10 runs, respectively (figure 3.11a), and achieved accuracies of 84.0%, 78.4%, 83.6%, and 77.9%, respectively (figure 3.7c).

Complementarity of Type 1 and 2 methylation markers Interestingly, 94.9% and 96.5% of the Type 1 and 2 markers are non-overlapping, respectively (figure 3.8b). While Type 1 markers were selected based on the average methylation rates in individual regions, Type 2 markers require strong read-level signals in only a (possibly very small) subset of tumor samples, to capture tumor heterogeneity.

Specifically, Type 1 methylation markers are hyper and hypomethylated regions selected with the R package limma [86]. These markers on average show (e.g) hypomethylation in normal samples and hypermethylation in tumor samples. In contrast, the Type 2 methylation markers are based on α values. These markers must show (e.g) strongly consistent hypomethylation in normal samples and have elevated hypermethylation in some tumor samples. For Type 2 markers, no sophisticated deconvolution methods are needed to identify tumor markers. Since the normal reads show consistent methylation, any (e.g) hypermethylated read can be marked as a tumor read. For Type 1 markers, the difference between tumor and normal is not as strong, but the signal is more consistent.

Examples of Type 1 markers in normal cfMethylSeq and solid tissue samples can be seen in figure 3.9 as a heatmap of methylation (β) values in Type 1 markers. Although in general the trend is (e.g) hypomethylation in normal samples and hypermethylation in tumor samples, neither is at the extreme. In contrast, Type 2 markers show strong signal in the normal samples only, and strong opposite signal in only a small subset of tumor samples.

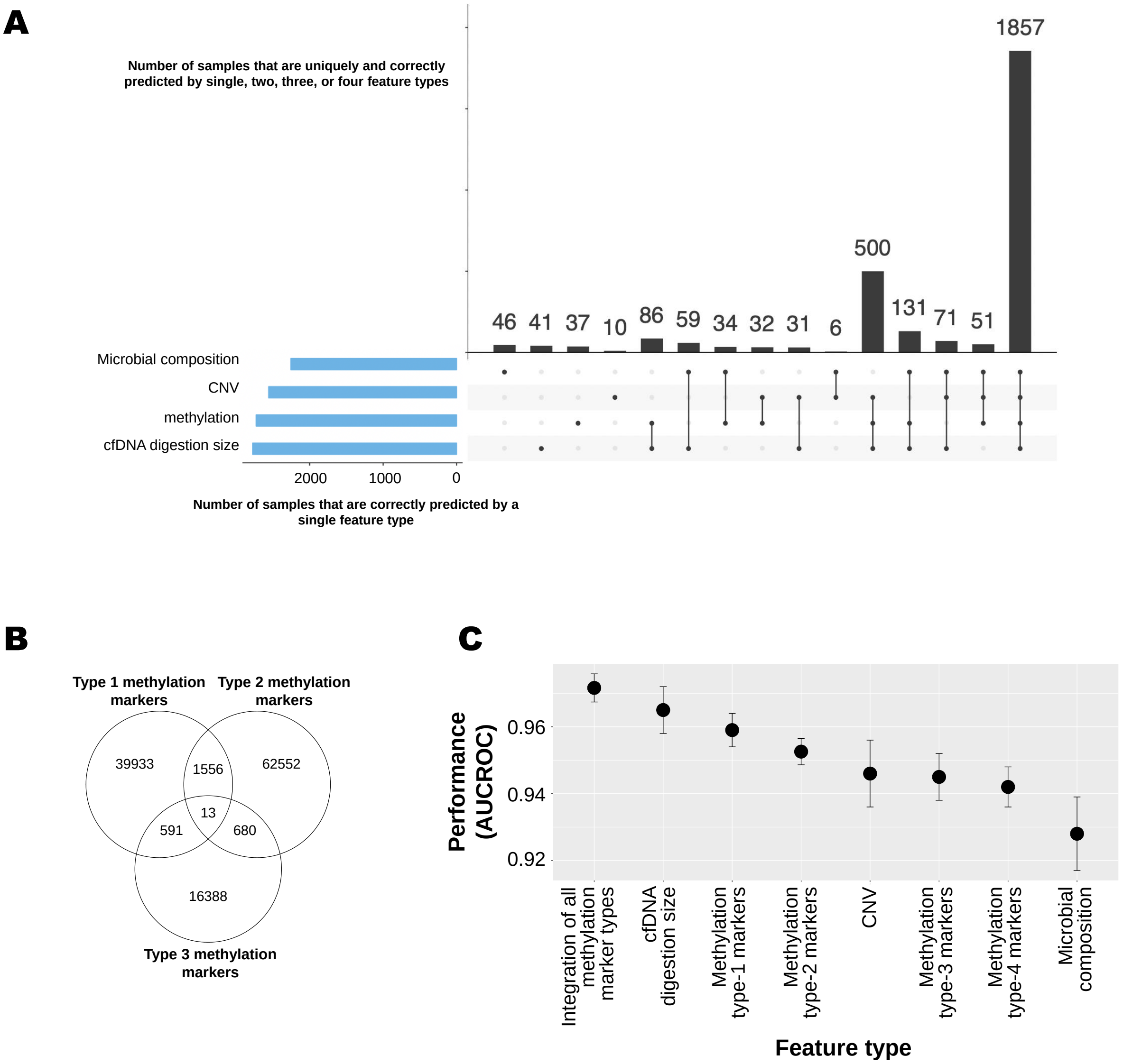

Figure 3.8: Synergistic analysis of multimodal cancer signals for cancer detection. (A) Visualization of intersecting sets of plasma samples in the LOOCV set that can be correctly predicted by each feature type at a false positive rate of 5 (B) Overlap between different methylation markers that are used for cancer detection. (C) Performance of individual feature types.

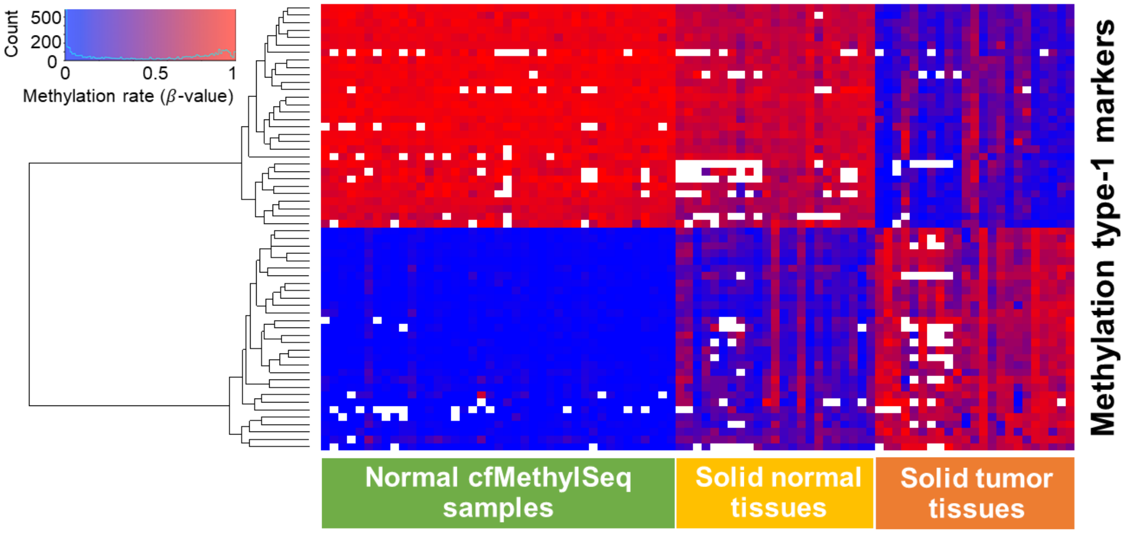

Figure 3.9: Heatmap illustrating marker discovery in Type 1 methylation markers. Color represents β value: red indicates high methylation, blue indicates low methylation. White indicates a missing value. Markers first must show opposite average methylation between solid tumor tissues and their matched normal tissues (right, middle panels). Then, opposite signals must be shown between solid tumor tissues and a set of normal cfMethylSeq samples (left panel).

The heatmap in figure 3.10 shows the count of hypermethylated reads (darker colors are higher numbers of hypermethylated reads) in regions that were selected as Type 2 markers (hypermethylated in tumor case).

Since the strategies used to define these markers select very different types of markers, there is little overlap between the regions selected (figure 3.8b). An example of a region that would be selected in both types of features would be a marker that is totally hypomethylated in all normal cfMethylSeq samples and totally hypermethylated in all tumor samples. In contrast, a region that is about 30% methylated in all normal samples and about 60% methylated in all tumor samples would be selected as a Type 1 marker but not as a Type 2 marker. A region that is 0% methylated in all normal samples but 50% methylated in only two tumor samples would be selected as a Type 2 marker but not as a Type 1 marker.

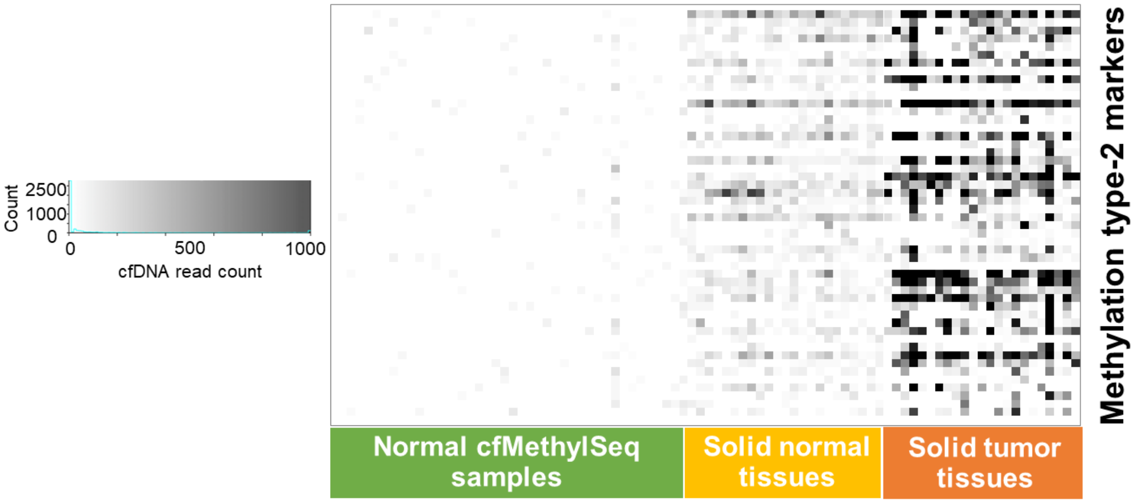

Figure 3.10: Heatmap illustrating marker discovery in Type 2 methylation markers, shown for the hypermethylated in tumor case. Color indicates the count of reads with α values equal to 1 (100% methylated), with darker colors indicating higher counts. Markers must have over 50% of reads with α values equal to 1 in at least one solid tumor sample and less than 25% of reads with α values equal to 1 in its matched normal sample (right, middle panels). Then, a set of normal cfMethylSeq samples (left panel) must show almost all reads with $\alpha < 0.5$.

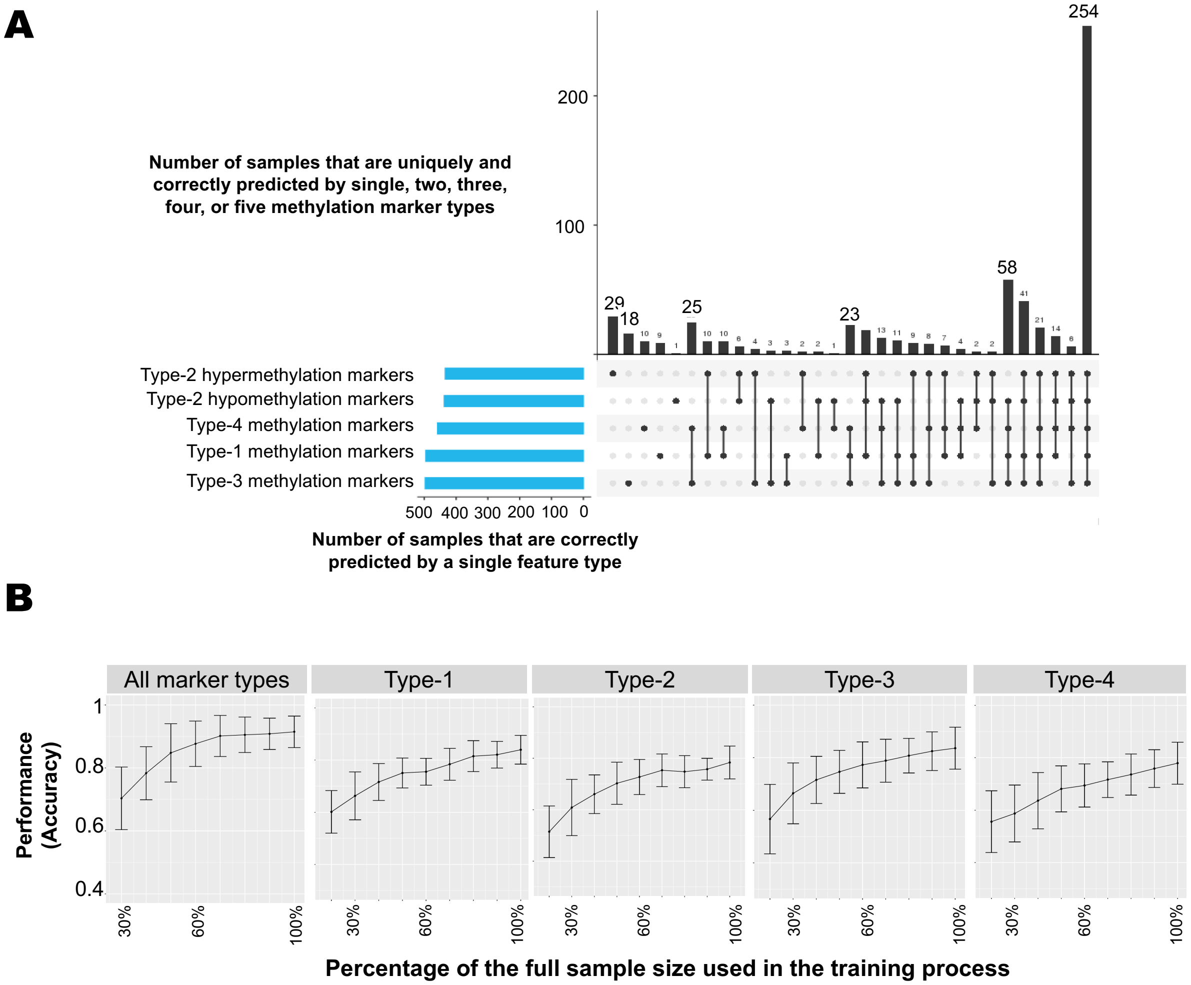

Figure 3.11: Synergistic analysis of multimodal cancer signals for cancer typing. (A) Visualization of intersecting sets of plasma samples in the independent validation set that can be correctly predicted by each methylation marker type. (B) Performance in the independent validation set increases as training sample size (fraction of LOOCV samples used) increases.

As illustrated from the TOO results (figure 3.11), the complementary signals these markers capture increase accuracy when an ensemble learner is used.

3.2.2.4 Impact of training sample size on performance

In their training of neural networks to automatically de-identify electronic health records, Dernoncourt et al. [19] showed that as training size increased, performance increased. We sought to reproduce this in our own results. To see the effect of using, for example, only 10% of the training set, 10% of the LOOCV samples were used to train a model, and this model was applied to all samples in the independent set. This was done 10 times over the 10 splits, at intervals of 10%.

