Decoding DNA Methylation: Insights into Age-Related Diseases and Transcription Factor Dynamics

by © Gastonguay Siu
Copy of

A thesis submitted to the School of Graduate Studies
Degree of Master of Science in Medicine
Human Genetics and Genomics in
the Division of BioMedical Sciences

Feb 2025*
St. John's Memorial University of Newfoundland

Acknowledgement

I would like to express my deepest gratitude to those who have supported and guided me throughout the completion of my master's thesis.

First and foremost, I would like to thank my wife, River Stone, for her unwavering love and support. Your encouragement and patience have been my foundation throughout this journey. I am also profoundly grateful to my mother, Bing Kin Siu, and my grandmother for their continuous support and belief in me. Their encouragement has been invaluable. I would like to extend my heartfelt thanks to my dog, Timer, whose companionship provided emotional support during challenging times. Your presence brought me comfort and joy.

In the realm of academia, I am immensely thankful for the guidance and support of my supervisor, Dr. Touati Benoukraf. Your insights and feedback have been instrumental in shaping this thesis. I am also grateful to Thomas J. Belbin, my manuscript co-author, for your collaboration and contributions. I would especially like to acknowledge that Chapter 4, regarding the YY1-TAF1 transcription factor, has been further developed into a manuscript currently in the pipeline for publication.

Special thanks to my academic colleagues, Elizabeth Trofimenkoff and Matthew Dyer, for their support and camaraderie. Our discussions and shared experiences have enriched my academic journey. I would also like to acknowledge Prof. Marc R. Roussel for introducing me to computational modelling, a key component of my research.

I am grateful to Dr. Karen Doody, Director of ARC-NL, for her support and the ARC-NL graduate fellowship awarded in September 2023. This financial support was crucial in enabling my research. Additionally, I am honoured to have received the Genetics Award for Best MSc

Presentation at the 2023 Biomedical Symposium organized by Memorial University of Newfoundland.

I would also like to thank my lab members for their collaborative spirit and assistance throughout my research. Your input and support have been vital. Finally, I acknowledge Memorial University of Newfoundland and the School of Graduate Studies for providing the resources and environment necessary for my research.

Thank you all for your unwavering support and encouragement. This thesis was only possible with each and every one of you.

Abstract

This thesis investigates the role of DNA methylation in age-related diseases by improving patient classification and exploring transcription factor interactions in the context of methylation. We developed an enhanced k-means clustering algorithm integrated with an expectation-maximization (EM) framework to filter irrelevant CpG sites and reduce noise in patient classification. The algorithm's performance is benchmarked against traditional k-means clustering to assess its efficacy in stratifying patients and gain meaningful insights. Our method improved patient stratification based on survival outcomes by applying to Kidney Renal Papillary Cell Carcinoma (KIRP) data. It revealed hypermethylated profiles resembling the CpG Island Methylator Phenotype (CIMP), which is associated with poor prognosis. Clinical features better represent the population, and tumour stages are more realistically linked to survival outcomes after grouping using the EM k-means algorithm. An analysis of the Fraction Genome Altered (FGA) and mutation rates suggests that post-cancer development survival rates appear more closely tied to large-scale genomic instabilities induced by aberrant methylation than point mutations. Gene Ontology (GO) and KEGG pathway analyses identified critical pathways in cell adhesion, signal transduction, and BMP signalling influencing tumour behaviour and metastasis. With slight modifications, the EM k-means algorithm was used to predict patient survival based on the methylation patterns. Although there was a large difference between the training and testing populations due to sampling variability, the results were promising, indicating a potential as a diagnostic method for treatment plans.

Extending our approach to Alzheimer's disease (AD), we uncovered genes associated with differentially methylated regions (GADMR) in AD patients not identifiable through traditional Braak staging or simple k-means clustering. The algorithm classified patients into

pseudo-intermediate and pseudo-advanced groups, with significant overlaps in known AD-associated genes and pathways. An age analysis demonstrated that machine learning-classified patients exhibited increased chronological and genetic ages (DNA methylation aberrations) correlated with AD progression risk.

Furthermore, we explored the molecular interplay between transcription factors Ying-Yang 1 (YY1) and TATA-Binding Factor 1 (TAF1). Employing TFregulomeR for motif distribution, genomic location, and Gene Ontology (GO) analysis on the YY1-TAF1 pair, we investigated methylation changes by comparing three scenarios: both TFs present together, and each TF alone by excluding the partner's peaks. From 152 conditions showing significant methylation variations, we focused on the YY1-TAF1 pair in GM12878, H1-hESC, and SK-N-SH cell lines. We found that TAF1 will not co-bind to YY1 when the YY1 binding motif is methylated at the third residue (cytosine or guanine) or when methylation is impossible due to the third residue not supporting it. Although the motif for co-binding peaks is cell-specific, stronger cytosine conservation at the third residue is observed where YY1-TAF1 co-binding occurs. Additionally, TAF1 binding to YY1 depends on an unmethylated state at this site. Our GO analysis reveals that the co-binding of YY1 and TAF1 expands the set of GO terms compared to YY1 alone, indicating a synergistic effect in regulating cellular processes. The extent of YY1-TAF1 co-binding at promoter-TSS sites varies by cell type, being more extensive in cells capable of differentiation. Specifically, co-binding in GM12878 cells correlates with functions related to protein synthesis and RNA processing, while in H1-hESC and SK-N-SH cells, it associates with a broader range of enrichments. Conversely, TAF1 alone shows the opposite pattern, suggesting that cellular needs and differentiation potential influence binding patterns.

In conclusion, this thesis presents a novel computational framework for improving patient classification based on DNA methylation patterns. It sheds light on the molecular mechanisms of transcription factors in the context of methylation specificity. The enhanced EM k-means algorithm demonstrates potential as a diagnostic tool for personalized treatment plans, while the insights into YY1 and TAF1 interactions contribute to understanding gene regulation in disease progression. These findings have significant implications for developing targeted therapies and highlight the importance of methylation dynamics in age-related disease mechanisms.

General Summary

This thesis explores the effects of DNA methylation, focusing on two main areas. The first approach uses machine learning to study age-related diseases like Kidney Renal Papillary Cell Carcinoma (KIRP) and Alzheimer's Disease (AD). The second examines how transcription factors YY1 and TAF1 interact.

Chapter 1 includes an introduction and literature review. Chapter 2 focuses on my research project that enhances KIRP prognosis prediction using advanced clustering techniques to analyze DNA methylation patterns, improving survival predictions. Chapter 3 adapts this approach to better classify AD patients by identifying age-related methylation changes. Chapter 4 investigates the interaction between YY1 and TAF1, showing that their co-binding leads to lower methylation levels and increased gene activity, highlighting TAF1's role in enhancing YY1 function.

Table of Contents

List of Abbreviations

5hmC: 5-Hydroxymethylcytosine

5mC: 5-Methylcytosine

AD: Alzheimer's Disease

AJCC: American Joint Committee on Cancer

ANNs: Artificial Neural Networks

ANOVA: Analysis of Variance

API: Application Programming Interface

ARE: Antioxidant Response Element

bp: Base Pair

BRCA: Breast Invasive Carcinoma

BS: Bisulphite

ccRCC: Clear Cell Renal Cell Carcinoma

CGI: CpG Island

CHG: Cytosine-Huanine-Guanine

CHH: Cytosine-Huanine-Huanine

ChIP-seq: Chromatin Immunoprecipitation Sequencing

CIMP: CpG Island Methylator Phenotype

Cl: Cluster

COAD: Colorectal Adenocarcinoma

CpG: Cytosine-phosphate-Guanine

DEG: Differentially Expressed Genes

DMR: Differentially Methylated Regions

DNAm: DNA Methylation

DR: Detection Rate

EM: Expectation Maximization

EOAD: Early-Onset Alzheimer's Disease

ESC: Embryonic Stem Cells

FDR: False Discovery Rate

FGA: Fraction of Genome Altered

FRET: Fluorescence Resonance Energy Transfer

GABA: Gamma-Aminobutyric Acid

GADMR: Genes Associated with Differentially Methylated Regions

GDC: Genomic Data Commons

GEO: Gene Expression Omnibus

GO: Gene Ontology

GSE: Gene Expression Series

GTRD: Gene Transcription Regulation Database

H3K27ac: Histone H3 Lysine 27 Acetylation

H3K27me3: Histone H3 Lysine 27 Trimethylation

H3K4me1: Histone H3 Lysine 4 Monomethylation

HEK293: Human Embryonic Kidney 293 Cells

HM450k: Illumina Infinium Human Methylation 450K

HMMs: Hidden Markov Models

IDAT: Illumina Data Array Table

Inr: Initiator Element

IQR: Interquartile Range

KEGG: Kyoto Encyclopedia of Genes and Genomes

KIRC: Kidney Renal Clear Cell Carcinoma

KIRP: Kidney Renal Papillary Cell Carcinoma

KM: Kaplan-Meier

LOAD: Late-Onset Alzheimer's Disease

LFC: Log Fold Change

LUAD: Lung Adenocarcinoma

LUSC: Lung Squamous Cell Carcinoma

lncRNA: Long Non-Coding RNA

MBD-seq: Methylated DNA Binding Domain Sequencing

MeDIP-seq: Methylated DNA Immunoprecipitation Sequencing

ML: Machine Learning

MRE: Mean Relative Error

mTOR: Mechanistic Target of Rapamycin pathway

NCI GDC: National Cancer Institute Genomic Data Commons

NCBI: National Center for Biotechnology Information

NFTs: Neurofibrillary Tangles

OR: Odds Ratio

OxBS: Oxidative Bisulphite

PAAD: Pancreatic Adenocarcinoma

PAMP: Pathogen-Associated Molecular Pattern

PI3K/AKT: Phosphatidylinositol 3-Kinase/Protein Kinase B Pathway

PSCs: Pluripotent Stem Cells

PWMs: Position Weight Matrices

RAAS: Renin-Angiotensin-Aldosterone System

RCC: Renal Cell Carcinoma

RNA-Seq: RNA Sequencing

RSEM: RNA-Seq by Expectation-Maximization

ROS: Reactive Oxygen Species

RSKC: Robust Sparse K-means Clustering

SAM: S-Adenosyl Methionine

SVMs: Support Vector Machines

TCGA: The Cancer Genome Atlas

TET: Ten-Eleven Translocation

TGF-β: Transforming Growth Factor Beta

TLR: Toll-Like Receptor

TPM: Transcripts Per Million

TSS: Transcription Start Site

WGBS: Whole-Genome Bisulphite Sequencing

WGCNA: Weighted Gene Co-expression Network Analysis

WSS: Within-Cluster Sum of Squares β

List of Figures

CHAPTER 1: Introduction

1.1 General Concepts

1.1.1 Overview Epigenetic Mechanisms

The incidence of age-related diseases, including cancer, cardiovascular diseases, and neurodegenerative disorders, increases with aging. Aberrant changes in the epigenome have been identified as key risk factors for these diseases. Epigenomics explores changes in gene expression and cellular phenotype that do not stem from alterations in the DNA sequence itself. According to the epigenetic progenitor model, age-related diseases and cellular dysfunctions originate from modifications in epigenetic marks rather than from genetic mutations [1], [2], [3]. As cells proliferate, progenitor cells harbouring accumulated epigenetic errors can lead to differentiated cells with dysfunctional gene expression patterns, thereby contributing to the onset of age-related diseases. Given that epigenetic changes are potentially reversible, unlike genetic mutations, they represent a promising focus for research. This thesis aims to elucidate methylation signatures that can be used as biomarkers in diagnostics and screenings of age-related diseases.

1.1.1.1 Overview of RNA-mediated modification & Histone modification

One form of epigenetic modification is mediated by RNA, which includes miRNAs and other non-coding RNAs. These miRNAs function primarily in the post-transcriptional regulation of gene expression. They bind to complementary sequences on mRNA resulting in gene silencing either through mRNA degradation or repression of translation. Long Non-Coding

RNAs (lncRNAs) can regulate gene expression via chromatin modification, transcription, and post-transcriptional processing. Other types of RNA may also induce epigenetic change through similar techniques.

Another significant type of epigenetic change involves histone modification [1], [2]. These modifications occur on the histone proteins around which DNA is wrapped, forming a structure known as nucleosomes. Changes to the histone proteins can influence the accessibility of DNA to transcription factors and other machinery necessary for gene expression, thereby altering the transcription rate of many genes. Epigenetic histone alteration can occur via histone acetylation, methylation, phosphorylation, and ubiquitination.

1.1.1.2. DNA Methylation

DNA methylation is one of the most well-known epigenomic modifications; aberrant DNA methylation refers to changes in the pattern of DNA methylation that deviate from what is typically seen in normal cells. As we age, we continue to accumulate more and more aberrant DNA methylations in a process called epigenetic drift. There is a discrepancy between age-related DNA methylation and chronological age, which may predict future disease and general mortality rates. Aberrant DNA methylation patterns have been implicated in various diseases, including cancer, cardiovascular disease, and neurological disorders like Alzheimer's [3], [4].

In the context of cancer, DNA methylation can be gained or lost at specific genomic locations, contributing to malignancy. For instance, DNA methylation gains at the promoter regions of tumour suppressor genes can lead to their inactivation [1]. In contrast, DNA methylation loss in some genes has been shown to contribute to tumour malignancy. Furthermore, metastasis, the process where a tumour invades and spreads to different tissues, is cancer's most lethal attribute, and initial findings indicate that metastasis initiation could be

epigenetically controlled. For example, metastasis initiation could be epigenetically controlled through the loss of DNA methylation in the SNAI1 gene or the inactivation of TRIM29, which results in reduced DNA methylation of TWIST.

DNA methylation is an epigenetic modification that involves adding a methyl group to the carbon-5 cytosine residue in DNA, forming 5-methylcytosine (5mC). The family of DNA methyltransferases (DNMTs) catalyze this process by transferring a methyl group from S-adenyl methionine (SAM) to cytosine. There are four members of the DNMT family: DNMT1, DNMT2, DNMT3A, and DNMT3B [5], [6]. DNMT1 is critical in maintaining DNA methylation patterns during cell division [7]. It uses hemi-methylated CpG DNA strands as a template to re-methylate cytosines in both strands. In contrast, DNMT3A and DNMT3B deposit de novo methylation patterns during development or in response to environmental stimuli and are known as de novo DNMTs [8]. DNMT2's function remains largely unknown, but some evidence suggests it plays a role in tRNA methylation during spermatogenesis [9]. DNA hydroxymethylation is a modification of DNA that involves adding a hydroxyl group to the 5-methylcytosine base, forming 5-hydroxymethylcytosine (5hmC) [10], [11] 5hmC is an intermediate of active demethylation catalyzed by a group of enzymes called ten-eleven translocation (TET) proteins.

.