As was found in [19], the performance of our cancer detection and typing models also increase with training size. As shown in figures 3.12a and 3.11b, as the training sample size increases (i.e., from 10%, 20%, . . . , 100% of the original size), the average performance of both cancer detection and TOO models on the same independent validation cohort increases, and the performance variance over 10 runs decreases. This result holds for the overall model as well as individual feature types, indicating our models do not over-fit the data and would have more power with larger numbers of training samples.

3.2.2.5 The optimal number of markers increases with training sample size

Among all the feature types used, the numbers of Type 1, 2, and 3 methylation markers are unconventionally high, e.g. over 40,000 markers for Type 1 markers. We show that the independent validation performances of both cancer detection and typing increase with the number of input markers, and performance plateaus when higher numbers of markers are used. We further showed that with more training samples used, more markers are needed to reach the best independent validation performances (figures 3.12b, 3.13a), testifying to the advantage of using the entire methylome rather than a small-panel-based approach. In addition, we found that Type 3 methylation markers (tissue markers) can achieve higher independent validation performance as the number of tissue types used to train the markers

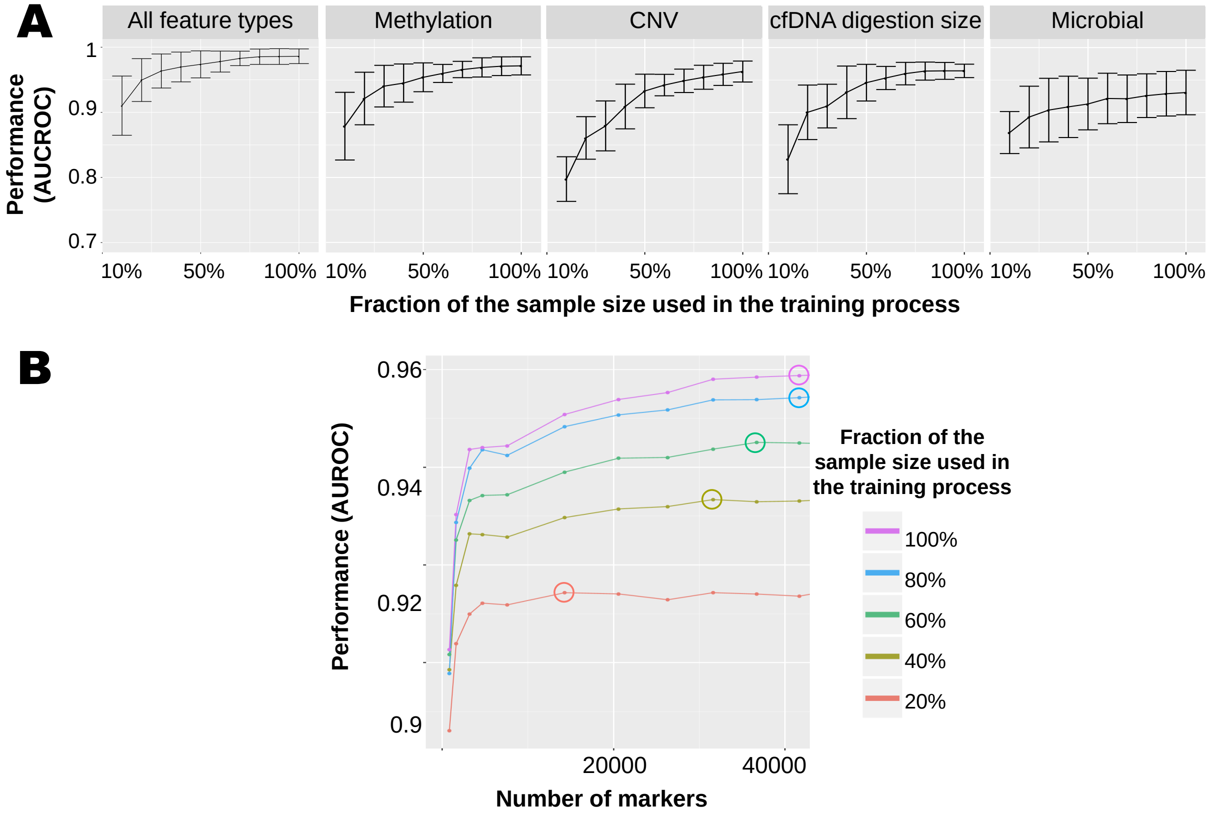

Figure 3.12: Cancer detection performance with relation to training sample size. (A) Independent validation performance using all feature types (leftmost panel) and individual feature types (methylation, CNV, cfDNA digestion size, and microbial composition, 4 rightmost panels), with increasing training sample size. Different proportions of all LOOCV samples are used to train the model, from 10% to 100%. (B) Increased training sample size not only improves the independent validation performance, but also achieves higher performance when marker number increases. Performance (y-axis) increases as the number of Type 1 markers (x-axis) increases, until reaching a plateau. The plateau is reached at higher numbers of markers as training size (indicated by color) increases.

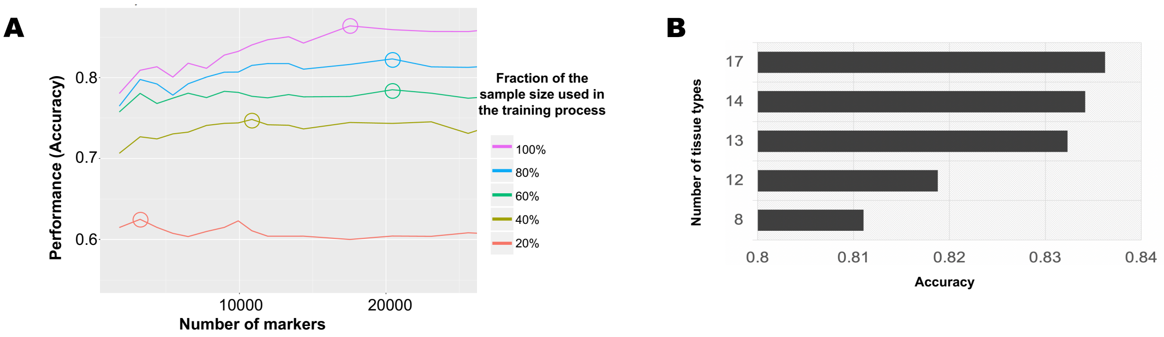

Figure 3.13: Cancer typing performance with relation to training sample size and number of markers. (A) Increased numbers of Type 1 markers (x-axis) achieve higher performance (y-axis) until reaching a plateau (circled). As more training samples are used (indicated by color), the optimal number of markers needed to reach the plateau increases. (B) Increasing the number of tissue types used to derive Type 3 markers yields higher cancer typing accuracy.

increases (figure 3.13b).

3.3 Methods

3.3.1 Sample collection and exclusion

Plasma samples from subjects with cancer were collected from patients at UCLA's hospitals and purchased from BioPartners, Inc. (Woodland Hills, CA). Plasma samples from subjects with cirrhosis were collected from patients at UCLA's hospitals. Plasma samples from healthy individuals were collected from UCLA's Institute for Precision Health and purchased from BioPartners, Inc. and BioChain Institute, Inc. (Newark, CA). Solid normal and tumor tissue samples were collected from UCLA's Translational Pathology Core Laboratory and purchased from BioPartners, Inc., Biochain Institute, Inc., Origene, Inc. (Rockville, MD), and Gundersen Health System (La Crosse, WI).

Healthy samples purchased from BioPartners met the following criteria: no cancer, drug addiction, or auto-immune diseases in medical history, no signs of acute disease, and no HIV, HCV, HBV, or syphilis. Healthy samples from UCLA's Institute for Precision health met the following criteria: no cancer, organ transplant, hepatitis, pancreatitis, cirrhosis, pancreatitis, sepsis, pregnancy, diabetes, or NASH in medical history. Healthy plasma samples purchased from BioChain had no diseases in their medical history. cfDNA was extracted from plasma samples with the QIAGEN QIAamp circulating nucleic acid kit (Catalog # 55114, Germantown, MD) following their instructions. The amount of starting material was 5-10ml plasma for non-cancer controls and 1-5ml plasma for cancer samples. The solid tissue gDNA samples were extracted with QIAGEN blood and tissue kit (Catalog # 69506). 10-100 ng tissue was used to extract gDNA from each sample.

cfMethylSeq samples from plasma samples obtained before 2017 and from problematic labs were excluded. Solid tissue samples that came from the same individual as cfMethylSeq samples were excluded. Some solid tissue samples were excluded based on copy number analysis.

Specifically, a bed file was made representing hg19 in 1 MB windows using `bedtools makewindows` [83]:

```
bedtools makewindows -g hg19.genome -w 1000000 > hg19_1MB.bed
```

This bed file was intersected with each sample's annotated bed file to get counts of mapped, deduplicated fragments in each 1MB window:

```
bedtools intersect -a hg19_1MB.bed -b sample_annotated_bed -c -F .5 >sample_1MB_counts.bed
```

To normalize, these 1 MB read counts were divided by the average 1 MB read counts of normal solid tissue samples sequenced in the same batch using the same library preparation methods. The normalized 1 MB profiles were visually examined and compared to the Broad Institute's Firehose Database of Gistic2 profiles for TCGA samples [72]. Solid tumor samples were excluded if they did not match the profiles in the Firehose database, and normal tissue samples were excluded if they were not flat.

3.3.2 Library construction, sequencing, and data processing

The DNA libraries of both cfMethylSeq (for cfDNA) and RRBS (for tissue genomic DNA) were sequenced with 150 bp paired-end reads using HiSeqX (Illumina) by Genewiz, Inc. (South Plainfield, NJ). cfMethylSeq libraries were prepared as described in chapter 2. The raw reads were preprocessed as depicted in detail in chapter 2; briefly we performed three steps to preprocess the cfMethylSeq and RRBS data. (Step 1) Removal of custom adapters with UMIs followed by read trimming. (Step 2) Sequence alignment, deduplication and methylation calling. We used Bismark [51] to align the trimmed reads to the reference genome hg19 (GRCh37 Genome Reference Consortium Human Reference 37 (GCA 000001405.1)). Then Umi-Grinder [50] was used to remove PCR duplicates based on the UMI labels (in the read names, allowing 4 mismatches in the total 16 bp UMI). Bismark [51] methylation extractor was then used to call methylation in the mapped, deduplicated reads. (Step 3) The chromosome-wise sequence alignment statistics and whole-methylome methylation statistics of CpG islands, CpG shores, gene promoters and repetitive regions were summarized from the individual read information obtained in Step 2. (Step 4) The mapping locations of R1 and R2 were merged to form one fragment.

3.3.3 Identification of methylation markers

3.3.3.1 Subsetting data to theoretical RRBS regions

Other groups have compared methylation between samples using average methylation levels (β values) in genomic bins [12, 46, 59]. This compensates for low sequencing coverage: smaller bins (e.g. individual CpG sites) require higher depth to reliably call methylation, whereas broad genomic bins can contain sufficient numbers of mapped fragments at low sequencing depths. However, average methylation levels in genomic bins cannot detect small deviations, e.g. 1 fully unmethylated read falling in a region with 99 other reads that were fully methylated would yield an average methylation of 99%. CancerDetector's α value [59], illustrated in figure 3.14, and Guo et al.'s methylation haplotype load [25] alleviated this issue by using methylation calls at the fragment level. An advantage of RRBS and cfMethylSeq, in

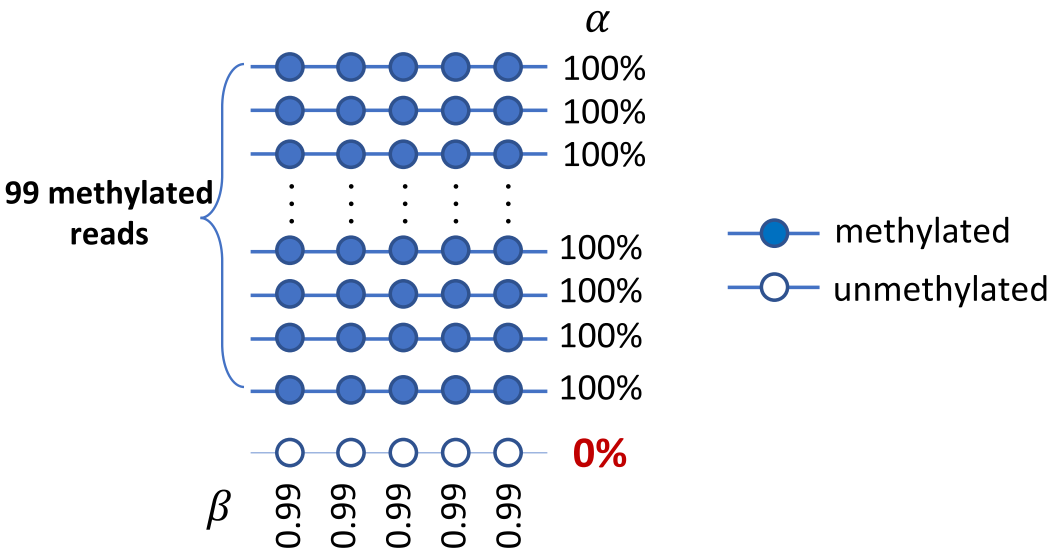

Figure 3.14: Illustration of α values vs. β values. β values measure methylation vertically, averaging methylation across all CpG sites covering a locus. α values, in contrast, measure methylation horizontally, at the read level.

addition to increased coverage for reduced cost, is that MspI digestion creates known mapping locations that the majority of fragments will map to. These MspI-produced fragments create natural genomic bins for methylation analysis; fragments will span the entire bin and be directly comparable across samples. Unlike broad genomic bins, a cfMethylSeq fragment can never fall in more than one theoretical RRBS region and will not have CpG sites outside of the bin boundaries.

Theoretical RRBS regions on the positive strand were defined by in silico digesting the hg19 reference genome with MspI. All occurrences of the MspI cut site “CCGG” were found in the hg19 reference genome fasta file. The string “CCGG” was replaced with “C—CGG” and the genome was split on the character “—”. These fragments and their genomic locations were written to a file, which was then size selected for fragments between 0 and 350 bp in length. The reference genome represents the positive strand of the genome. Fragment locations were shifted forward by 2 bp at both the start and end locations to get mapping coordinates on the negative strand. Positive and negative mapping locations were counted as separate fragments. In total there are 2,178,790 theoretical RRBS regions in hg19. Each theoretical RRBS region was assigned an index.

Mapped, deduplicated fragments that intersected exactly with these theoretical RRBS regions were extracted from the annotated bed files using bedtools (`-f 1 -F 1 -s` options) [83]. On average 85% of mapped, deduplicated fragments met this criteria. After subsetting to exact RRBS fragments, a python script was used to extract the average methylation of each fragment, referred to as the α value. For example, a fragment with 10 sequenced CpG sites, 4 of which are methylated, would have $\alpha = 0.4$. Because most cfMethylSeq and RRBS fragments have α values of either 0 or 1 (see figure 3.17 in section 3.3.3.3), fragments were stratified into 5 groups: $\alpha = 1$, $\alpha > 0.5$, $\alpha = 0.5$, $\alpha < 0.5$, and $\alpha = 0$.

An R script was used to aggregate this information by the theoretical RRBS region index. Ultimately a file was produced for each sample with a line for each of the possible 2,178,290 theoretical RRBS regions covered. Each line contains a count of how many sequenced fragments had $\alpha = 1$, $\alpha = 0$, $\alpha < 0.5$, $\alpha > 0.5$, and $\alpha = 0.5$ for that sample.

Table 3.2 shows the first 10 lines of one such file. Line two indicates that, in this sample, theoretical RRBS region 3 (chr1 10496 10588 +) had 17 fully methylated fragments, 0 fragments with no methylation, 1 fragment with less than half CpGs methylated, 40 fragments with more than half the CpGs methylated, and 0 fragments with exactly half the CpGs methylated. A missing index indicates there were no fragments at all in this sample—in this sample there were no fragments falling in marker indices 1 or 2, for example.

3.3.3.2 Merging theoretical RRBS regions by strand

MspI cuts DNA at CCGG sites, leaving CG overhangs on either end. This means that in the final library, there is a difference in the strand coverage for CpG sites falling in MspI digestion sites. This is illustrated in figure 3.15. The original DNA fragment, before MspI digestion, contains 4 CpG sites: A, B, C, and D. After MspI digestion, the fragment that

theoretical RRBS region index	$\alpha = 1$	$\alpha = 0$	$\alpha < 0.5$	$\alpha > 0.5$	$\alpha = 0.5$
3	17	0	1	40	0
4	30	0	2	38	1
54	1	0	0	1	0
122	5	1	0	6	0
126	1	0	0	0	0
129	3	1	1	1	0
130	2	0	0	1	0
131	0	0	1	0	0
143	102	1	0	0	3

Table 3.2: Example file format for α value stratification in theoretical RRBS regions for a cfMethylSeq sample. Each line contains a count of how many sequenced fragments had the column-specified α value in each theoretical RRBS region (row)

maps to the positive strand will yield methylation calls for CpG sites A, B, and C, whereas the fragment that maps to the negative strand will yield methylation calls for CpG sites B, C, and D.

Because the mapping locations are shifted by two bp between the top and bottom strand, the subsetting of the annotated bed file into theoretical RRBS regions takes the strand information into account (i.e., these are counted as two separate regions). However, in our analyses, the methylation calls between the two strands are highly correlated. For downstream analyses, we merged theoretical fragments 1 and 2 into one region.

Table 3.3 shows an example output file after merging. Original markers 3 and 4 were merged into one marker, now indexed as 2. The sum of reads counts from markers 3 and 4 is now in marker 2. Some markers, such as original marker 54, only had read counts from marker 54, not marker 53. New marker 27 includes counts from original markers 53 and 54, but since 53 had no reads the counts are the same for new marker 27 as original marker 54. These processed files are used for Type 2 methylation marker identification, as described

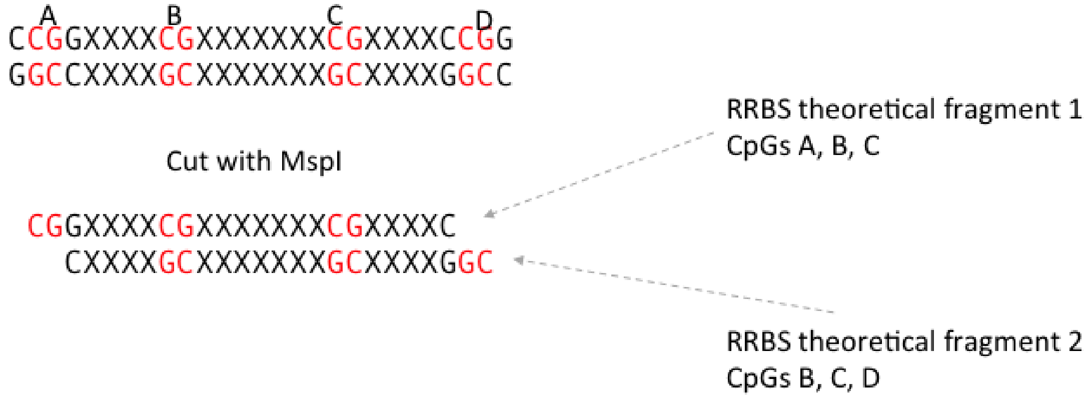
A
B
C
D
CCGGXXXXCGXXXXXXXCGXXXXCCGG
GGCCXXXXGCXXXXXXXGCXXXXGGCC
Cut with MspI
CGGXXXXCGXXXXXXXCGXXXXC
CXXXXGCXXXXXXXGCXXXXGGC
RRBS theoretical fragment 1
CpGs A, B, C
RRBS theoretical fragment 2
CpGs B, C, D

Figure 3.15: Overview of theoretical RRBS regions with respect to the original DNA fragment prior to MspI digestion. After MspI digestion, CpG sites falling in MspI digestion sites (CCGG; A and D in the diagram) will only be covered by fragments mapping to the positive (CpG A) or negative (CpG D) stands. These are counted as separate theoretical RRBS regions.

merged theoretical RRBS region index	$\alpha = 1$	$\alpha = 0$	$\alpha < 0.5$	$\alpha > 0.5$	$\alpha = 0.5$
2	47	0	3	78	1
27	1	0	0	1	0
61	5	1	0	6	0
63	1	0	0	0	0
65	5	1	1	2	0
66	0	0	1	0	0
72	102	1	0	0	3

Table 3.3: Example file format for α value stratification in merged theoretical RRBS regions for a cfMethylSeq sample. Each line contains a count of how many sequenced fragments had the column-specified α value in each merged theoretical RRBS region (row), where merged theoretical RRBS regions combine the paired theoretical RRBS region indices mapping to the respective top and bottom strands (see figure 3.15)

below.

3.3.3.3 Using processed data to identify markers

To identify the methylation markers, we generated RRBS data from 251 samples of solid tumors and adjacent normal tissues for liver, lung, colon, and stomach cancer. We also used cfMethylSeq data from cfDNA of 41 non-cancer individuals. In addition, we used array-based methylome profiles (Illumina 450K) from TCGA [114], and RRBS-based methylomes of normal tissues from GEO [5] (figure 3.16). We extract four types of methylation markers for cancer detection and typing, following different marker discovery principles. For the first three marker types, the unit regions are (merged) theoretical RRBS regions or groups of theoretical RRBS regions as described above. The unit regions of the Type 4 markers are 1 MB equal-sized regions in the hg19 genome (referred to as bins). All theoretical RRBS regions and 1 MB regions falling in chromosomes X, Y and M or without mapped reads across all samples are excluded from marker discovery. Specifically,

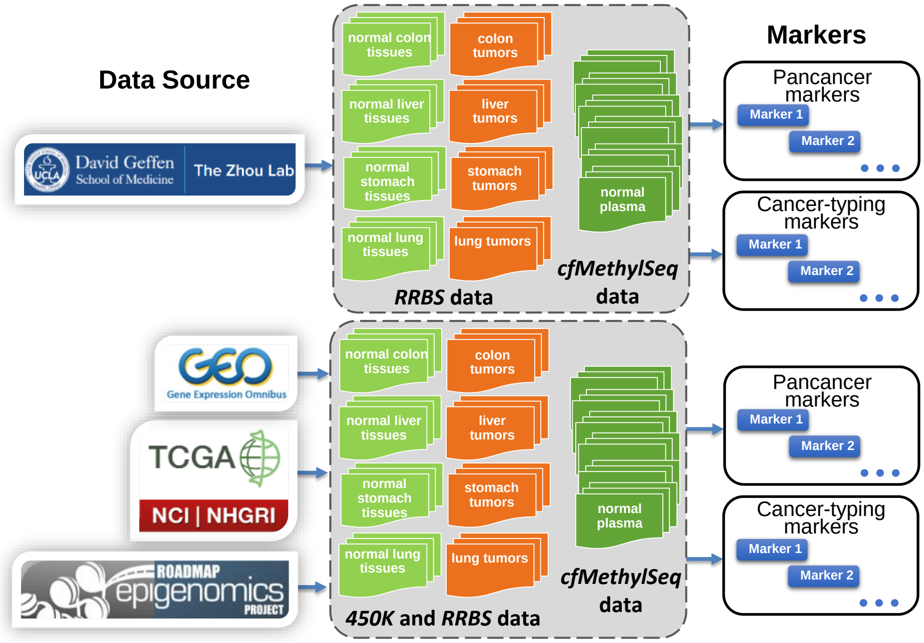

Figure 3.16: Overview of samples and data sources for identifying methylation markers

- β-value-based tumor markers (Type 1): These are regions that show differential methylation between tumors and normal samples. Specifically, the limma method [86] (R package limma version 3.42.0) was used to select genomic regions, in which the methylation rate (i.e. β-value, $\beta = \frac{\texttt{\# methylated CpGs in all mapped reads within the region}}{\texttt{\# CpGs in all mapped reads within the region}}$) differentiates not only between solid tumors and adjacent normal samples, but also between solid tumors and the set-aside set of 41 non-cancer cfDNA samples. These regions are selected separately for each cancer type. The empirical Bayes moderated t-test relative to a log2-fold-change threshold 1.0 was used in limma. To adjust for age differences in methylation, markers are first found between solid matched normal and tumor tissue. Specifically, the top 150,000 hypermethylated and hypomethylated markers are identified for colon adenocarcinoma (COAD) and their matched normal samples, liver hepatocellular carcinoma (LIHC) and their matched normal samples, lung adenocarcinoma (LUAD) and their matched normal samples, lung squamous cell carcinoma (LUSC) and their matched normal samples, and the top 225,000 hypermethylated and hypomethylated markers only for stomach adenocarcinoma (STAD) and their matched normal samples. Then the final marker set is the union of the top 6,000 hypermethylated and top 6,000 hypomethylated markers identified for COAD, LIHC, LUAD, LUSC, and the top 9,000 hypermethylated and hypomethylated markers only for STAD, selected from the initial pools of 150,000 hyper and hypomethylated markers for COAD, LIHC, LUAD, LUSC, and 225,000 for STAD. Because STAD markers generally have lower differential power (i.e., lower fold change) and fewer training samples than other tumor types, we identified more markers only for STAD. We generated and used the RRBS data of 101 tumor samples (19 COAD tumors, 23 LIHC tumors, 21 LUAD tumors, 23 LUSC tumors, 15 STAD tumors) and their adjacent normal tissues, and the cfMethylSeq data of 41 non-cancer cfDNA samples, to identify 41,493 RRBS-based tumor markers on average (over 10 random selections of 41 non-cancer cfDNA samples). Since this method uses the β-values, it can be applied to the Illumina-450K array data of TCGA. That is, we employed the same method to array data of 149 tumor samples (38 COAD tumors, 49 LIHC tumors, 29 LUAD tumors, 40 LUSC tumors, 2 STAD tumors) and their normal adjacent tissue, and the cfMethylSeq data of 41 non-cancer subject's cfDNA samples to identify 18,190 array-based tumor markers on average (over 10 random selections of 41 non-cancer cfDNA samples). Note that (1) only those unit regions with at least 10 cfMethylSeq reads in 70% of 41 non-cancer subject's cfDNA samples are used as the marker candidates, (2) Illumina-450K array data cover a smaller number of unit regions than the RRBS data, and therefore we selected a smaller number of tumor markers, i.e., the top 5,000 hypermethylated and hypomethylated markers for COAD, LIHC, LUAD and LUSC, and the top 7,500 hypermethylated and hypomethylated markers only for STAD, and (3) the missing values of solid tissue samples were imputed using the nearest neighbor averaging method in the R package "impute". The overview of data sources is depicted in figure 3.16. Note that all Type 1 markers were derived by strictly comparing the tumors and their adjacent normal tissues, in order to minimize the age influence [64].

- α-value-based tumor markers (Type 2):

  In contrast to Type-1 tumor markers that compare the population-averaged methylation measurement of a region (i.e., β-values which are the averaged methylation state over all reads in a region), this marker type compares a different methylation measurement, the methylation rate of a single sequencing read, so called α-value [59] ($\alpha=\frac{\texttt{\# methylated CpGs in a mapped read}}{\texttt{\# CpGs in a mapped read}}$). This is a sequencing-read-level measure, a higher resolution than the region-population-level measure (β-value). α values harness the power of multiple CpG sites on a read to increase detection of differential methylation. While a single hypomethylated read in a background of mostly methylated reads would not lead to a differential β value, it can be easily detected using α values (see figure 3.14).

Rationale Several methods in the literature employ PCR-based approaches to detect cancer in cfDNA, where the signal may be very low in tumor samples, but consistently not observed in normal samples [81, 88, 120, 60, 111, 3]. These strategies often target a small number of gene promoters or transcription start sites that are unmethylated in normal samples and highly methylated in cancer samples. After treating the DNA sample with bisulfite conversion, PCR primers are used that will only selectively amplify the methylated version of the target. If a methylation-specific PCR product is observed over a certain limit, then the methylated target was considered present [30]. These tests are simple to perform and do not require sequencing, but do require prior knowledge of the targets of interest [34]. These tests are highly specific but have low sensitivity; i.e. rarely will a methylated product be found in a non-cancerous case, but many cancer samples will lack the methylation product.

In our cfMethylSeq data, which covers over 75% of all gene promoters and CpG islands in the genome, we reasoned we could apply this highly specific strategy at a much larger scale. Briefly, markers were selected that were consistently highly methylated or unmethylated in a set of cfMethylSeq normal samples, and showed the opposite signal in a (possibly very small) subset of solid tumor samples. We used the concept of α value [59], to amplify signals at the read level. These markers might not display noticeable methylation differences between tumor and normal samples at a population level. This type of marker cannot be found with 450k data and is further strengthened by our theoretical RRBS regions which are comparable across samples and are often entirely unmethylated or methylated.

Distribution of α values in cfMethylSeq and RRBS samples To mimic the results seen in PCR-based cancer detection assays, Type 2 markers should be totally methylated in cancer and totally unmethylated in normal samples, or vice versa. Most α values are either 0 or 1 (figure 3.17). A cutoff of α being greater than 0.5 or less than 0.5 should be sufficient for determining if a fragment is mostly methylated or unmethylated, because in most cases $\alpha > 0.5$ implies the fragment is totally methylated, and $\alpha < 0.5$ implies the fragment is totally unmethylated. This cutoff is used in the marker discovery pipelines below. Figure 3.18 illustrates the strategy of read-level tumor marker discovery.