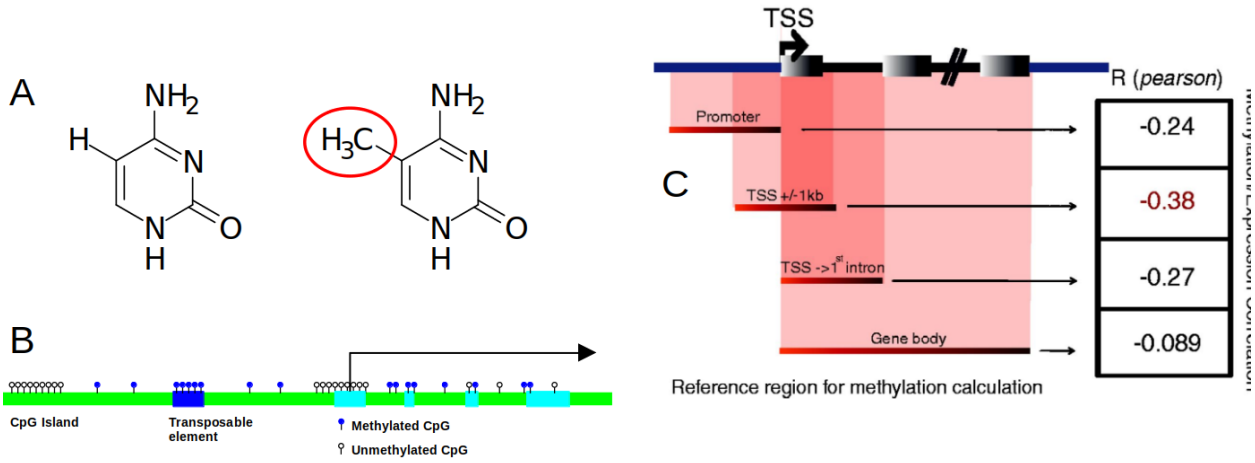


Figure 1. Overview of DNA Methylation: (A) Chemical structures of cytosine and 5-methylcytosine, illustrating the addition of a methyl group to the cytosine base. (B) Typical DNA methylation landscape in mammals, highlighting CpG islands, transposable elements, and the density of CpG sites represented as a lollipop plot (source:). (C) Breakdown of different DNA regions and their methylation-expression correlations, with Pearson (R) values shown for the promoter, transcription start site (TSS), and gene body regions (Permission for re-use through a Creative Commons CC-BY-NC license: T. Benoukraf, S. Wongphayak, L. H. A. Hadi, M. Wu, and R. Soong, "GBSA: a comprehensive software for analyzing whole genome bisulphite sequencing data," NUCLEIC ACIDS Res., vol. 41, no. 4, Feb. 2013, doi: 10.1093/nar/gks1281).

Often, the addition of a methyl group onto cytosine occurs next to guanine, forming what is commonly referred to as CpG dinucleotide [4]. CpG dinucleotide clusters, known as CpG islands, are sections of DNA that are at least 200 bp long and contain a CG concentration greater than 50%. Approximately 60% of mammalian promoters have CpG islands, most of which are unmethylated. DNA methylation at CpG islands can lead to gene silencing by preventing transcription factors and other regulatory proteins from binding to the DNA, resulting in decreased gene expression [12], [13]. DNA methylation also impacts splicing, nucleosome positioning, and transcription factor recruitment to DNA. Furthermore, DNA methylation can occur in non-CpG contexts, such as CHG and CHH (where H represents A, T, or C). Non-CpG methylation has been observed in a few cell types, such as neurons during brain tissue development and neuron differentiation, embryonic stem cells, and induced pluripotent stem cells. Additionally, X chromosome inactivation appears to be partly regulated by non-CpG cytosine methylation [14].

Differential methylation can occur at CpG island (CGI) shores. These shores refer to genomic regions immediately adjacent to CpG islands extending up to 2 kilobases (kb) away from the CpG island itself [3], [5], [15]. This change in methylation influences gene expression and can contribute to the development of various diseases. Similar to the methylation of CpG islands, a typically active gene may become silenced when heavily methylated at its CpG shore, potentially disrupting regular cellular function and fostering disease development. The effects of shore methylation are not limited to these proximal regions. Indeed, methylation at CpG shores situated in intragenic regions, within the gene but outside the protein-coding sequence, or in gene deserts, which are expansive genomic regions with sparse or no genes, could harbour additional regulatory roles that are not yet fully understood. Emerging research underscores the role of

differential methylation at CpG island shores in the onset of numerous diseases, including colon cancer and neurological disorders.

Differential methylation of CpG sites in relation to their nearest gene also helps elucidate genomic expression [16]. CGIs are often found near the TSS of genes, especially housekeeping genes that are consistently expressed in most cells. When CpG islands near the TSS are unmethylated, the associated gene is likely to be transcribed and, therefore, active. Conversely, when the CpG islands near the TSS are methylated, the gene is usually silenced because the methylation physically impedes the transcription machinery from accessing the gene, preventing its transcription [1], [5], [17]. Differential methylation within the gene body refers to changes in DNA methylation patterns within gene coding and non-coding regions. These changes can affect gene expression, which can play a role in various diseases, including cancer and neurological disorders [3], [5]. The impact of differential methylation on gene expression is complex. It depends on several factors, such as the location of the methylated cytosine base within the gene, the type of methylation pattern (e.g., hypermethylation or hypomethylation), and other epigenetic modifications. Overall, differential methylation within the gene body is an important area of research in understanding the role of epigenetics in disease development and progression.

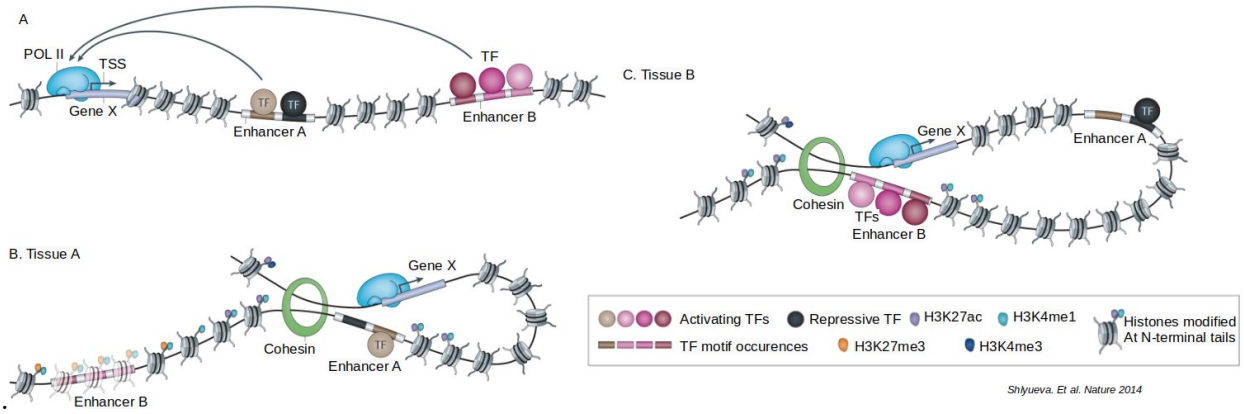
Enhancers are regions of DNA that are responsible for controlling gene expression. They are situated at variable distances from promoters and are key to regulating gene expression in development and cell function [16]. The methylation status of enhancers is closely connected to their function. In general, these regions have relatively variable methylation. Differential methylation of enhancer regions can lead to changes in gene expression, which can impact cellular processes and contribute to the development of diseases such as cancer. Researchers use DNA methylation arrays to identify enhancer regions and analyze the data using statistical

methods such as ANOVA or t-tests. The impact of differential methylation on enhancer regions is a topic of ongoing research. Still, this epigenetic modification plays a critical role in regulating gene expression and controlling cellular processes.

1.1.2 Transcription Factors

Variation of genetic expression is critical to adaptability to our cellular environment and differentiation in multicellular life. Transcription factors, proteins that bind to specific DNA domains known as motifs, play a key role in variability seen in genetic expression [18], [19]. They do so by modulating the rate at which genes are transcribed. They control cellular processes such as differentiation, development, and response to environmental cues. Transcription factors regulate gene expression by interacting with distinct DNA sequences termed regulatory elements, such as promoters, enhancers, and silencers, each having a specialized role in transcription.

Enhancers are distinct genomic regions containing binding sites for transcription factors that can upregulate the transcription of a target gene from its transcription start site (TSS) [20]. Despite their potential distances from target genes, active enhancers in specific tissues are bound by activating transcription factors and brought into close proximity to their target promoters through DNA looping, often mediated by cohesin and other protein complexes (Figure 2). This complex interaction enhances gene transcription and is marked by distinct biochemical features. For instance, active enhancers and promoters show nucleosome depletion and specific histone modifications such as histone H3 lysine 4 monomethylation (H3K4me1) and H3K27 acetylation (H3K27ac). Conversely, inactive enhancers might be characterized by repressive markers like the H3K27me3 mark, associated with Polycomb proteins, or by binding repressive transcription factors.