Markers that are hypermethylated in tumor The 41 non-cancer cfDNA samples were used in the initial marker selection. All solid tumor and normal RRBS matched pair samples passing the exclusion step were used in the initial marker selection. In the initial pass, markers were selected that displayed consistent hypomethylation in the normal cfMethylSeq reference samples but the opposite signal in a subpopulation of solid tumor tumor RRBS samples. Specifically, two types of theoretical RRBS regions were selected as initial markers. Final markers are the intersection of regions satisfying a and b:

(a) regions where there were no fragments with $\alpha \geq 0.5$ in $\geq 90\%$ of normal cfMethylSeq samples

(b) regions where at least 1 solid tumor sample had >50% of its fragments (in that region) totally methylated ($\alpha = 1$) and its matched normal sample had $< 25\%$ of its fragments (in that region) totally methylated ($\alpha = 1$))

The 90% thresholds were out of all samples that had at least 10x coverage, for example,

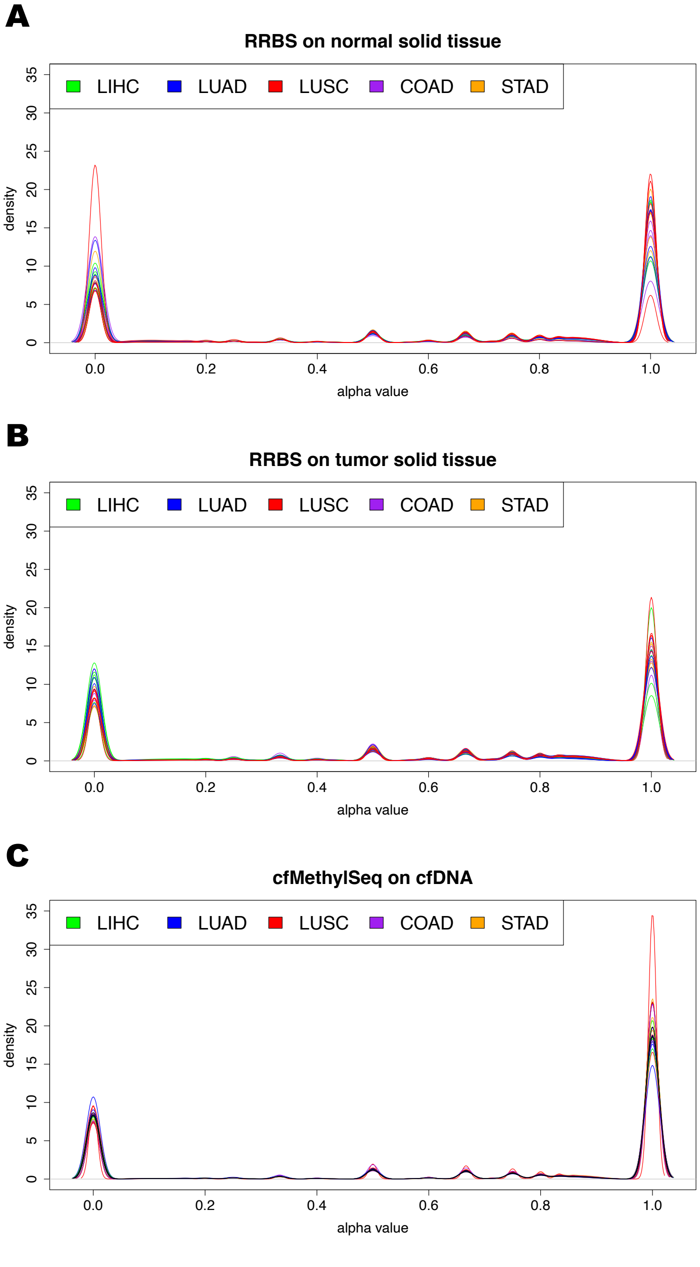

Figure 3.17: α values in RRBS and cfMethylSeq data. (A) normal solid tissue, (B) tumor solid tissue, and (C) cfMethylSeq data

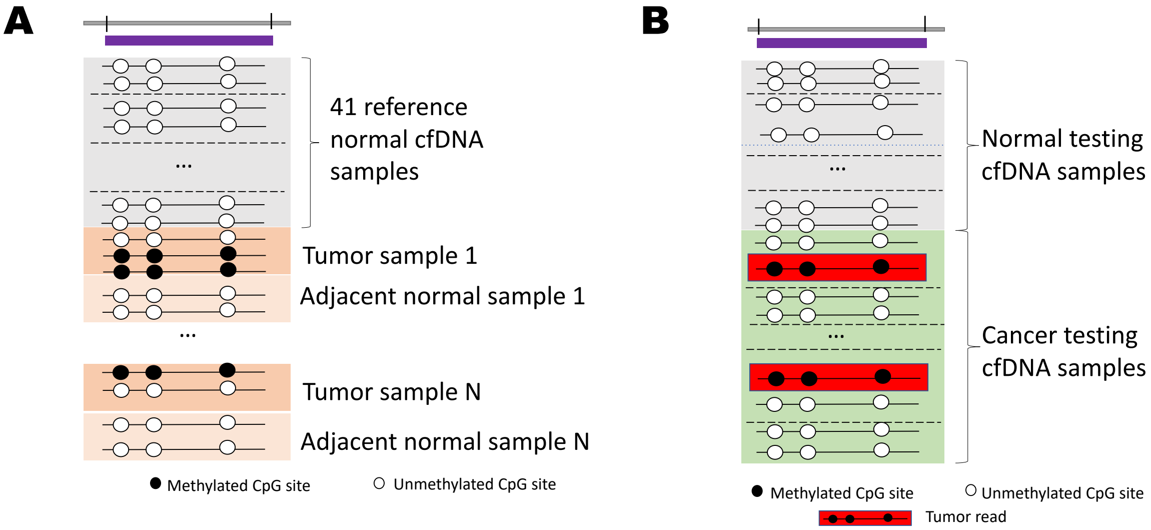

Figure 3.18: Conceptual illustration of (A) Type 2 marker discovery strategy, shown for the hypermethylated in tumor case, and (B) read deconvolution in Type 2 hyper markers. In (A), the region shown is selected as a Type 2 hyper marker because it is unmethylated in most of the normal cfMethylSeq samples and there is at least one solid tumor/normal adjacent pair where the solid tumor has $> 50\%$ of its reads with $\alpha = 1$ and the adjacent normal sample has less than 25% of its reads with $\alpha = 1$. In (B), hypermethylated reads ($\alpha > .5$, shown in red) in this region are considered tumor reads

of the cfMethylSeq normal samples, if only 21 had coverage in a theoretical RRBS region, 19 must have $\alpha \geq 0.5$ to be a potential Type 2 marker. If 3 had $\alpha < 0.5$, this would indicate only 18/21=88% of normal samples had $\alpha \geq 0.5$, and requirement (a) would not be satisfied. One of the 20 non-covered cfMethylSeq samples may have had, say, 5 fragments falling in this region, but it would be counted as no coverage because it did not meet the minimum coverage requirement of 10 fragments. Markers for LIHC, LUAD, LUSC, COAD, and STAD were found separately and then aggregated to form one set of initial pancancer markers; any marker that was a cancer marker in at least 1 tumor type was used.

Using our RRBS data of 101 tumor samples and their adjacent normal tissues, we identified 9,374 α-value-based hypermethylation markers on average (over 10 random selections of the 41 non-cancer cfDNA samples).

Markers that are hypomethylated in tumor Type 2 markers that are hypomethylated in cancer are defined using bigger regions than the Type 2 hypermethylated in cancer markers. The rationale is that hypomethylated markers are more broad in the genome and not as specifically located as hypermethylated markers [12]. Superbins are defined that group theoretical RRBS regions together that are within a certain distance apart until the accumulated size of the region exceeds a threshold. Specifically, theoretical RRBS regions are grouped together that are less than 200 bp apart until the final region exceeds 1000 bp. Superbins can overlap. Superbins that were less than 100 bp were removed.

The 41 non-cancer cfDNA samples were used in the initial marker selection. All solid tumor and normal RRBS matched pair samples passing the exclusion step were used in the initial marker selection. In the initial pass, markers were selected that displayed consistent hypermethylation in the normal cfMethylSeq reference samples but the opposite signal in a subpopulation of solid tumor tumor RRBS samples.

Specifically, two types of superbins were selected as initial markers, final markers are the intersection of superbins satisfying a and b:

(a) regions where there were no fragments with $\alpha \leq 0.5$ in $\geq 70\%$ of normal cfMethylSeq samples

(b) regions where at least 1 solid tumor sample had $> 25\%$ of its fragments (in that region) somewhat unmethylated ($\alpha \leq 0.5$) and its matched normal sample had $< 10\%$ of its fragments (in that region) somewhat unmethylated ($\alpha \leq 0.5$)

The 70% thresholds were out of all samples that had at least 10x coverage, as in the hypermethylated markers. Markers for LIHC, LUAD, LUSC, COAD, and STAD were selected separately and then aggregated to form one set of initial pancancer markers; any marker that was a cancer marker in at least 1 tumor type was used.

Using our RRBS data and their adjacent normal tissues, we identified 32,501 α-value-based hypomethylation markers on average (over 10 random selections of the 41 non-cancer cfDNA samples).

- β-value-based tissue markers (Type 3): These are regions that have differential methylation between tissue types. Specifically, the limma method [86] (R package limma version 3.42.0) was used to select genomic regions in which the β-value differentiates not only between one tissue type and all other tissue types (i.e., one-vs-rest for tissue-type-specific markers), but also between pairs of tissue types (i.e., one-vs-one for tissue-type-pair comparison markers). The empirical Bayes moderated t-test relative to a log2-fold-change threshold 1.0 was used in limma. Then the final marker set is the union of the top 200 one-vs-rest markers identified for each tissue type and top 30 one-vs-one markers identified for each tissue type pair. Note that because stomach markers have lower differential power (i.e., lower fold change) than other tumor types, we identified the top 400 one-vs-one markers and the top 200 one-vs-rest markers only for stomach tissue type. Since this method uses the β values, it can be applied to both methylation sequencing data and Illumina-450K array data from TCGA [114]. Using the RRBS data of 17 tissue types with 217 samples (1 adipose, 3 b-cell, 9 brain, 38 colon, 3 esophagus, 4 granulocyte, 2 heart, 4 kidney, 25 liver, 68 lung, 4 monocyte, 13 neutrophil, 7 pancreas, 1 small intestine, 1 spleen, 26 stomach, 8 t-cell) collected from

GEO [5], we identified 17,672 RRBS-based tumor markers. Using the TCGA Illumina-450K array data of 7 tissue types with 391 normal tissues (38 colon, 205 kidney, 50 liver, 74 lung, 13 neutrophil, 9 pancreas, 2 stomach) and RRBS data on 19 samples from 4 blood cell types (3 b-cell, 8 t-cell, 4 granulocyte, 4 monocyte) collected from GEO [5], we identified 9,818 array-based tumor markers.

- β-value-based 1 MB bins (Type 4): In contrast to Type 1, 2, and 3 markers that use theoretical RRBS regions or groups of theoretical RRBS regions as unit marker regions, Type 4 markers use equal-sized 1MB bins in the hg19 genome as markers. With these markers, we aim to capture the broad hypomethylation observed in cancer [12, 20]. The methylation rate (i.e., β value) of all reads falling in repeat regions in a bin is calculated as the value of this bin in the profile. Repeat regions are defined by UCSC table browser [103].

3.3.4 Marker profile generation

3.3.4.1 Computing Type 1 and 3 methylation marker profiles

We developed an algorithm to deconvolute cfDNA reads in two contexts: (1) deconvolute reads into tumor-derived reads or background reads; and (2) deconvolute reads into reads from different tissues. This algorithm extends our previous tissue-deconvolution algorithm [18] by adding an unknown class to absorb reads that are not likely to belong to any known classes. Given the methylation signatures of T classes in a methylation marker, without loss of generalization, we assume that a cfDNA read falling in this region can be assigned a probability of coming from each of these T classes. Those cfDNA reads that have multiple classes with high probabilities are considered to have ambiguous class memberships, and we assume they may come from an unknown class. This algorithm includes three steps:

1. **Calculate class-specific likelihoods.** To assess how well the joint methylation status of multiple CpG sites on a read fits the methylation signature of a class t, we calculated the class-specific likelihood that a cfDNA read came from a class t, $P(read|class\ t)$,

by the method used in our CancerDetector paper [59]. T likelihoods are calculated for each cfDNA read, each corresponding to one of the T classes. If the fold change between the highest likelihood and the second highest likelihood is less than a threshold 2, this cfDNA read is considered an ambiguous read, otherwise it is considered unambiguous.

2. **Estimate the overall class composition.** Given N_T unambiguous reads of a cfDNA sample, we denote the cfDNA composition of T known classes as a vector $\Theta = (\theta_1, \theta_2, \cdots, \theta_T)$, which satisfies $\sum_{t=1}^{T} \theta_t = 1$. We want to estimate Θ by maximizing the log-likelihood $\log P(all\ reads|\Theta, \mathbf{T\ classes})$. This is a maximum likelihood estimation problem, which can be solved by the EM algorithm. Assuming the independence of each read, $P(all\ reads|\Theta, \mathbf{T\ classes}) = \prod_{i=1}^{N_T} P(read^{(i)}|\Theta, \mathbf{T\ classes})$. For calculating $P(read^{(i)}|\Theta, \mathbf{T\ classes})$, we introduce a latent random variable $z^{(i)}$ for each read $read^{(i)}$ to indicate which class it comes from; i.e., $z^{(i)} = t$ where $(1 \leq\ t \leq\ T)$. We can then expand the likelihood $P(read^{(i)}|\Theta, \mathbf{T\ classes})$ of $read^{(i)}$ as follows:

$$P\left(read^{(i)}\middle|\Theta, \mathbf{T\ classes}\right) = \sum_{t=1}^{T} P\left(read^{(i)}\middle|\ z^{(i)} = t, class\ t\right) P(z^{(i)} = t|\Theta) \tag{3.1}$$

$$= \sum_{t=1}^{T} \theta_t P\left(read^{(i)}\middle|class\ t\right) \tag{3.2}$$

Here $P(z^{(i)} = t|\Theta)$ is the prior probability that each read $read^{(i)}$ belongs to class t, so we have $P\left(z^{(i)} = t\middle|\Theta\right) = \theta_t$. Let $q_i\left(t\right) = P(z^{(i)} = t|read^{(i)})$ be the posterior probability. According to the EM algorithm, we have the following iterative steps for optimization. In the E-step, $q_i\left(t\right)$ is estimated by the posterior probability of $z^{(i)}$ given read $read^{(i)}$, the methylation marker, and the composition Θ calculated from the last iteration. In the M-step, given the estimated $q_i\left(t\right)$, we estimate the cfDNA composition Θ using the maximum likelihood.

$$\texttt{E-step:}\ q_i\left(k\right) \leftarrow\ P\left(z^{(i)} = k\middle|\Theta, read^{(i)}, \mathbf{T\ classes}\right) = \frac{\theta_t P\left(read^{(i)}\middle|\ t\right)}{\sum_{t=1}^{T} \theta_t P\left(read^{(i)}\middle|\ t\right)}$$

$$\texttt{M-step:}\theta_t \leftarrow \frac{\sum_{i=1}^{N} q_i\left(t\right)}{\sum_{t=1}^{T}\sum_{i=1}^{N} q_i\left(t\right)}, \quad \textit{where } 1 \leq \ t \leq \ T$$

Intuitively, the E-step updates "soft labels" of each read and the M-step updates the overall composition by summing "soft labels" over all reads. A random Θ is used to initialize this procedure, and iteration continues until Θ converges to the estimated solution $\hat{\Theta}$. Since the EM algorithm converges to a local optimum, we repeat the EM procedure with several random initializations, and take the best solution as final. After obtaining the composition of T known classes for the unambiguous reads, we can calculate the overall composition of both the T known origin-types for all cfDNA reads and the ambiguous reads. They are calculated as below:

$$\texttt{For unknown class } u: \quad \theta_u = \frac{N_u}{N}$$

$$\texttt{For known class } t: \quad \theta_t \leftarrow \left(1-\theta_u\right)\theta_t \quad , \quad \textit{where } 1 \leq \ t \leq \ T$$

where N and N_u are the number of total and ambiguous reads, respectively, that belong to the marker regions.

3. **Calculate normalized class-specific read counts in each marker.** We count the number of class-specific reads in each marker as the sum of the class-specific posterior probability of all cfDNA reads within the marker, i.e.,

$$count_t(marker) = \sum_{unambiguous\ reads\ in\ marker} \theta_t P(read|t)$$

for the class t and $count_u(marker) = \frac{ambiguous\ read\ count\ within\ marker}{total\ read\ count\ within\ marker}$ for the unknown class. We then normalize a read count by the sample's sequencing depth and used its logarithm transformation as the final input value, i.e., $\widehat{count}_{t\ or\ u}(marker) = \log 10^9 \frac{count_{t\ or\ u}(marker)}{raw\ read\ count\ of\ genome}$, because the value after logarithm transformation was empirically shown to follow the normal distribution and thus proved to have better prediction performance than the value before logarithm transformation.

4. **Generate the read-count profile for a cfDNA sample.** The normalized read counts of all markers are concatenated into a vector profile, and this is used as the input profile for the cfDNA sample.

As aforementioned, this method was applied in two contexts: Type 1 methylation markers use tumor-read deconvolution, and Type 3 methylation markers use tissue-read deconvolution. Specifically, for the tumor-read deconvolution, we assume that a cfDNA read comes from a tumor type, normal plasma, or the unknown class. Therefore, we performed five different tumor-read deconvolution processes, each corresponding to one of five tumor types (COAD, LIHC, LUAD, LUSC, and STAD). Only the tumor read counts, $\widehat{count}_{tumor}\,(marker)$, are used in the input profile. For tissue deconvolution, we assume a cfDNA read comes from one of the normal tissue types or the unknown class. All tissue read counts, $\widehat{count}_{tissue}\,(marker)$, are used in the input profile. In this study, the Type 3 markers generated from RRBS data use 17 normal tissue types (adipose, b-cell, brain, colon, esophagus, granulocyte, heart, kidney, liver, lung, monocyte, neutrophils, pancreas, small intestine, spleen, stomach, and t-cell), while the Type 3 markers generated from 450k array data use 11 normal tissue types (b-cell, colon, granulocyte, kidney, liver, lung, monocyte, neutrophils, pancreas, stomach, and t-cell).