.

.

Figure 2. Enhancers are distinct genomic regions (or the DNA sequences thereof) that contain binding site sequences for transcription factors (TFs) and that can upregulate (that is, enhance) the transcription of a target gene from its transcription start site (TSS). A) Along the linear genomic DNA sequence, enhancers can be located at any distance from their target genes, which makes their identification challenging. B&C) In a given tissue, active enhancers (Enhancer A in part b or Enhancer B in part c) are bound by activating TFs. They are brought into proximity of their target promoters by looping, which is thought to be mediated by cohesin and other protein complexes. Additionally, gene regulatory elements are distinguished by their nucleosome presence or absence and specific histone modifications: active promoters and enhancers exhibit nucleosome depletion and modifications like histone H3 lysine 4 monomethylation (H3K4me1) and H3K27 acetylation (H3K27ac). Inactive enhancers might be silenced by different mechanisms, such as the Polycomb protein-associated repressive H3K27me3 mark (part b) or repressive TF binding (part c). Permission for re-use through a SPRINGER NATURE LICENSE copyright: D. Shlyueva, G. Stampfel, and A. Stark, “Transcriptional enhancers: from properties to genome-wide predictions,” Nat. Rev. Genet., vol. 15, no. 4, pp. 272–286, Apr. 2014, doi: 10.1038/nrg3682).

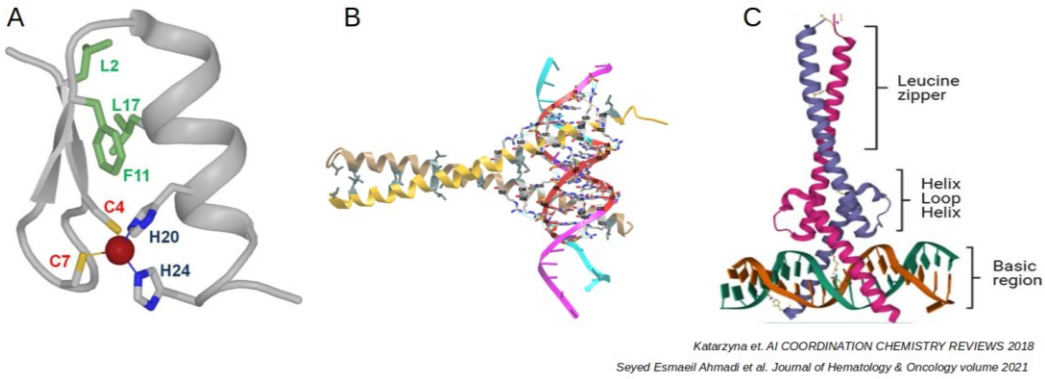
The additive action of different enhancers, each with cell-type- or tissue-specific activities, results in complex gene expression patterns. Transcription factors (TFs) act as molecular switches, which can either activate or repress the recruitment of RNA polymerase to specific genes, thereby controlling the gene transcription rate. [18], [19]. This "on-off" mechanism is not binary; it fine-tunes the gene expression level in response to cellular needs. The interplay between transcription factors adds to their complexity. TFs usually function within complex networks, where multiple TFs work together temporarily and chronologically. Co-binding of TFs can activate or silence domains, which would otherwise lead to the transcription factor acting differently. The combinatorial interactions of different transcription factors can lead to diverse transcriptional outcomes, allowing cells to respond with precision to various signals and cellular stresses. This complexity is further enhanced by post-translational modifications that can modulate transcription factors' activity, stability, or DNA-binding affinity [21].

Promoters, enhancers, and silencers are essential regulatory elements that transcription factors engage with [18]. Promoters are located immediately upstream of the gene and serve as the primary docking site for the transcriptional machinery. On the other hand, enhancers can be found thousands of base pairs away from the genes they influence. They boost the transcriptional activity of a gene by looping the DNA so that the enhancer comes into close proximity with its target gene’s promoter. Silencers are the antithesis of enhancers; they downregulate or inhibit transcription. Like enhancers, they can also be located considerably from the genes they influence.

There are three main structures in transcription factors that facilitate their binding onto DNA (Figure 3). The zinc finger is a dominant eukaryotic DNA-binding motif [22], [23], [24],

[25], [26], [26]. Discovered in transcription factor IIIA from the African clawed toad, this motif involves compact, small globular domains brought together by Zn^{2+} ions, facilitating DNA binding. The Zif268 protein from mice is an example that uses this motif, employing three zinc fingers arranged in a structure aptly complementing the DNA's major groove. Each finger forms highly specific hydrogen bonds with the DNA, leading to a modular system proficiently recognizing extended asymmetric base sequences.

.



.

.

Figure 3. Key Structural Motifs in Eukaryotic Transcription Factors A) illustrates the zinc finger motif, which uses zinc ions to stabilize its structure for DNA interaction. B) displays the leucine zipper motif. C) depicts the basic helix-loop-helix (bHLH) motif combined with a leucine zipper. (Permission for re-use through a Creative Commons CC-BY-NC license [26] and [27]: K. Kluska, J. Adamczyk, and A. Krezel, "Metal binding properties, stability and reactivity of zinc fingers," Coord. Chem. Rev., vol. 367, pp. 18–64, Jul. 2018, doi: 10.1016/j.ccr.2018.04.009; S. E. Ahmadi, S. Rahimi, B. Zarandi, R. Chegeni, and M. Safa, "MYC: a multipurpose oncogene with prognostic and therapeutic implications in blood malignancies," J. Hematol. Oncol.J Hematol Oncol, vol. 14, no. 1, Aug. 2021, doi: 10.1186/s13045-021-01111-4).

Another eukaryotic transcription factor feature is the leucine zipper. Heptad repeats (seven-residue pseudo repeating sequences) within these factors, particularly leucine at every seventh position, facilitate dimerization, creating a structure resembling a zipper [22], [24], [25], [28], [29]. The leucine zipper doesn't bind DNA but enables the DNA-binding motif's proper orientation. Many of these proteins encompass a DNA-binding region rich in basic residues right before the leucine zipper, leading to their classification as basic region leucine zipper (bZIP) proteins. One representative example is GCN4, where the protein's helices dimerize through the leucine zipper and then diverge to interact with DNA's significant grooves.

The basic helix-loop-helix (v) is another motif in eukaryotic transcription factors. Often followed by a leucine zipper, the bHLH motif's basic region, combined with the first helix, binds to the DNA major groove [22], [24], [25], [29]. Meanwhile, its second helix aids protein dimerization via coiled-coil formation. As visualized in the Max protein, the bHLH/Z dimer grips the DNA akin to forceps, where each basic region contacts specific DNA bases and phosphate groups, culminating in precise gene regulation.

Transcription factors are instrumental in cellular differentiation, the process by which unspecialized cells mature into specialized forms. For example, the presence or absence of specific TFs can trigger stem cells to differentiate into muscle cells, neurons, or other cell types. Pluripotency is a transient property of cells within the early embryo. This property can be captured in vitro as pluripotent stem cells (PSCs) at different developmental periods [9]. These cells initially appear in the blastocyst-stage embryo's inner cell mass and can be maintained later by reprogramming germ cells, resulting in embryonic germ cells similar to ESCs. This differentiation process is crucial for multicellular organism development and tissue repair [31], [32].

Transcription factors regulate gene expression for tissue and organ growth, ensuring timely cell proliferation and specialization [18], [30]. Mutations in transcription factors or their regulatory elements can result in diseases such as cancer, auto-immune diseases, diabetes, and mental disorders such as retinopathies [33]. For instance, the transcription factor SOX9 is critical in developing and differentiating chondrocytes, which are cells found in cartilage and have implications in disorders such as campomelic dysplasia, a skeletal malformation syndrome [34]. Transcription factors like SOX9 regulate genes responsible for the growth and development of tissues and organs, ensuring that cells proliferate and specialize at the appropriate times and locations.

1.1.3 Techniques for Measuring Genome-wide DNA Methylation Profiles

Genome-wide DNA methylation profiling has become an essential tool for understanding the role of epigenetic regulation in various biological processes and diseases. Several techniques have been developed to measure DNA methylation profiles on a genome-wide scale, each with advantages and limitations. Some more common methods include whole-genome bisulphite sequencing (WGBS), Illumina Infinium Methylation arrays, methylated DNA immunoprecipitation sequencing (MeDIP-seq), and Methylated DNA binding domain sequencing (MBD-seq).

WGBS is considered the gold standard for DNA methylation profiling [35]. This technique treats genomic DNA with sodium bisulphite, which converts unmethylated cytosines to uracil while methylated cytosines remain unchanged. Subsequently, the bisulphite-converted DNA is sequenced using high-throughput sequencing platforms. Bioinformatic analysis is employed to align the sequenced reads to a reference genome and determine the methylation status of individual cytosines by comparing the differences between the bisulphite-converted and

unconverted sequences. WGBS provides single-base resolution and covers the entire genome, but it is relatively expensive and requires a large amount of input DNA.

Illumina Infinium Methylation arrays have a high throughput, are cost-effective, and require less input DNA than bisulphite sequencing-based methods [36]. The Infinium Methylation arrays, such as the 450K and EPIC arrays, consist of thousands of CpG probes designed to interrogate specific CpG sites across the genome. Genomic DNA is bisulphite-converted, and the methylation status of the targeted CpG sites is determined by hybridization to the array probes, followed by single-base extension and fluorescence detection. These arrays provide information on many pre-selected CpG sites but do not offer single-base resolution or cover the entire genome.

MeDIP-seq is an antibody-based method that involves immunoprecipitation of methylated DNA fragments using an anti-5-methylcytosine antibody [37]. The enriched methylated DNA is then sequenced using high-throughput sequencing platforms. MeDIP-seq provides information on genome-wide methylation patterns but lacks single-base resolution and may have reduced sensitivity in regions with low CpG density.

Similar to MeDIP-seq, MBD-seq uses a protein domain (such as the MBD domain of MeCP2) with a high affinity for methylated DNA to enrich methylated DNA fragments [38]. The captured DNA is then sequenced to generate genome-wide methylation profiles. MBD-seq also lacks single-base resolution and may have reduced sensitivity in regions with low CpG density.

1.1.4 Tools and Techniques for Mapping Transcription Factor Binding Sites

Chromatin Immunoprecipitation followed by sequencing, commonly known as ChIP-seq, stands as a cornerstone technique in the modern realm of genomics, granting researchers the ability to map protein-DNA interactions on a genome-wide scale [39], [40], [41]. The essence of

ChIP-seq revolves around identifying where specific proteins, such as transcription factors or modified histones, bind to DNA. The procedure initiates with cross-linking, where DNA and its associated proteins are fixed using formaldehyde, ensuring their interactions are preserved. Post-cross-linking, cell lysis breaks open the cells, followed by sonication to fragment the DNA into smaller, manageable pieces. Immunoprecipitation is then employed, where specific antibodies selectively pull down the protein of interest, capturing the DNA it's bound to. Once this is achieved, the cross-links are reversed, reverting the protein-DNA complex to its native state, and the DNA is subsequently isolated. The culminating step involves sequencing these isolated DNA fragments, which illuminates the precise regions of the genome where the protein binds.

The vast potential of ChIP-seq is reflected in its multifaceted applications [39], [40], [41]. It is instrumental in identifying transcription factor binding sites, which are crucial for understanding how genes are turned on or off. Additionally, it facilitates the mapping of genome-wide histone modifications, offering insights into the epigenetic mechanisms that regulate gene expression. Beyond this, ChIP-seq plays a pivotal role in deciphering the intricate regulatory networks that govern gene expression, shedding light on the dynamic interplay between proteins and the genome. Post-sequencing, the obtained DNA fragments are analyzed against a reference genome, elucidating regions instrumental in regulating gene expression, pinpointing active enhancers, and identifying other functional DNA elements, thereby deepening our understanding of the genomic landscape and its regulatory intricacies.

TFregulomeR is a comprehensive database that integrates position weight matrices (PWMs) of transcription factor (TF) motifs from the MethMotif and Gene Transcription Regulation Database (GTRD) [42], [43]. MethMotif provides TF motifs enriched with methylation data, generated using MACS2 for ChIP-seq peak calling and MEME-ChIP for motif

analysis [41], [44]. DNA methylation profiles are also inferred using whole-genome bisulphite sequencing (WGBS). TFregulomeR incorporates data from GTRD, including ChIP-seq peaks and motifs centred within a 200 bp range of those peaks. Each PWM in TFregulomeR is given a unique ID and thoroughly annotated with details such as source, species, cell line, and other contextual factors.

1.1.5 Introduction to Machine Learning

1.1.5.1 fundamental concept of machine learning

Computational biology faces the challenge of extracting useful information from biological data [45]. Developing tools and methods to convert data into biological knowledge is essential. Machine learning, a powerful tool for analyzing complex biological datasets in bioinformatics, trains algorithms to make predictions or classifications without explicit programming. It consists of two phases: (i) estimating unknown dependencies from a dataset and (ii) using these dependencies to predict new system outputs [45]. Machine learning can be categorized into supervised and unsupervised algorithms useful for bioinformatics analysis.

Supervised learning trains algorithms on labelled datasets, with correct outputs or targets assigned. It learns the input-output relationship by finding patterns in labelled examples. Unsupervised learning uses unlabeled datasets, identifying patterns or structures independently. Two methods to solve biological problems using machine learning optimization principles are exact and approximate. Exact methods provide optimal solutions in specific cases, while approximate methods consistently offer solutions, although not always optimal. Machine learning models favour approximate methods with computational constraints, such as estimation maximization [46].

The curse of dimensionality refers to the difficulties in handling high-dimensional datasets. Data sparsity and computational complexity increase as dimensions increase, impacting algorithm effectiveness, training times, and performance. Visualizing and understanding data in high-dimensional spaces is challenging, rendering traditional data analysis techniques less effective.

Machine learning aims to create models for classification, prediction, and estimation [47]. Classification, the most common task, sorts data into pre-defined categories. Training errors occur during training, while generalization errors occur with new data. A good classification model should fit the training set and perform well on unseen data. Overfitting happens when a model fits training data too well but underperforms with new data. The ideal model complexity minimizes generalization error. The bias-variance decomposition analyzes the expected generalization error, summing the bias and variance components for a classification model's overall expected error.

1.1.5.2 Artificial Neural Networks

Artificial neural networks (ANNs) are one of the most well-known forms of machine learning algorithms and are often misused synonymously in media for machine learning. ANNs can handle various classification or pattern recognition problems by generating output as a combination of the input variables [48]. The process usually involves multiple hidden layers that represent the neural connections mathematically. Although ANNs are considered a gold standard method in several classification tasks, they have some drawbacks. ANNs can be computationally expensive and time-consuming to train, particularly for larger and more complex models in genomic analysis. Additionally, understanding how the model makes its predictions in genomic

analysis is challenging, if not impossible. Finally, ANNs are susceptible to overfitting, where the model is too closely tailored to the training data and performs poorly on new, unseen data.

1.1.5.3 Support Vector Machines

Support Vector Machines (SVMs) are supervised learning algorithms used in genomics for classification and regression analysis [45], [49]. SVMs work by finding the hyperplane that maximizes the margin between the two classes of data points or fits the data points as closely as possible in the case of regression analysis. In genomics, SVMs are particularly useful for analyzing high-dimensional data, such as gene expression. The input features are mapped into a higher-dimensional space using a kernel function, which can effectively separate the classes of data points. Different kernel functions have different properties and are suitable for different types of genomic data. However, the performance of SVMs in genomics depends on the choice of the kernel function and the value of hyperparameters. Additionally, SVMs can be computationally expensive when dealing with large genomic datasets. They may need to improve when the data is imbalanced or the classes overlap, a common issue in genomic data analysis.