3.3.4.2 Computing Type 2 methylation marker profiles

In each marker, we count the number of reads whose α values are opposite from the normal cfDNA background (figure 3.18b). For example, in Type 2 markers that are hypermethylated in tumors, any hypermethylated read ($\alpha > 50\%$) is counted as a tumor read. This read count is denoted as $\widehat{count}\,(marker)$. We then normalize this read count by the sample's sequencing depth and use its logarithm transformation as the final input value, i.e., $\widehat{count}\,(marker) = \log 10^9 \frac{count(marker)}{Number\ of\ all\ mapped\ reads\ in\ whole\ genome}$, because the value after logarithm transformation was empirically shown to follow the normal distribution and thus proved to have better prediction performance than the value before logarithm transformation.

3.3.4.3 Computing Type 4 methylation marker profiles

These markers are 1MB bins. We used the average methylation level of all repeat regions in each bin to form a vector as the input profile. We then normalized the value in each bin by standardizing it with the 30 reference non-cancer individuals' cfDNA samples, i.e., $\frac{\beta_{sample} - \mu_{reference}}{\sigma_{reference}}$, where β_{sample} is the average methylation level of a bin in the cfDNA sample, and $\mu_{reference}$ and $\sigma_{reference}$ are the mean and standard deviation of the average methylation level in the same bin among the 30 non-cancer reference cfDNA samples.

Age adjustment Methylation profiles are known to vary with age [38, 26, 113]. To offset any age bias between our normal and cancer samples, Type 4 features that were found to be associated with age were removed before classification. Specifically, the R package Weighted Correlation Network Analysis (WGCNA) was used to identify features that were positively and negatively associated with age [53]. The metaAnalysis function in WGCNA calculates two p-values: pValueHighScale and pValueLowScale to find markers that are consistently positively age related (as age increases, methylation increases) and negatively age related (as age increases, methylation decreases). Any markers with p-value less than 0.05 in either direction were removed.

After removal of the age associated features, Type 4 markers are z-score adjusted, i.e., $\frac{L_{sample} - \mu_{reference}}{\sigma_{reference}}$, where L_{sample} is the average methylation of repeat regions falling in a bin in the cfDNA sample, and $\mu_{reference}$ and $\sigma_{reference}$ are the mean and standard deviation of the average methylation of repeat regions falling in the same bin among the 30 reference non-cancer cfDNA samples. Type 1 and 2 markers involve matched tumor and normal tissues in the marker discovery steps. Since markers must be identified in tissues that are the same age, markers cannot be related to age. Therefore, no age adjustment is needed for these features even though they are methylation features. From the original 2,734 1 MB bins in the genome, age adjustment removed on average 267 bins. This resulted in 2,467 markers on average, over 10 random selections of the 41 reference non-cancer samples.

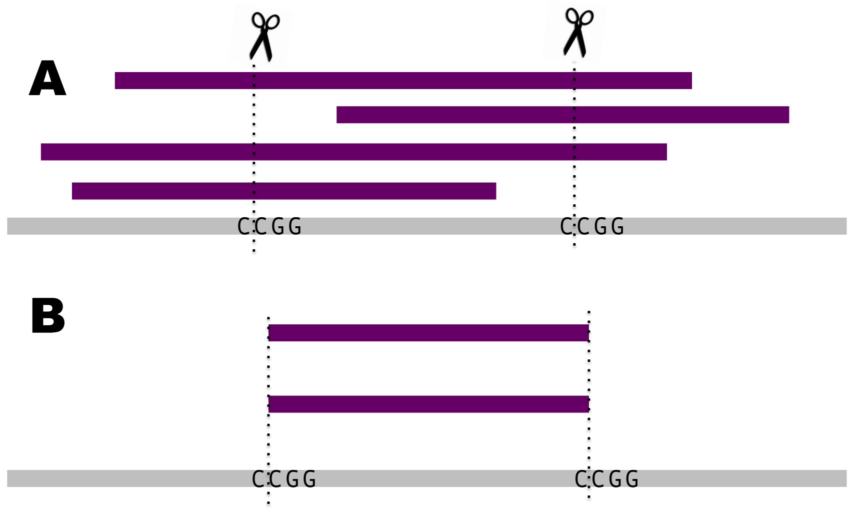

Figure 3.19: Overview of fragment length in cfMethylSeq. (A) cfDNA samples before digestion with MspI, oriented along the genome. CCGG sites are indicated with dotted lines. (B) after MspI digestion in the cfMethylSeq procedure, only two fragments remain. The other two fragments lack the ability to ligate with adapters and will not be sequenced.

3.3.4.4 Computing the cfDNA digestion size profiles

Figure 3.19 illustrates the effect of MspI digestion on fragment length. In our cfMethylSeq procedure, we only see the observed fragments (figure 3.19b) if the initial cfDNA fragment was long enough to cover both CCGG sites. In other papers that use these types of features [74, 66, 42], the fragment lengths would be coming from the initial set of fragments (figure 3.19a). We no longer have the original fragment length information. However, by performing in silico MspI digestion on cfDNA whole-genome sequencing data from Jiang et al. [42], we found that the digested fragment lengths still correlated highly with the original fragment lengths across all samples (mean correlation=0.925, p-value $< 2.2e^{-16}$). This implies that the distance between CCGG sites has an effect on fragment length. In fact, nucleosome positioning, which is thought to determine cfDNA fragment lengths, has been found to be strongly affected by DNA sequence [94].

For our digestion size features, we calculated the average length of all sequencing reads that fell into each 1 MB bin along the hg19 genome. We then normalized the average length in each bin by first dividing by the sample's median fragment length over all 1 MB bins, and then applying a z-score, i.e., $\frac{L_{sample}-\mu_{reference}}{\sigma_{reference}}$, where L_{sample} is the average read length of a bin in the cfDNA sample divided by the sample median, and $\mu_{reference}$ and $\sigma_{reference}$ are the mean and standard deviation of the average read length in the same bin among the 30 reference non-cancer people's cfDNA samples (also adjusted by their medians). Any bins that had no reads across all samples or fell in mitochondrial or sex chromosomes were removed, before normalization.

3.3.4.5 Computing the copy number variation profiles

Using 1 MB bins across the genome, we calculate the normalized read count per bin (number of all reads in a bin divided by the sample's total read count). We then standardize the count by the 30 reference non-cancer cfDNA samples, i.e., $\frac{C_{sample}-\mu_{reference}}{\sigma_{reference}}$, where C_{sample} is the normalized read count per bin in the cfDNA sample, and $\mu_{reference}$ and $\sigma_{reference}$ are the mean and standard deviation of the normalized read count in the same bin among the 30 reference non-cancer cfDNA samples. Bins that had no reads across any samples or fell in mitochondrial or sex chromosomes were removed, before normalization.

3.3.4.6 Computing the microbial profiles

Sequencing reads that did not align to the hg19 reference genome were remapped against 1,620 human-host microbial genomes, including 1,017 bacterial genomes, 1 archaea genome, 453 eukaryota genomes, and 149 viral genomes using Bismark [51] with default parameters. UMI-Grinder [50] was used to remove PCR duplicates based on mapping location and UMIs. After alignment, uniquely mapped reads were counted for each microbial genome. The abundance of a microbe in a cfDNA sample was calculated as the number of reads that uniquely mapped to its genome, divided by the total number of sequencing reads in the sample and the size of the microbial genome. The abundance was then scaled by a large integer (10^9) to avoid small floats. We then normalized each abundance by standardizing it with the 30 reference non-cancer cfDNA samples, i.e., $\frac{\tau_{sample}-\mu_{reference}}{\sigma_{reference}}$, where τ_{sample} is the abundance of a bacterial or viral genome in the cfDNA sample, and $\mu_{reference}$ and $\sigma_{reference}$ are the mean and standard deviation of the abundance in the same bacterial or viral genome among the 30 reference non-cancer cfDNA samples.

3.3.5 Sample split

The 479 cfDNA samples, from 204 non-cancer individuals and 42, 126, 67, 40 liver, lung, colon, and stomach cancer patients, respectively, were split into four sets: marker definition, reference distribution definition, leave-one-out cross validation (LOOCV) training/testing set, and an independent set (figure 3.20). The marker definition set is 41 non-cancer samples used for Type 1 and 2 methylation marker definition. The reference distribution set is 30 non-cancer samples used to normalize observed values using z-scores or to find and remove age bias in Type 4 methylation markers. 75% of the remaining 408 samples were put in the training/testing LOOCV set and the remaining 25% were put into the independent validation set. No samples are shared between sets. For robust performance evaluation, we repeated this split scheme 10 times and reported results averaged over the 10 runs.

3.3.6 Multimodal predictive model integrating heterogeneous and multiscale signal types

The conceptual illustration of the multi-view stacked learning model for cancer detection and TOO prediction is shown as a two-layer structure in figure 3.2 (in the Results section). The predictions from a set of Layer 1 base-models, each separately learned from an individual signal type, are used as input for training a Layer 2 meta-model. However, learning such a two-layer multimodal predictive model may run the risk of overfitting if we simply train the base-models on the full set of training samples, make predictions on the full set of test samples, and then use these predictions to train the meta-model. Therefore, we employed a complicated two-layer learning process, as illustrated in figure 3.21, to overcome the overfitting risk present in the simple learning process [100]. That is, we generate the training data for the Layer 2 meta-model by using only training samples in the base-models

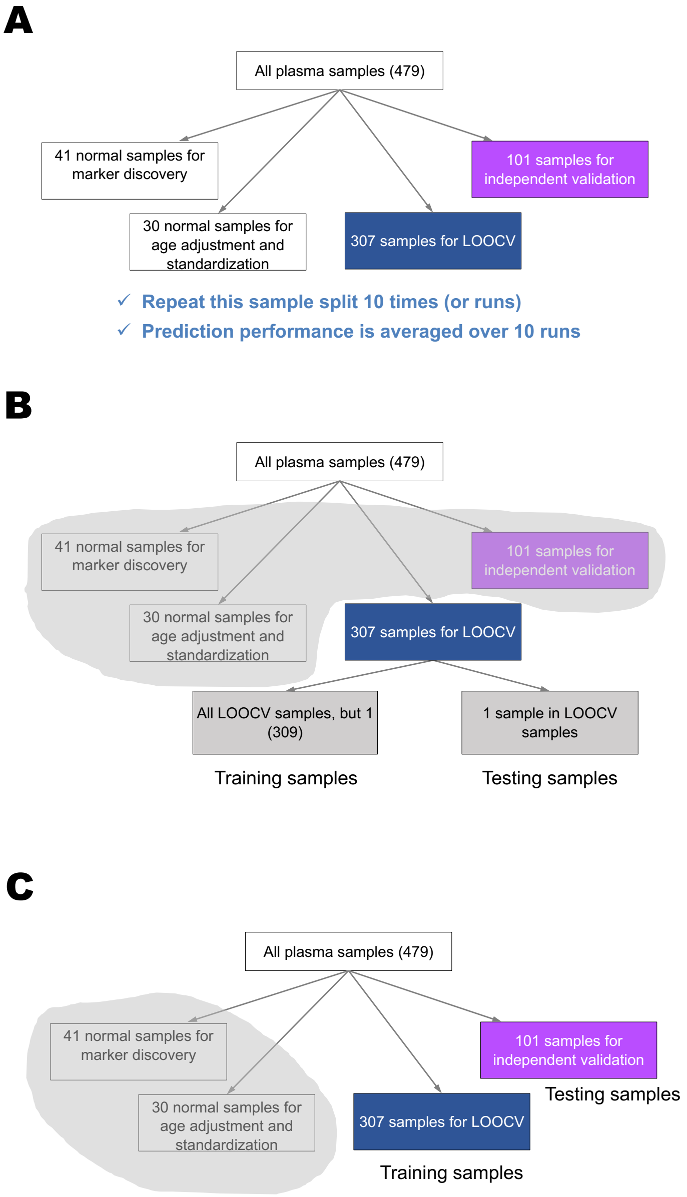
A
All plasma samples (479)
41 normal samples for marker discovery
30 normal samples for age adjustment and standardization
307 samples for LOOCV
101 samples for independent validation
✓ Repeat this sample split 10 times (or runs)
✓ Prediction performance is averaged over 10 runs
B
All plasma samples (479)
41 normal samples for marker discovery
30 normal samples for age adjustment and standardization
307 samples for LOOCV
101 samples for independent validation
All LOOCV samples, but 1 (309)
1 sample in LOOCV samples
Training samples
Testing samples
C
All plasma samples (479)
41 normal samples for marker discovery
30 normal samples for age adjustment and standardization
101 samples for independent validation
Testing samples
307 samples for LOOCV
Training samples

Figure 3.20: Overview of how plasma samples in the cohort are used for training and evaluating the predictive model. (A) Depiction of sample splitting into the 4 sets: marker discovery, age adjustment/standardization, LOOCV, and independent validation, (B) samples used for training and testing during LOOCV, and (C) samples used for training and testing during independent validation.

of Layer 1, avoiding the risk of seeing the test samples in Layer 1. This is implemented by splitting all training samples into 10 non-overlapping folds (i.e. partitions), and training the base model for each signal type on the samples in 9 folds and making predictions on the samples in the remaining 1 fold. This training and prediction process is repeated 10 times, each time using a different left out fold. These predictions, obtained by iterating 9-folds and 1-fold of the training samples, are called “out-of-fold predictions” (OOFPs), indicating a special way of using training samples to generate the prediction scores by themselves. The OOFPs are concatenated to form the new features for all training samples so they can be used as the training data for the Layer 2 meta-model, all without seeing the test samples. To generate the testing data of Layer 2, we train each Layer 1 base-model for each signal type using all training samples (i.e., all 10 folds), and make predictions on all test samples. These predictions then serve as the testing data for the Layer 2 meta-model.

During LOOCV, there is only one testing sample: every sample in the LOOCV set has its own full multimodal predictive model trained for it using all other samples in the LOOCV set. In independent validation, all samples in the independent validation set serve as the testing data, and all samples in the LOOCV set serve as the training data; only one full multimodal predictive model is built. No samples overlap between the LOOCV and independent validation sets.