1.1.5.4 k-means clustering

In high-throughput genomics research, the curse of dimensionality can make it challenging to identify meaningful patterns in datasets with more variables than observations. Traditional statistical techniques may not be sufficient to address this challenge. To overcome this, some researchers have developed methods such as feature selection to reduce the data's dimensionality while preserving important features and relationships between variables [46], [49], [50]. Unsupervised machine learning algorithms such as k-means clustering have also been

employed to partition the data into clusters with similar characteristics. K-means clustering is an unsupervised machine learning algorithm that aims to partition a given dataset into k clusters where each data point belongs to the cluster with the nearest mean. The algorithm randomly initializes k centroids and iterates through two steps until convergence. In the first step, each data point is assigned to the cluster with the nearest centroid. In the second step, the centroids are recalculated as the mean of all the data points in the cluster. The two steps are repeated until the assignments no longer change or a maximum number of iterations is reached. Applying k-means clustering in genomics aims to increase the effectiveness of analyzing high-dimensional datasets with few observations.

1.1.5.5 Estimation Maximization

Estimation Maximization (EM) is another popular iterative algorithm in machine learning used to maximize a posteriori (MAP) estimates of parameters in statistical models. EM is beneficial in epigenomic analysis, where some variables or parameters may be unobserved or hidden. EM alternates between two key steps: estimating the missing or hidden variables and updating the model parameters based on these estimates. The estimation of missing data is carried out in the E-step, where the expected value of the log-likelihood function is calculated using the current estimates of the model parameters. This provides an estimate of the missing or hidden data, which is then used to update the model parameters in the M-step. The process of alternating between the E and M-steps continues until the estimates of the model parameters converge to a local maximum or minimum. Carefully choosing the initial estimate of the model parameters is important, and regularization techniques should be used to avoid overfitting. In summary, EM is a powerful technique for identifying the parameters of complex statistical models, particularly when some variables are unobserved or hidden.

1.1.5.6 Technique Overview and Limitation in Machine Learning

Machine learning (ML) utilizes statistical theory and mathematical principles to create computational models that efficiently process data and make accurate inferences [45], [51]. Algorithm efficiency is as crucial as predictive accuracy due to computational constraints. Preprocessing techniques, such as dimensionality reduction, feature selection, and feature extraction, can enhance data analysis by modifying raw data. Dimensionality reduction, particularly for datasets with numerous features, can eliminate irrelevant features, reduce noise, and improve learning model accuracy. Machine learning can aid in extracting biological knowledge from complex datasets through supervised and unsupervised algorithms, useful for classification, prediction, and estimation. Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) are popular supervised algorithms used in genomics for classification and regression analysis. However, both are computationally expensive, and ANNs are prone to overfitting. Employing preprocessing techniques can boost the accuracy and efficiency of learning models.

1.1.6 Background on Kidney Renal Papillary Cell Carcinoma (KIRP)

1.1.6.1 Overview of Cancer

Cancer arises from the uncontrolled growth and spread of abnormal cells within the body, originating from changes or mutations in a cell's genetic material [52]. These mutations can occur spontaneously or be induced by external factors such as radiation, chemicals, or viruses. Genetic changes disrupt normal cell growth and division processes, allowing cells to proliferate uncontrollably and form a tissue mass called a tumour. Tumours can be classified as benign or malignant. Benign tumours are non-cancerous and do not invade neighbouring tissues or spread

to distant organs. In contrast, malignant tumours are cancerous, having the ability to invade surrounding tissues and spread to other parts of the body through a process called metastasis. This invasive nature of malignant tumours is a key factor that makes cancer a life-threatening disease.

In cancer biology, tumours are generally defined as enlarged tissue swellings caused by excessive epithelial cell growth. Although "tumour" and "cancer" are often used synonymously and will be used synonymously in this paper, it is crucial to distinguish between benign and malignant tumours [52]. Benign tumours are encapsulated by fibrous tissue and do not invade neighbouring tissues, while malignant tumours lack structure, infiltrate nearby tissues, and are classified as cancer. Tumour metastasis can interfere with organ function and result in patient death, making early detection and treatment critical for improved outcomes.

Cancer is a complex and diverse disease that arises from abnormal cell growth and behaviour. While there are various types of cancer, they all share several common features that differentiate them from normal cells and provide insight into their underlying mechanisms. Hanahan and Weinberg (2000) identified six key cancer features outlined below and illustrated in figure 4 [53]:

.

.

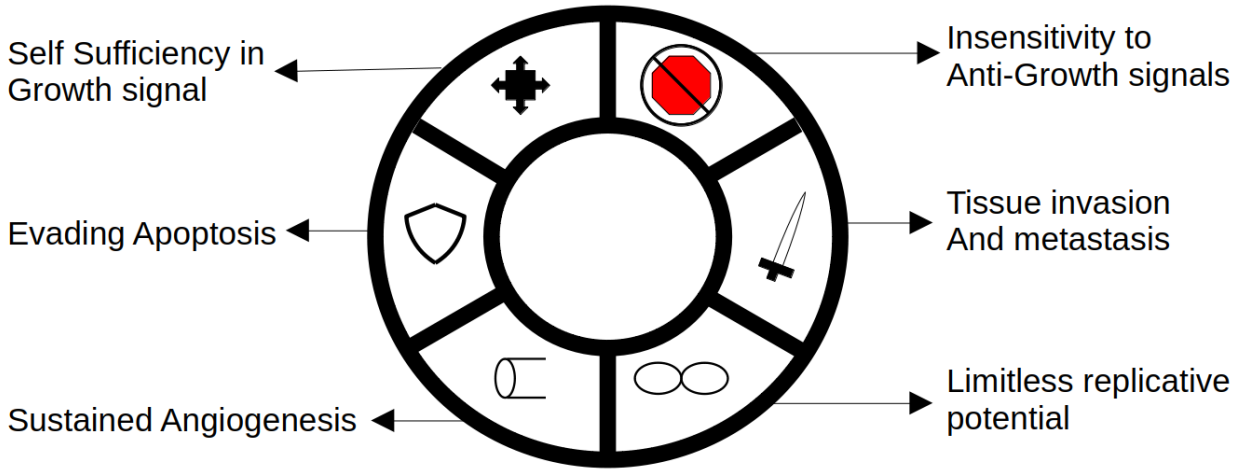


Figure 4. The six acquired capabilities of cancer: insensitivity to anti-growth signals, tissue invasion and metastasis, limitless replicative potential, sustained angiogenesis, evasion of apoptosis, self-sufficiency in growth signals,

1. Self-sufficiency in growth signals: Normal cells require external growth signals to divide. However, cancer cells can produce their growth signals or become less dependent on them, promoting continuous cell division. Various factors, such as mutations in oncogenes or loss of tumour suppressor genes, can drive this feature.
2. Insensitivity to anti-growth signals: Cancer cells can become resistant to signals that usually inhibit cell growth or initiate cell death, allowing them to continue growing unchecked. This resistance can be due to genetic mutations or epigenetic changes that affect the expression of key genes involved in cell cycle regulation.
3. Evasion of apoptosis: Apoptosis is the process of programmed cell death that eliminates damaged or unwanted cells. Cancer cells often develop mechanisms to avoid apoptosis, such as overexpression of anti-apoptotic proteins or downregulation of pro-apoptotic proteins, enabling them to survive and multiply.
4. Limitless replicative potential: Normal cells can divide a finite number of times before they enter a state called senescence. Cancer cells can bypass this limit by activating telomerase or other telomere maintenance mechanisms, allowing them to replicate indefinitely.
5. Sustained angiogenesis: Angiogenesis is the formation of new blood vessels. Cancer cells can stimulate angiogenesis to ensure a continuous supply of nutrients and oxygen, supporting their growth and survival. This feature is crucial for tumour growth and progression.
6. Tissue invasion and metastasis: Malignant tumours lose adhesion capabilities and gain the ability to invade surrounding tissues and spread to distant organs through blood or lymphatic vessels. This process, called metastasis, is a key characteristic of cancer and a

major contributor to its severity. The mechanisms underlying metastasis are complex and involve numerous factors, including the expression of specific genes and interactions between cancer cells and the surrounding microenvironment.