We implemented two multimodal predictive models, each for a different prediction task:

1. *Model 1: Cancer detection model.* In Layer 1, a base classifier is trained for each of the feature profiles generated from (1) tumor-read counts of RRBS-derived Type 1 markers, (2) tumor-read counts of array-derived Type 1 methylation markers, (3) tumor-read counts of Type 2 hypermethylation markers, (4) tumor-read counts of Type 2 hypomethylation markers, (5) tissue-read counts of RRBS-derived Type 3 methylation markers, (6) tissue-read counts of array-derived Type 3 methylation markers, (7) average methylation rates (β-values) of Type 4 methylation markers, (8) CNV, (9) cfDNA digestion size, and (10) microbial abundances. For the microbial abundance profile, Topçuoğlu et al. [102] has demonstrated that random forest achieves superior predic-

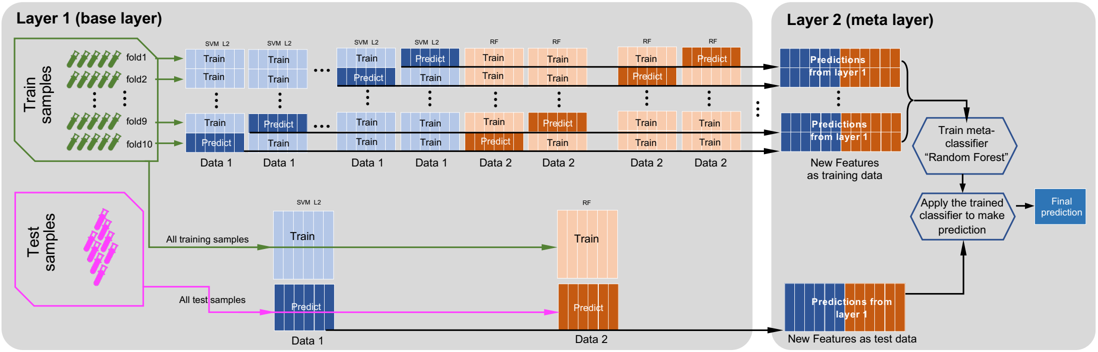

Figure 3.21: Illustrative flowchart of the learning and prediction process using the two-layer stacked model. In the base layer (Layer 1), given the training samples and data types (due to limited space, only two data types are shown here), we split the training samples into 10 equal-size folds to make the "out-of-fold predictions (OOFPs)" that are used as Layer 2 training data. These OOFPs are Layer 1 prediction scores for the left out fold. This OOFP process is repeated 10 times, by choosing each of the 10 folds as the training data and the remaining 9 folds as the OOFPs, until all training samples have OOFPs. The OOFPs for all training samples form the training data for the Layer 2 meta-classifier. Simultaneously, all training samples (i.e., all 10 folds) are used to train base models for use on the testing samples. The testing sample prediction scores from these base-models form the testing data for the Layer 2 meta-classifier.

tion performance. Therefore, in Layer 1 we use a random forest binary classifier with 2000 trees and $mtry = 5\sqrt{\#features}$ (i.e., the number of features randomly sampled as candidates at each node split) for microbial abundances profile, while all other feature profiles in Layer 1 use Linear Support Vector Machine (LSVM) classifiers with the L2 penalty function and the C parameter set to $C = 0.5$. In Layer 2, a random forest model with 2000 trees and $mtry = \sqrt{\#features}$ is used as the binary classifier to combine the predictions scores from all Layer 1 models and make the final prediction. The output prediction score of Layer 2 is the probability of getting cancer; specifically it is the fraction of the 2000 trees that voted "cancer" in the random forest model. For learning and evaluating this model, the plasma cfDNA samples from non-cancer subjects are regarded as the negative class, while those from cancer patients are regarded as the positive class. The higher the prediction score, the more likely the subject has cancer.

2. *Model 2: Cancer typing model:* The same ten feature profiles used in cancer detection are used for TOO prediction. However, instead of using binary classifiers, the one-vs-rest multiclass configuration of the same classifier is used. In Layer 1, a binary classifier is first learned for each class versus all the other classes. Then, all binary classifier predictions are compared, and the class with the highest prediction score is the most likely predicted class. Specifically, all feature profiles in Layer 1 use the LSVM-based multi-class classifier with the L2 penalty function and the C parameter $C = 0.5$, except for the microbial composition profile which uses a random forest multi-class classifier with 2000 trees and mtry=$5\sqrt{\#features}$. In Layer 2, a random forest multi-class classifier with 2000 trees and mtry=$\sqrt{\#features}$ is used to combine the predictions scores from all Layer 1 multiclass classifiers to make the final prediction. The output prediction score of Layer 2 is the cancer-type-membership probability for each cancer type; the cancer type with the highest membership probability is the predicted cancer type. For learning and evaluating this model, the plasma cfDNA samples of cancer patients of all five cancer types (COAD, LIHC, LUAD, LUSC, and STAD) are predicted to be from one of four classes: colon cancer (COAD), liver cancer

(LIHC), lung cancer (LUAD or LUSC), and stomach cancer (STAD).

These two models were performed sequentially, each corresponding to a prediction task:

1. We applied the cancer detection model (Model 1) to predict if a subject has cancer or not. When the prediction score is less than a threshold, the subject is predicted as non-cancer, otherwise as cancer. We selected the threshold to be one false positive non-cancer sample; this translates to 99% specificity in the LOOCV set and 97% specificity in the independent validation set.

2. We performed cancer typing only for the cancer samples that were predicted to have cancer in step 1. In this step, the cancer typing model (Model 2) predicts four cancer-type-membership probabilities, each corresponding to a cancer type (colon, liver, lung and stomach cancer). The cancer type with the largest membership probability is the cancer type predicted by the model. However, in some cases two or more cancer types receive similar large membership probabilities, indicating the predictive model cannot effectively determine the tissue of origin. To alleviate this, we used the fold change between the highest membership probability and the second highest membership probability as a metric to measure cancer type prediction confidence. The higher this confidence is, the more certain we are in the cancer type prediction. No cancer type prediction was assigned for those patients whose cancer typing confidence was less than a threshold of 2. For samples with high cancer typing confidence, the predicted TOO is the cancer type with the highest cancer-type-membership probability.

3.3.7 Performance evaluation of multimodal predictive models

The cancer detection and typing models are evaluated using either LOOCV or independent validation. For cancer detection (Step 1), the AUROC (Area Under the Receiver Operating Characteristic curve) and the sensitivity at a certain specificity are the two most popular performance metrics to assess binary classification [21]. For cancer typing (Step 2), the overall accuracy, i.e., accuracy=$\frac{\#correctly\ predicted\ samples}{total\ \#\ samples}$ is the most widely used measure [91].

We use a confusion matrix to further break down the overall accuracy into specific cancer types for the correctly and incorrectly predicted samples. Using the confusion matrix, we can also calculate the precision for each cancer type, defined as $precision(cancer\ type) = \frac{\#samples\ correctly\ predicted\ as\ this\ cancer\ type}{total\ \#\ samples\ predicted\ as\ this\ cancer\ type}$. Due to the limited sample size in the independent validation set, here we generated the confusion matrix by accumulating scores over all 10 sets.

3.3.7.1 Reduced training size and marker number evaluation

For evaluation on the independent set, one model is trained on all samples that make up the LOOCV set as described above. This one model is applied to every sample in the independent set. To see if the training size affected the prediction performance, subsets of the LOOCV set were used to train the models. All models were still applied to the full independent set. For example, the LOOCV set is 307 samples; full independent results (100% training size) have one model trained on all 307 samples. To measure performance at 20% training size, 61 of the 307 LOOCV set samples are randomly selected to train a full model. The model is still trained in the same way, just using the reduced number of training samples. This one final model is applied to all samples in the independent set. This reduction in training size was performed 10 times, once for each random split, for proportions 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%.

3.3.8 Complementarity analysis

To see if different features types were complementary to each other, we wondered whether there were some samples that could only be classified by one feature type but not others. Since there are four types of methylation features, a stacked classifier was used that combined all methylation features. Specifically, the same framework for cancer detection (Model 1) as described above is used, but instead of using all 10 feature types, the CNV, microbial, and digestion size features are removed. The output of this stacked classifier (using random forest) was used to compare to Layer 1 results from CNV (SVM L2), digestion size (SVM

L2), and microbial features (random forest).

For cancer detection, a sample was considered classified correctly by a base level classifier if it was correctly predicted at a false positive rate of 5. Each sample in each of the 10 random splits was evaluated for correct predictions using each base level classifier. The graph in figure 3.8a shows, for each sample in each split, how many were classified correctly by each feature combination. There are 10 splits, with about 307 LOOCV samples in each split. Each split/sample is treated independently even though the samples could overlap between splits (i.e., the numbers add up to around 3070 even though there are not this many unique samples).

3.4 Discussion

Due to the heterogeneous nature of cancer and the wide range of features that have been used to perform cancer detection and typing in cfDNA, it is evident that a highly successful method should aim to combine as many disparate features as possible. In this chapter, we develop the first integrative cancer detection and typing model incorporating these diverse aspects of cancer. This is only possible because of the experimental protocol developed in chapter 2. The methylation features used herein use both small and large scale genomic regions. Currently the only commercially available technology to obtain such measurements in cfDNA is WGBS, which would have been cost-prohibitive for a study of this size. Alternatively, if a targeted panel were used to capture the small, highly specific methylation features, we would sacrifice all broad information contained in the digestion size, copy number, microbial, and Type 4 methylation markers which provide key complementary information for our model.

Inspired by the highly sensitive PCR-based cancer detection methods in the literature, we designed the Type 2 methylation markers: a novel kind of methylation feature that mimics this effective PCR approach. These PCR-based methods often target gene promoters. Since our cfMethylSeq procedure covers roughly 75% of gene promoters and CpG islands at moderate depth, we reasoned we could apply this principle at a much larger scale. Using our previously published α value [59] in conjunction with this idea allowed us to identify markers that would be undiscoverable using population-level average methylation measurements such as 450k. By using other broad measurements, as well as capturing broad signals from large genomic regions, our method can capture pervasive changes observed in cancer [12]. Combining all these features together using an ensemble learning approach, the stacked classifier offsets any bias in any individual markers. Ultimately, we were able to achieve $85.6 \pm 6.7\%$ sensitivity at 99% specificity for cancer detection. This is considerably higher than results from other groups such as cancerSEEK [14] (55.1% sensitivity) or Grail [65] (62% sensitivity).

In total, our integrated experimental and computational system, CancerRadar, overcomes major challenges in cfDNA-based early cancer detection including the low fraction of tumor DNA in cfDNA and the molecular heterogeneity of cancer. Our cfMethylSeq assay not only cost-effectively captures the cfDNA methylome, but also provides genome-wide profiles of tumor CNV and cfDNA fragment lengths, as well as microbial abundances in blood. Since our patient cohort is dominated by non-metastatic cancer patients, our data demonstrates the feasibility of using CancerRadar in a screening setting. Note that our control samples are not restricted to healthy individuals, but also include patients of various non-cancer diseases (e.g. cirrhosis, pancreatitis, hepatitis, diabetes, etc.), reflecting practical scenarios.

Although methylation, CNV, cfDNA fragment size, and microbial abundances have each been used to detect cancer in the literature [64, 15, 46, 117, 27, 121, 82, 76] this study is the first to systematically compare their prediction power using cfDNA samples from the same cancer cohort. Among all the features used, as expected, methylation contributed the most information for detecting and locating cancer. To exploit the power of methylation for cancer detection and typing, we integrated our own data with a large amount of public data to identify four types of methylation markers with different characteristics; and we expanded our previous read-level deconvolution algorithm [59] to further enhance its power in accurately identifying trace tumor signals and also to identify elevated tissue cfDNA signals. Next to methylation, the second most powerful feature for cancer detection is cfDNA digestion size, which reflects cfDNA fragment length. CNV and microbiome signatures also helped to increase the prediction power, especially in late stage patients.

Our data showed that as training sample sizes increase, the detection power of CancerRadar increases, and so do the numbers of required markers. The CancerRadar system assesses a comprehensive set of biomarkers, encompassing the large epigenetic and genetic landscape of diverse cancer etiologies, allowing continued refinement and expansion as training cohorts grow. In fact, the comprehensive information provided by cfMethylSeq can potentially serve multiple purposes in diagnosis and prognosis, such as predicting disease progression, stage, etiology, or use in therapy optimization.

CHAPTER 4

Conclusions

Early detection of cancer holds our greatest hope for increasing cancer survival. Liquid biopsies based on the presence of tumor-derived cfDNA in the blood are an attractive means to achieve this goal. While cfDNA is already clinically used for noninvasive prenatal testing [44], its clinical use for cancer screening has faced major challenges. Namely, the amount of tumor-derived cfDNA in the blood is very small, requiring sophisticated experimental and/or computational techniques to increase or identify tumor signal. In addition, cancer is a heterogeneous disease encompassing broad and specific genomic changes. Experimental techniques often suffer a tradeoff between profiling a small panel of regions in depth or shallowly surveying the genome. cfDNA has displayed a number of unique biological features, such as its fragmented nature being related to nucleosome positioning [90], gene expression [104], and 3D genome organization [66], but at the same time the poorly understood origins and properties of cfDNA have hindered library preparation techniques and sometimes led to unexpected results [107]. In this work, we present an integrated computational and experimental system that addresses these challenges.

In chapter 2, we develop cfMethylSeq, a novel protocol for profiling methylation in cfDNA. cfMethylSeq, an adaptation of RRBS for cfDNA, profiles CpG-dense regions of the genome, yielding more than 12 fold enrichment over WGBS in CGIs. While RRBS has been used to profile methylation cost-effectively in solid tissues, its use in cfDNA is limited because of cfDNA's fragmented nature. Our protocol adapts the RRBS technique to work on fragmented DNA. We showed that cfMethylSeq profiled the regions of interest, reliably called methylation, and was reproducible. Our cfMethylSeq procedure enables cfDNA methylomes to be profiled inexpensively.

In chapter 3, we make use of this cfMethylSeq data to detect and type cancer in cfDNA. A major advantage of our cfMethylSeq method is its ability to profile both highly specific methylation features as well as broad genomic features. We integrated these features together using an ensemble machine learning algorithm, and achieved 89.1% sensitivity in cancer detection at 97% specificity, with an overall AUC of 0.986 in the independent validation set, and an accuracy of 91.5% in locating the cancer's tissue of origin. Our results outperform existing methods [15, 65], and our method easily scales to larger sample sizes and training sets.

Although our results are promising, there is still room for improvement in many areas. While the cfMethylSeq procedure presented in chapter 2 offers much-needed cost effective methylation profiling in cfDNA, the input amounts required to build the library (10ng) are still relatively large. Improvements could be made by further optimizing the protocol to prevent sample loss, such as using enzymatic alternatives to bisulfite conversion [115]. Additionally, the current adapter synthesis strategy produces fragments that all have the sequence "TGACT" in bases 9-13. Non-random distribution of bases in the beginning of reads is known to cause issues with cluster localization on Illumina machines [9]. Using PhiX to increase diversity could mitigate this, but this will raise sequencing costs. Others have used dark cycles to overcome this issue, where these non-random bases are essentially skipped during sequencing [9]. However, in our cfMethylSeq procedure and in traditional RRBS, the sequence "CGG" or "TGG" appears at the beginning of each read (after the UMI for cfMethylSeq). Since this low-diversity sequence contains an informative CpG site, it cannot be skipped using a dark cycle. Alternatively, we could use multiple adapters that have different UMI lengths, e.g. some 7 bp or some 6 bp. Mixing these adapters together would shift the fixed sequence in some of the reads, and avoid the Illumina machine reading the same sequence in cycles 9-13. Bioinformatically, the length of the original UMI can be deduced by finding the location of the fixed sequence "TGACT" in the read.

Coverage of desired genomic regions in cfMethylSeq or traditional RRBS can be manipulated by using restriction enzymes other than or in addition to MspI [24]. For example, while MspI yields the highest CpG:fragment ratio of any single enzyme, HaeIII (GG↓CC) comes in second, and profiles about twice as many CpGs. Note that since HaeIII does not contain a CpG site in its cut site, every fragment is no longer guaranteed to contain a CpG site, thereby increasing sequencing cost and leading to wasted data [69]. Alternatively, a double enzyme digest can be used. It has been found that digestion with MspI and ApekI (C↓CWGC) can yield increased coverage in CpG islands, CGI shores, and introns. Drawbacks of the double-enzyme approach are the increased sequencing cost and greater fragmentation of the MspI digested fragments [108]. In principle, the different enzymes or combination of enzymes could be tailored to the needs of the experiment.

While the results of our computational cancer detection and typing framework presented in chapter 3 outperform existing methods, we caution that our sample sizes could be increased. Our LOOCV set contains on average 100 non-cancer samples and 207 cancer samples, while our independent validation sets contain on average 34 non-cancer samples and 67 cancer samples. The small number of non-cancer samples in the independent set especially hinders our ability to evaluate our results at high specificities. Additional samples at high risk for certain cancer types, such as those with HBV for liver cancer [12] or those with benign lung nodules for lung cancer [70] could be included in the non-cancer samples in greater numbers to better reflect real-world screening scenarios. All of our cfMethylSeq and RRBS libraries were prepared by two individuals and sequenced on HiSeqX Illumina machines. In the future, to reflect more practical settings of a widespread clinical test, it would be better to use combinations of different sequencing machines, e.g. NovaSeq 6000 to ensure the method is robust to any differences caused by technical specifications of the sequencers (e.g. dye chemistries) and their resulting biases (e.g. read quality scores, depth of coverage) [122]. Different individuals at different labs preparing the cfMethylSeq and RRBS libraries could also serve a similar purpose.

The stacked classifier framework could be improved by adding additional machine learning methods in the Layer 1 classifiers. Currently all features use SVM classifiers, with the exception of microbial features which use random forest. Classifiers such as logistic regression, XGBoost, K nearest neighbors, and/or Naive Bayes with different parameter settings could be used in addition to SVM and random forest in Layer 1. Then, rather than using the output of a random forest model in Layer 2 as the final prediction score, multiple classifiers could be used in Layer 2 and then integrated together in a third layer for the final prediction score. In this way, in addition to the different features providing complementary information for each sample, the different machine learning methods offer multiple view points of the same feature, further boosting accuracy. Covariates such as gender and age could be added to models to further reduce their bias. When collecting future samples, it would be advantageous to collect as much clinical information as possible, allowing for additional covariates such as BMI or smoking history to be incorporated into the models as well.

Although cfDNA cancer detection faces many difficulties, much progress has been made in its entrance into the clinical space. Herein, we presented an experimental and computational method that will bring us closer to making clinical cfDNA cancer screening a reality.

REFERENCES

[1] Charu C. Aggarwal. *Data Classification: Algorithms and Applications.* Chapman Hall/CRC, 1st edition, 2014.

[2] Catherine Alix-Panabieres. Perspective: The future of liquid biopsy. *Nature*, 579(7800):S9–S9, 2020.

[3] Adrian Ally, Miruna Balasundaram, Rebecca Carlsen, Eric Chuah, Amanda Clarke, Noreen Dhalla, Robert A Holt, Steven JM Jones, Darlene Lee, Yussanne Ma, et al. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*, 169(7):1327–1341, 2017.

[4] Paul Angulo, David E Kleiner, Sanne Dam-Larsen, Leon A Adams, Einar S Bjornsson, Phunchai Charatcharoenwitthaya, Peter R Mills, Jill C Keach, Heather D Lafferty, Alisha Stahler, et al. Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. *Gastroenterology*, 149(2):389–397, 2015.

[5] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995, 2012.

[6] Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M Le, David Delano, Lu Zhang, Gary P Schroth, Kevin L Gunderson, et al. High density dna methylation array with single cpg site resolution. *Genomics*, 98(4):288–295, 2011.

[7] Babraham Bioinformatics. Reduced representation bisulfite-seq –a brief guide to rrbs. `http://www.bioinformatics.babraham.ac.uk/projects/bismark/RRBS_Guide.pdf`. Accessed: 2020-11-18.

[8] Babraham Bioinformatics. Trim galore. `https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`.

[9] Patrick Boyle, Kendell Clement, Hongcang Gu, Zachary D Smith, Michael Ziller, Jennifer L Fostel, Laurie Holmes, Jim Meldrim, Fontina Kelley, Andreas Gnirke, et al. Gel-free multiplexed reduced representation bisulfite sequencing for large-scale dna methylation profiling. *Genome biology*, 13(10):R92, 2012.

[10] Abel Jacobus Bronkhorst, Vida Ungerer, and Stefan Holdenrieder. The emerging role of cell-free dna as a molecular marker for cancer management. *Biomolecular detection and quantification*, 17:100087, 2019.

[11] Jacob J Chabon, Emily G Hamilton, David M Kurtz, Mohammad S Esfahani, Everett J Moding, Henning Stehr, Joseph Schroers-Martin, Barzin Y Nabet, Binbin Chen, Aadel A Chaudhuri, et al. Integrating genomic features for non-invasive early lung cancer detection. *Nature*, 580(7802):245–251, 2020.

[12] KC Allen Chan, Peiyong Jiang, Carol WM Chan, Kun Sun, John Wong, Edwin P Hui, Stephen L Chan, Wing Cheong Chan, David SC Hui, Simon SM Ng, et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma dna bisulfite sequencing. *Proceedings of the National Academy of Sciences*, 110(47):18761–18768, 2013.

[13] Deanna M Church, Valerie A Schneider, Tina Graves, Katherine Auger, Fiona Cunningham, Nathan Bouk, Hsiu-Chuan Chen, Richa Agarwala, William M McLaren, Graham RS Ritchie, et al. Modernizing reference genome assemblies. *PLoS biology*, 9(7):e1001091, 2011.

[14] Joshua D Cohen, Lu Li, Yuxuan Wang, Christopher Thoburn, Bahman Afsari, Ludmila Danilova, Christopher Douville, Ammar A Javed, Fay Wong, Austin Mattox, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, 359(6378):926–930, 2018.

[15] Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel, Vilmos Adleff, Daniel C Bruhm, Sarah Østrup Jensen, Jamie E Medina, Carolyn Hruban, James R White, et al. Genome-wide cell-free dna fragmentation in patients with cancer. *Nature*, 570(7761):385–389, 2019.

[16] Emily Crowley, Federica Di Nicolantonio, Fotios Loupakis, and Alberto Bardelli. Liquid biopsy: monitoring cancer-genetics in the blood. *Nature reviews Clinical oncology*, 10(8):472, 2013.

[17] Andries De Koker, Ruben Van Paemel, Bram De Wilde, Katleen De Preter, and Nico Callewaert. A versatile method for circulating cell-free dna methylome profiling by reduced representation bisulfite sequencing. *bioRxiv*, page 663195, 2019.

[18] Giorgia Del Vecchio, Qingjiao Li, Wenyuan Li, Shanthie Thamotharan, Anela Tosevska, Marco Morselli, Kyunghyun Sung, Carla Janzen, Xianghong Zhou, Matteo Pellegrini, et al. Cell-free dna methylation and transcriptomic signature prediction of pregnancies with adverse outcomes. *Epigenetics*, pages 1–20, 2020.

[19] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606, 2017.

[20] Melanie Ehrlich. Dna hypomethylation in cancer cells. *Epigenomics*, 1(2):239–259, 2009.

[21] Christopher M Florkowski. Sensitivity, specificity, receiver-operating characteristic (roc) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical Biochemist Reviews*, 29(Suppl 1):S83, 2008.

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

[23] María Gallardo-Gómez, Sebastian Moran, María Páez de la Cadena, Vicenta Soledad Martínez-Zorzano, Francisco Javier Rodríguez-Berrocal, Mar Rodríguez-Girondo, Manel Esteller, Joaquín Cubiella, Luis Bujanda, Antoni Castells, et al. A new approach to epigenome-wide discovery of non-invasive methylation biomarkers for colorectal cancer screening in circulating cell-free dna using pooled samples. *Clinical epigenetics*, 10(1):1–10, 2018.

[24] Hongcang Gu, Zachary D Smith, Christoph Bock, Patrick Boyle, Andreas Gnirke, and Alexander Meissner. Preparation of reduced representation bisulfite sequencing libraries for genome-scale dna methylation profiling. *Nature protocols*, 6(4):468–481, 2011.

[25] Shicheng Guo, Dinh Diep, Nongluk Plongthongkum, Ho-Lim Fung, Kang Zhang, and Kun Zhang. 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):635–642, 2017.

[26] Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, 49(2):359–367, 2013.

[27] Xiaoke Hao, Huiyan Luo, Michal Krawczyk, Wei Wei, Wenqiu Wang, Juan Wang, Ken Flagg, Jiayi Hou, Heng Zhang, Shaohua Yi, et al. Dna methylation markers for diagnosis and prognosis of common cancers. *Proceedings of the National Academy of Sciences*, 114(28):7414–7419, 2017.

[28] Timothy Hardy, Mujdat Zeybel, Christopher P Day, Christian Dipper, Steven Masson, Stuart McPherson, Elsbeth Henderson, Dina Tiniakos, Steve White, Jeremy French, et al. Plasma dna methylation: a potential biomarker for stratification of liver fibrosis in non-alcoholic fatty liver disease. *Gut*, 66(7):1321–1328, 2017.

[29] Jennifer Harrow, France Denoeud, Adam Frankish, Alexandre Reymond, Chao-Kung Chen, Jacqueline Chrast, Julien Lagarde, James GR Gilbert, Roy Storey, David Swarbreck, et al. Gencode: producing a reference annotation for encode. *Genome biology*, 7(1):1–9, 2006.

[30] Yutaka Hashimoto, Timothy J Zumwalt, and Ajay Goel. Dna methylation patterns as noninvasive biomarkers and targets of epigenetic therapies in colorectal cancer. *Epigenomics*, 8(5):685–703, 2016.

[31] Hikoya Hayatsu, Yusuke Wataya, Kazushige Kai, and Shigeru Iida. Reaction of sodium bisulfite with uracil, cytosine, and their derivatives. *Biochemistry*, 9(14):2858–2865, 1970.

[32] Ellen Heitzer and Michael R Speicher. One size does not fit all: size-based plasma dna diagnostics. *Science Translational Medicine*, 10(466), 2018.

[33] Ellen Heitzer, Peter Ulz, Jelena Belic, Stefan Gutschi, Franz Quehenberger, Katja Fischereder, Theresa Benezeder, Martina Auer, Carina Pischler, Sebastian Mannweiler, et al. Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. *Genome medicine*, 5(4):1–16, 2013.

[34] James G Herman, Jeremy R Graff, SBDN Myöhänen, Barry D Nelkin, and Stephen B Baylin. Methylation-specific pcr: a novel pcr assay for methylation status of cpg islands. *Proceedings of the national academy of sciences*, 93(18):9821–9826, 1996.

[35] Holger Heyn and Manel Esteller. Dna methylation profiling in the clinic: applications and challenges. *Nature Reviews Genetics*, 13(10):679–692, 2012.

[36] Benjamin Hing, Enrique Ramos, Patricia Braun, Melissa McKane, Dubravka Jancic, Kellie LK Tamashiro, Richard S Lee, Jacob J Michaelson, Todd E Druley, and James B Potash. Adaptation of the targeted capture methyl-seq platform for the mouse genome identifies novel tissue-specific dna methylation patterns of genes involved in neurodevelopment. *Epigenetics*, 10(7):581–596, 2015.

[37] Emily Hodges, Andrew D Smith, Jude Kendall, Zhenyu Xuan, Kandasamy Ravi, Michelle Rooks, Michael Q Zhang, Kenny Ye, Arindam Bhattacharjee, Leonardo Brizuela, et al. High definition profiling of mammalian dna methylation by array capture and single molecule bisulfite sequencing. *Genome research*, 19(9):1593–1605, 2009.

[38] Steve Horvath, Paolo Garagnani, Maria Giulia Bacalini, Chiara Pirazzini, Stefano Salvioli, Davide Gentilini, Anna Maria Di Blasio, Cristina Giuliani, Spencer Tung, Harry V Vinters, et al. Accelerated epigenetic aging in down syndrome. *Aging cell*, 14(3):491–495, 2015.

[39] Øistein Ihle and Terje E Michaelsen. Efficient purification of dna fragments using a protein binding membrane. *Nucleic Acids Research*, 28(16):e76–e76, 2000.

[40] Infinium®. Field guide to methylation methods.

[41] Taylor J Jensen, Sung K Kim, Zhanyang Zhu, Christine Chin, Claudia Gebhard, Tim Lu, Cosmin Deciu, Dirk van den Boom, and Mathias Ehrich. Whole genome bisulfite sequencing of cell-free dna and its cellular contributors uncovers placenta hypomethylated domains. *Genome biology*, 16(1):78, 2015.

[42] Peiyong Jiang, Carol WM Chan, KC Allen Chan, Suk Hang Cheng, John Wong, Vincent Wai-Sun Wong, Grace LH Wong, Stephen L Chan, Tony SK Mok, Henry LY Chan, et al. Lengthening and shortening of plasma dna in hepatocellular carcinoma patients. *Proceedings of the National Academy of Sciences*, 112(11):E1317–E1325, 2015.

[43] Peiyong Jiang, KC Allen Chan, and YM Dennis Lo. Liver-derived cell-free nucleic acids in plasma: Biology and applications in liquid biopsies. *Journal of hepatology*, 71(2):409–421, 2019.

[44] Peiyong Jiang and YM Dennis Lo. The long and short of circulating cell-free dna and the ins and outs of molecular diagnostics. *Trends in Genetics*, 32(6):360–371, 2016.

[45] Peiyong Jiang, Tingting Xie, Spencer C Ding, Ze Zhou, Suk Hang Cheng, Rebecca WY Chan, Wing-Shan Lee, Wenlei Peng, John Wong, Vincent WS Wong, et al. Detection and characterization of jagged ends of double-stranded dna in plasma. *Genome Research*, 30(8):1144–1153, 2020.

[46] Shuli Kang, Qingjiao Li, Quan Chen, Yonggang Zhou, Stacy Park, Gina Lee, Brandon Grimes, Kostyantyn Krysan, Min Yu, Wei Wang, et al. Cancerlocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free dna. *Genome biology*, 18(1):53, 2017.

[47] Donna Karolchik, Angela S Hinrichs, Terrence S Furey, Krishna M Roskin, Charles W Sugnet, David Haussler, and W James Kent. The ucsc table browser data retrieval tool. *Nucleic acids research*, 32(suppl_1):D493–D496, 2004.

[48] Scott R Kennedy, Michael W Schmitt, Edward J Fox, Brendan F Kohrn, Jesse J Salk, Eun Hyun Ahn, Marc J Prindle, Kawai J Kuong, Jiang-Cheng Shen, Rosa-Ana Risques, et al. Detecting ultralow-frequency mutations by duplex sequencing. *Nature protocols*, 9(11):2586, 2014.

[49] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

[50] Felix Krueger. Umi-grinder. `https://github.com/FelixKrueger/Umi-Grinder`.

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

[52] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[53] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[54] Eun-Joon Lee, Junfeng Luo, James M Wilson, and Huidong Shi. Analyzing the cancer methylome through targeted bisulfite sequencing. *Cancer letters*, 340(2):171–178, 2013.

[55] Jong-Hun Lee, Sung-Joon Park, and Kenta Nakai. Differential landscape of non-cpg methylation in embryonic stem cells and neurons caused by dnmt3s. *Scientific reports*, 7(1):1–11, 2017.

[56] Roni Lehmann-Werman, Daniel Neiman, Hai Zemmour, Joshua Moss, Judith Magenheim, Adi Vaknin-Dembinsky, Sten Rubertsson, Bengt Nellgård, Kaj Blennow, Henrik Zetterberg, et al. Identification of tissue-specific cell death using methylation patterns of circulating dna. *Proceedings of the National Academy of Sciences*, 113(13):E1826–E1834, 2016.

[57] SA Leon, B Shapiro, DM Sklaroff, and MJ Yaros. Free dna in the serum of cancer patients and the effect of therapy. *Cancer research*, 37(3):646–650, 1977.

[58] Qing Li, Masako Suzuki, Jennifer Wendt, Nicole Patterson, Steven R Eichten, Peter J Hermanson, Dawn Green, Jeffrey Jeddeloh, Todd Richmond, Heidi Rosenbaum, et al. Post-conversion targeted capture of modified cytosines in mammalian and plant genomes. *Nucleic acids research*, 43(12):e81–e81, 2015.