1.1.6.2 Fundamental Biological Understanding of Kidney Function

The kidneys are a pair of vital, bean-shaped organs in the lower back region on either side of the spine [19], [54]. They play a critical role in maintaining the body's homeostasis by filtering waste products, excess nutrients, and fluids from the blood to produce urine, which is then eliminated from the body. The kidneys carry out several essential functions to ensure overall health, and their complex and intricate mechanisms involve numerous physiological processes:

- Blood volume and pressure regulation: Kidneys control fluid excretion and produce hormones like renin, involved in the renin-angiotensin-aldosterone system (RAAS), regulating blood pressure and blood vessel constriction.
- Electrolyte balance: The kidneys maintain electrolyte balance (sodium, potassium, calcium, phosphate) by regulating ion reabsorption and secretion in nephrons, their functional units. The renal tubular epithelium, cells lining renal tubules within nephrons, actively transport ions between blood and forming urine. Electrolytes are vital for physiological functions, including muscle contractions, nerve impulses, and maintaining the body's acid-base balance.
- Acid-base balance: The kidneys help maintain the body's acid-base balance by excreting hydrogen ions and reabsorbing bicarbonate ions through complex processes in the renal

tubular epithelium. This regulation helps maintain the blood's pH levels within a narrow range, essential for proper cellular function, enzyme activity, and overall homeostasis.

- Waste excretion: The kidneys excrete metabolic waste products (urea, creatinine, uric acid) by filtering them through glomeruli in nephrons, then combining them with excess water and substances in renal tubules to form urine. The renal tubular epithelium actively transports waste products from the blood into the urine, preventing toxic substance accumulation in the body. This waste removal process is essential for maintaining overall health.
- Hormone production: Kidneys produce hormones like erythropoietin (EPO), stimulating red blood cell production, and calcitriol (active vitamin D) for calcium absorption and bone health.
- Gluconeogenesis: The kidneys are involved in gluconeogenesis, which generates glucose from non-carbohydrate sources such as amino acids, lactate, and glycerol. This function is particularly important during fasting or prolonged exercise when the body requires additional glucose to maintain normal blood sugar levels.

The renal tubular epithelium's location within the nephrons and its involvement in various kidney functions emphasize its crucial role in maintaining the body's overall health and homeostasis. The renal tubule is a long, convoluted structure emerging from the glomerulus and is functionally divided into three parts. The first part, the proximal convoluted tubule (PCT), remains in the renal cortex near the glomerulus. The second part, the loop of Henle or the nephritic loop, forms a descending and ascending limb that passes through the renal medulla. The third part, the distal convoluted tubule (DCT), is also in the renal cortex. The DCT, the final

segment of the nephron, connects and empties its contents into collecting ducts lining the medullary pyramids. These collecting ducts gather contents from multiple nephrons and fuse as they enter the papillae of the renal medulla. Figure 5 illustrates a more detailed look at the nephron of the kidney. Any impairment in kidney function, including disruptions in the renal tubular epithelium, can lead to imbalances and complications that affect various physiological systems, such as the cardiovascular, nervous, and musculoskeletal systems. For example, electrolyte imbalances can cause irregular heartbeats, muscle weakness, and neurological issues. Impaired waste excretion may lead to toxin accumulation, causing fatigue, nausea, and kidney failure. Additionally, hormone production disruptions can affect red blood cell production and bone health. Therefore, maintaining renal tubular epithelium function is essential for overall well-being and preventing severe health complications.

.

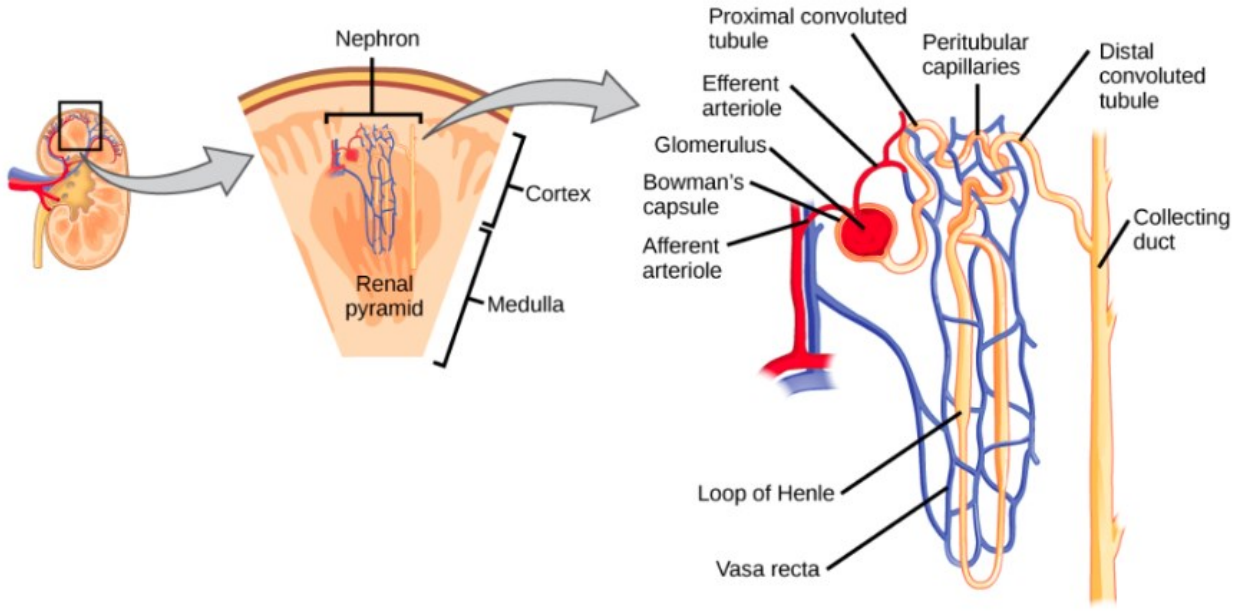


Figure 5. The nephron is the functional unit of the kidney. The glomerulus and convoluted tubules are located in the kidney cortex, while collecting ducts are located in the pyramids of the medulla (Permission for re-use through a Creative Commons CC-BY-NC license [55]: C. Rye, R. Wise, V. Jurukovski, J. DeSaix, J. Choi, and Y. Avissar, Biology. Houston, Texas: OpenStax, 2016.)

1.1.6.3 Overview of kidney renal papillary cell carcinoma and kidney cancer

Kidney Renal Papillary Cell Carcinoma (KIRP) is a subtype of renal cell carcinoma (RCC) originating from a renal tubular epithelium mutation, accounting for most primary kidney cancers. RCC development is linked to the disruption of metabolic pathways, with clear cell RCC being associated with genetic changes such as the loss of von Hippel Lindau's (VHL) gene function [56]. When the VHL gene is lost, silenced, or mutated, it can lead to the abnormal stabilization of hypoxia-inducible factors (HIFs). Many target genes of HIF-α encode proteins that contribute to renal carcinogenesis. Apart from the VHL-HIF pathway, RCC is marked by alterations in several other signalling pathways, including mTOR, PI3K/AKT, and Wnt pathways. These pathways regulate cell cycle control and apoptosis, thereby driving tumour growth and progression [57], [58], [59]. According to the epigenetic progenitor model, epigenetic changes could be powerful substitutes for genetic mutations.

Other forms of kidney cancer include transitional cell carcinoma (TCC), which develops in the renal pelvis, the part of the kidney that collects urine and funnels it to the ureter. It originates in the urothelium, the lining of the urinary tract. There is also Wilms tumour, also known as nephroblastoma, which typically occurs in the renal parenchyma, the kidney's functional tissue, most commonly in children. Nephroblastoma is associated with abnormal kidney development. Renal sarcoma arises in the kidney's connective tissue or blood vessels and is often more aggressive than other tumours. Other forms of kidney cancers include oncocytoma, angiomyolipoma, medullary carcinoma, and cystic nephroma.