[59] Wenyuan Li, Qingjiao Li, Shuli Kang, Mary Same, Yonggang Zhou, Carol Sun, Chun-Chi Liu, Lea Matsuoka, Linda Sher, Wing Hung Wong, et al. Cancerdetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free dna methylation sequencing data. *Nucleic acids research*, 46(15):e89–e89, 2018.

[60] Wenhua Liang, Yue Zhao, Weizhe Huang, Yangbin Gao, Weihong Xu, Jinsheng Tao, Meng Yang, Lequn Li, Wei Ping, Hui Shen, et al. Non-invasive diagnosis of early-stage lung cancer using high-throughput targeted dna methylation sequencing of circulating tumor dna (ctdna). *Theranostics*, 9(7):2056, 2019.

[61] Zachary Lippman, Anne-Valérie Gendrel, Michael Black, Matthew W Vaughn, Neilay Dedhia, W Richard McCombie, Kimberly Lavine, Vivek Mittal, Bruce May, Kristin D Kasschau, et al. Role of transposable elements in heterochromatin and epigenetic control. *Nature*, 430(6998):471–476, 2004.

[62] Delphine Lissa and Ana I Robles. Methylation analyses in liquid biopsy. *Translational lung cancer research*, 5(5):492, 2016.

[63] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. Human dna methylomes at base resolution show widespread epigenomic differences. *nature*, 462(7271):315–322, 2009.

[64] MC Liu, GR Oxnard, EA Klein, C Swanton, MV Seiden, Minetta C Liu, Geoffrey R Oxnard, Eric A Klein, David Smith, Donald Richards, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free dna. *Annals of Oncology*, 2020.

[65] Minetta C Liu, Arash Jamshidi, Oliver Venn, Alexander P Fields, M Cyrus Maher, Gordon Cann, Hamed Amini, Samuel Gross, Joerg Bredno, Meredith Miller, et al. Genome-wide cell-free dna (cfdna) methylation signatures and effect on tissue of origin (too) performance., 2019.

[66] Yaping Liu, Tzu-Yu Liu, David E Weinberg, Brandon W White, Chris J De La Torre, Catherine L Tan, Anthony D Schmitt, Siddarth Selvaraj, Vy Tran, Louise C Laurent, et al. Spatial co-fragmentation pattern of cell-free dna recapitulates in vivo chromatin organization and identifies tissues-of-origin. *BioRxiv*, page 564773, 2019.

[67] YM Dennis Lo, Noemi Corbetta, Paul F Chamberlain, Vik Rai, Ian L Sargent, Christopher WG Redman, and James S Wainscoat. Presence of fetal dna in maternal plasma and serum. *The lancet*, 350(9076):485–487, 1997.

[68] P Mandel. Les acides nucleiques du plasma sanguin chez 1 homme. *CR Seances Soc Biol Fil*, 142:241–243, 1948.

[69] Daniel B Martinez-Arguelles, Sunghoon Lee, and Vassilios Papadopoulos. In silico analysis identifies novel restriction enzyme combinations that expand reduced representation bisulfite sequencing cpg coverage. *BMC research notes*, 7(1):534, 2014.

[70] Annette McWilliams, Martin C Tammemagi, John R Mayo, Heidi Roberts, Geoffrey Liu, Kam Soghrati, Kazuhiro Yasufuku, Simon Martel, Francis Laberge, Michel Gingras, et al. Probability of cancer in pulmonary nodules detected on first screening ct. *New England Journal of Medicine*, 369(10):910–919, 2013.

[71] Alexander Meissner, Andreas Gnirke, George W Bell, Bernard Ramsahoye, Eric S Lander, and Rudolf Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. *Nucleic acids research*, 33(18):5868–5877, 2005.

[72] Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz. Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology*, 12(4):R41, 2011.

[73] Joshua Moss, Judith Magenheim, Daniel Neiman, Hai Zemmour, Netanel Loyfer, Amit Korach, Yaacov Samet, Myriam Maoz, Henrik Druid, Peter Arner, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free dna in health and disease. *Nature communications*, 9(1):1–12, 2018.

[74] Florent Mouliere, Dineika Chandrananda, Anna M Piskorz, Elizabeth K Moore, James Morris, Lise Barlebo Ahlborn, Richard Mair, Teodora Goranova, Francesco Marass, Katrin Heider, et al. Enhanced detection of circulating tumor dna by fragment size analysis. *Science translational medicine*, 10(466), 2018.

[75] Shalima S Nair, Phuc-Loi Luu, Wenjia Qu, Madhavi Maddugoda, Lily Huschtscha, Roger Reddel, Georgia Chenevix-Trench, Martina Toso, James G Kench, Lisa G Horvath, et al. Guidelines for whole genome bisulphite sequencing of intact and ffpet dna on the illumina hiseq x ten. *Epigenetics & chromatin*, 11(1):24, 2018.

[76] Deborah Nejman, Ilana Livyatan, Garold Fuks, Nancy Gavert, Yaara Zwang, Leore T Geller, Aviva Rotter-Maskowitz, Roi Weiser, Giuseppe Mallel, Elinor Gigi, et al. The

human tumor microbiome is composed of tumor type–specific intracellular bacteria. *Science*, 368(6494):973–980, 2020.

[77] Aaron M Newman, Scott V Bratman, Jacqueline To, Jacob F Wynne, Neville CW Eclov, Leslie A Modlin, Chih Long Liu, Joel W Neal, Heather A Wakelee, Robert E Merritt, et al. An ultrasensitive method for quantitating circulating tumor dna with broad patient coverage. *Nature medicine*, 20(5):548, 2014.

[78] Aaron M Newman, Alexander F Lovejoy, Daniel M Klass, David M Kurtz, Jacob J Chabon, Florian Scherer, Henning Stehr, Chih Long Liu, Scott V Bratman, Carmen Say, et al. Integrated digital error suppression for improved detection of circulating tumor dna. *Nature biotechnology*, 34(5):547, 2016.

[79] Masaki Okano, Daphne W Bell, Daniel A Haber, and En Li. Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

[80] Nelly Olova, Felix Krueger, Simon Andrews, David Oxley, Rebecca V Berrens, Miguel R Branco, and Wolf Reik. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting dna methylation data. *Genome biology*, 19(1):1–19, 2018.

[81] Akira Ooki, Zahra Maleki, Jun-Chieh J Tsay, Chandra Goparaju, Mariana Brait, Nitesh Turaga, Hae-Seong Nam, William N Rom, Harvey I Pass, David Sidransky, et al. A panel of novel detection and prognostic methylated dna markers in primary non–small cell lung cancer and serum dna. *Clinical Cancer Research*, 23(22):7141–7152, 2017.

[82] Gregory D Poore, Evguenia Kopylova, Qiyun Zhu, Carolina Carpenter, Serena Fraraccio, Stephen Wandro, Tomasz Kosciolek, Stefan Janssen, Jessica Metcalf, Se Jin Song, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature*, 579(7800):567–574, 2020.

[83] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[84] Wolf Reik. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425–432, 2007.

[85] Unrivaled Assay Reproducibility. Infinium® humanmethylation450 beadchip.

[86] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[87] Jason P Ross, Keith N Rand, and Peter L Molloy. Hypomethylation of repeated dna sequences in cancer. *Epigenomics*, 2(2):245–269, 2010.

[88] Katja U Schneider, Dimo Dietrich, Michael Fleischhacker, Gunda Leschber, Johannes Merk, Frank Schäper, Henk R Stapert, Erik R Vossenaar, Sabine Weickmann, Volker Liebenberg, et al. Correlation of shox2 gene amplification and dna methylation in lung cancer tumors. *BMC cancer*, 11(1):102, 2011.

[89] Shu Yi Shen, Rajat Singhania, Gordon Fehringer, Ankur Chakravarthy, Michael HA Roehrl, Dianne Chadwick, Philip C Zuzarte, Ayelet Borgida, Ting Ting Wang, Tiantian Li, et al. Sensitive tumour detection and classification using plasma cell-free dna methylomes. *Nature*, 563(7732):579–583, 2018.

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

[91] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[92] Clare Stirzaker, Phillippa C Taberlay, Aaron L Statham, and Susan J Clark. Mining cancer methylomes: prospects and challenges. *Trends in Genetics*, 30(2):75–84, 2014.

[93] Ravid Straussman, Deborah Nejman, Douglas Roberts, Israel Steinfeld, Barak Blum, Nissim Benvenisty, Itamar Simon, Zohar Yakhini, and Howard Cedar. Developmental programming of cpg island methylation profiles in the human genome. *Nature structural & molecular biology*, 16(5):564, 2009.

[94] Kevin Struhl and Eran Segal. Determinants of nucleosome positioning. *Nature structural & molecular biology*, 20(3):267, 2013.

[95] Thomas M Stubbs, Marc Jan Bonder, Anne-Katrien Stark, Felix Krueger, Ferdinand von Meyenn, Oliver Stegle, and Wolf Reik. Multi-tissue dna methylation age predictor in mouse. *Genome biology*, 18(1):1–14, 2017.

[96] Kun Sun, Peiyong Jiang, KC Allen Chan, John Wong, Yvonne KY Cheng, Raymond HS Liang, Wai-kong Chan, Edmond SK Ma, Stephen L Chan, Suk Hang Cheng, et al. 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, 2015.

[97] Kun Sun, Peiyong Jiang, Suk Hang Cheng, Timothy HT Cheng, John Wong, Vincent WS Wong, Simon SM Ng, Brigette BY Ma, Tak Y Leung, Stephen L Chan, et al. Orientation-aware plasma cell-free dna fragmentation analysis in open chromatin regions informs tissue of origin. *Genome research*, 29(3):418–427, 2019.

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

[99] Charles Swanton and Stephan Beck. Epigenetic noise fuels cancer evolution. *Cancer cell*, 26(6):775–776, 2014.

[100] Jiliang Tang, Salem Alelyani, and Huang Liu. Data classification: algorithms and applications. *Data Mining and Knowledge Discovery Series, CRC Press*, pages 37–64, 2014.

[101] Ai Ling Teh, Hong Pan, Xinyi Lin, Yubin Ives Lim, Chinari Pawan Kumar Patro, Clara Yujing Cheong, Min Gong, Julia L MacIsaac, Chee-Keong Kwoh, Michael J Meaney, et al. Comparison of methyl-capture sequencing vs. infinium 450k methylation array for methylome analysis in clinical samples. *Epigenetics*, 11(1):36–48, 2016.

[102] Begüm D Topçuoğlu, Nicholas A Lesniak, Mack T Ruffin, Jenna Wiens, and Patrick D Schloss. A framework for effective application of machine learning to microbiome-based classification problems. *Mbio*, 11(3), 2020.

[103] Cath Tyner. The ucsc genome browser coordinate counting systems. `http://genome.ucsc.edu/blog/the-ucsc-genome-browser-coordinate-counting-systems/`. Accessed: 2019-10-22.

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

[105] Ruben Van Paemel, Andries De Koker, Charlotte Vandeputte, Lieke van Zogchel, Tim Lammens, Geneviève Laureys, Jo Vandesompele, Gudrun Schleiermacher, Mathieu Chicard, Nadine Van Roy, et al. Minimally invasive classification of paediatric solid tumours using reduced representation bisulphite sequencing of cell-free dna: a proof-of-principle study. *Epigenetics*, pages 1–13, 2020.

[106] Joaquim SL Vong, Peiyong Jiang, Suk-Hang Cheng, Wing-Shan Lee, Jason CH Tsang, Tak-Yeung Leung, KC Allen Chan, Rossa WK Chiu, and YM Dennis Lo. Enrichment of fetal and maternal long cell-free dna fragments from maternal plasma following dna repair. *Prenatal diagnosis*, 39(2):88–99, 2019.

[107] Joaquim SL Vong, Jason CH Tsang, Peiyong Jiang, Wing-Shan Lee, Tak Yeung Leung, KC Allen Chan, Rossa WK Chiu, and YM Dennis Lo. Single-stranded dna library preparation preferentially enriches short maternal dna in maternal plasma. *Clinical chemistry*, 63(5):1031–1037, 2017.

[108] Junwen Wang, Yudong Xia, Lili Li, Desheng Gong, Yu Yao, Huijuan Luo, Hanlin Lu, Na Yi, Honglong Wu, Xiuqing Zhang, et al. Double restriction-enzyme digestion improves the coverage and accuracy of genome-wide cpg methylation profiling by reduced representation bisulfite sequencing. *BMC genomics*, 14(1):11, 2013.

[109] Kangli Wang, Xianfeng Li, Shanshan Dong, Jialong Liang, Fengbiao Mao, Cheng Zeng, Honghu Wu, Jinyu Wu, Wanshi Cai, and Zhong Sheng Sun. Q-rrbs: a quantitative reduced representation bisulfite sequencing method for single-cell methylome analyses. *Epigenetics*, 10(9):775–783, 2015.

[110] Peter M Warnecke, Clare Stirzaker, John R Melki, Douglas S Millar, Cheryl L Paul, and Susan J Clark. Detection and measurement of pcr bias in quantitative methylation analysis of bisulphite-treated dna. *Nucleic acids research*, 25(21):4422–4426, 1997.

[111] Jorja D Warren, Wei Xiong, Ashley M Bunker, Cecily P Vaughn, Larissa V Furtado, William L Roberts, John C Fang, Wade S Samowitz, and Karen A Heichman. Septin 9 methylated dna is a sensitive and specific blood test for colorectal cancer. *BMC medicine*, 9(1):133, 2011.

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

[113] Carola Ingrid Weidner, Qiong Lin, Carmen Maike Koch, Lewin Eisele, Fabian Beier, Patrick Ziegler, Dirk Olaf Bauerschlag, Karl-Heinz Jöckel, Raimund Erbel, Thomas Walter Mühleisen, et al. Aging of blood can be tracked by dna methylation changes at just three cpg sites. *Genome biology*, 15(2):R24, 2014.

[114] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.

[115] Louise Williams, Yanxia Bei, Heidi E Church, Nan Dai, Eileen T Dimalanta, Laurence M Ettwiller, TC Evans, Bradley W Langhorst, Janine G Borgaro, Shengxi Guan, et al. Enzymatic methyl-seq: the next generation of methylome analysis. *NEB expressions*, 2019.

[116] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[117] Rui-hua Xu, Wei Wei, Michal Krawczyk, Wenqiu Wang, Huiyan Luo, Ken Flagg, Shaohua Yi, William Shi, Qingli Quan, Kang Li, et al. Circulating tumour dna methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nature materials*, 16(11):1155–1161, 2017.

[118] Pengyi Yang, Yee Hwa Yang, Bing B Zhou, and Albert Y Zomaya. A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5(4):296–308, 2010.

[119] Müjdat Zeybel, Timothy Hardy, Yi K Wong, John C Mathers, Christopher R Fox, Agata Gackowska, Fiona Oakley, Alastair D Burt, Caroline L Wilson, Quentin M Anstee, et al. Multigenerational epigenetic adaptation of the hepatic wound-healing response. *Nature medicine*, 18(9):1369–1377, 2012.

[120] Chenzi Zhang, Wenjun Yu, Lin Wang, Mingna Zhao, Qiaomei Guo, Shaogang Lv, Xiaomeng Hu, and Jiatao Lou. Dna methylation analysis of the shox2 and rassf1a panel in bronchoalveolar lavage fluid for lung cancer diagnosis. *Journal of Cancer*, 8(17):3585, 2017.

[121] Ning Zhang, Meng Wang, Peiwei Zhang, and Tao Huang. Classification of cancers based on copy number variation landscapes. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1860(11):2750–2755, 2016.

[122] Li Zhou, Hong Kiat Ng, Daniela I Drautz-Moses, Stephan C Schuster, Stephan Beck, Changhoon Kim, John Campbell Chambers, and Marie Loh. Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. *Scientific reports*, 9(1):1–16, 2019.