KIRP, a subtype of RCC, begins in the cells lining the small tubes in the kidney, known as the renal papillae. These renal papillae are small structures that extend into the kidney's collecting system and facilitate urine transport from the kidney to the bladder [60]. KIRP

accounts for 15-20% of all kidney cancers and is more common in men than women. It typically affects people over 55, and some risk factors include smoking, obesity, high blood pressure, and a family history of kidney cancer. Compared to other RCC subtypes, KIRP is less invasive and generally has a more favourable prognosis [61].

DNA methylation, an epigenetic modification, plays a significant role in the development and progression of KIRP. Aberrant methylation patterns have been found in KIRP tumours, leading to changes in gene expression that contribute to tumour development and progression [62], [63]. These changes may be potential biomarkers for diagnosis, prognosis, and treatment response prediction. One study, which includes analysis of the TCGA KIRP data, emphasizes the integration of various molecular platforms for improved molecular classification of different cancer types [64].

Understanding the various signalling pathways and molecular mechanisms behind the development and progression of RCC subtypes, including KIRP, is crucial for advancing diagnosis, prognosis, and treatment. Further research is required to comprehend the mechanisms underlying these epigenetic changes and develop novel diagnostic and therapeutic strategies targeting DNA methylation in KIRP. Using a machine learning approach, researchers can uncover novel patterns and relationships between various molecular factors, KIRP and other age-related diseases, which might have remained hidden using traditional analytical methods. Moreover, the algorithm's ability to analyze multivariate data enables it to identify complex interactions between different pathways, thereby creating a valuable understanding of the disease's underlying mechanisms. The insights gained from this advanced analysis could pave the way for developing more effective diagnostic tools, personalized treatments, and preventive

strategies, ultimately improving patient outcomes and quality of life for those affected by age-related diseases.

1.1.7 Background on Alzheimer's Disease

1.1.7.1 Fundamentals of Neurobiology

The nervous system is a highly intricate and interconnected network that relies on specialized cells known as neurons to transmit and process information. These basic functional units of the nervous system are responsible for sending electrical/chemical signals through their distinct parts, including the soma or cell body, dendrites, axons, and synapses [65], [66]. The soma contains the nucleus and other essential cellular components, while dendrites are branch-like structures that receive signals from other neurons. The axon is a long projection that transmits electrical signals, and synapses are junctions between neurons where information exchange occurs.

Synaptic transmission is the primary method of communication between neurons, occurring when an action potential (signal) travels down the presynaptic neuron's axon. This process triggers the release of neurotransmitters, which are chemical messengers, into the synaptic cleft [65], [66]. These molecules then diffuse across the cleft and bind to receptors on the postsynaptic neuron. Depending on the neurotransmitter and receptor type, this binding can either excite or inhibit the postsynaptic neuron, generating a new action potential and facilitating the flow of information.

Neurotransmitters are pivotal in transmitting signals across synapses [65], [66]. Some common neurotransmitters include glutamate, which is generally excitatory; GABA, which is inhibitory; dopamine, which modulates reward and motivation; serotonin, which regulates mood;

and acetylcholine, which is involved in learning and memory. Receptors are specialized proteins found on the surface of neurons that bind to neurotransmitters. The binding of a neurotransmitter to its receptor can trigger a variety of cellular responses, such as depolarization in the postsynaptic neuron, which creates an action potential or activates intracellular signalling pathways.

The brain organizes neurons into intricate circuits and networks, which process and integrate information from various sources to support multiple functions, including sensory perception, motor control, memory, learning, and decision-making. These complex networks comprise diverse neuronal populations with specific properties and connectivity patterns. The coordinated activity among interconnected neurons within these circuits enables the brain to process and respond to information with remarkable speed and efficiency.

Glial cells are non-neuronal cells in the nervous system that play crucial roles in supporting, protecting, and nourishing neurons [65], [66]. There are three main types of glial cells: astrocytes, oligodendrocytes, and microglia. Astrocytes help maintain the extracellular environment, regulate neurotransmitter levels, and contribute to the blood-brain barrier. Oligodendrocytes produce the myelin sheath that insulates axons, enabling faster signal transmission. Microglia are the brain's primary immune cells, contributing to inflammation, clearing cellular debris, and responding to injury or infection.

Neuroplasticity refers to the brain's remarkable ability to adapt and change its structure and function in response to experiences, learning, or injury [65]. This dynamic process involves various cellular and molecular mechanisms, including the formation and strengthening of synaptic connections through long-term potentiation, the generation of new neurons (neurogenesis), primarily in the hippocampus, and the pruning of unused or weak connections.

Neuroplasticity is crucial for learning, memory, cognitive flexibility, and recovery from brain injury.

Memory and learning are interconnected cognitive processes that involve encoding, storing, and retrieving information. These processes rely on changes in synaptic connections, a phenomenon known as synaptic plasticity [65], [66]. The hippocampus and surrounding medial temporal lobe structures play a crucial role in forming and consolidating long-term memories. Other brain regions, such as the prefrontal cortex and amygdala, contribute to memory's emotional and contextual aspects.

In conclusion, the fundamentals of neurobiology encompass the structure and function of neurons, synaptic transmission, neurotransmitters and their receptors, neuronal circuits and networks, glial cell function, neuroplasticity, and the processes of learning and memory. Understanding these principles is essential for studying age-related cognitive decline and neurodegenerative disorders, such as Alzheimer's, and developing novel therapeutic strategies to address these complex and multifaceted conditions.

1.1.7.2 Overview of Alzheimer's Disease

Alzheimer's disease (AD) is a progressive and irreversible neurodegenerative disorder affecting the brain. It is the most common cause of dementia, accounting for 60-80% of all cases [67], [68]. AD is classified into two distinct categories: early-onset AD (EOAD), which accounts for about 5% of the AD population, and late-onset AD (LOAD), which accounts for the other 95% of AD patients. EOAD is a Mendelian pattern disease, whereas LOAD is genetically complex and associated with several genes. The apolipoprotein E (APOE) gene is LOAD's most important genetic risk factor. Although there is currently no known cure for Alzheimer's disease, treatments are available that can help manage the symptoms and slow the progression of the

disease [69]. The exact causes of Alzheimer's disease remain a topic of research; it is believed to be caused by a combination of genetic, environmental, and lifestyle factors. A family history of the disease, age, and specific genetic mutations are known risk factors for developing the disease. Other potential risk factors include head injuries, high blood pressure, and high cholesterol levels.

The symptoms of Alzheimer's disease typically begin with mild memory loss, difficulty with language, and problems with decision-making and problem-solving. As the disease progresses, these symptoms worsen, and other symptoms may develop, such as changes in mood and behaviour, difficulty with basic activities of daily living, and eventually, complete loss of communication and mobility. In the later stages of the disease, patients may require round-the-clock care.

Alzheimer's is a complex disorder with multiple biological and molecular mechanisms involved in its development. Amyloid plaques and neurofibrillary tangles within the brain most often characterize it [70], [71]. Amyloid plaques consist of beta-amyloid (Aβ42/40) aggregates derived from amyloid precursor protein (APP) by β-secretase 1 (BACE-1) and γ-secretase cleavage. However, in Alzheimer's disease, beta-amyloid proteins accumulate in the spaces between neurons, forming clusters visible as amyloid plaques. This disrupts communication between neurons by interfering with the normal transmission of signals between neurons and triggering the formation of neurofibrillary tangles.

Neurofibrillary tangles (NFTs) consist of abnormal tau protein clumps that develop inside neurons in the brain [72]. Tau is a protein that usually helps stabilize the neurons' structure by binding to microtubules, which are long, thin fibres that provide support and transport materials within the neuron. In addition to inferring neuron function, NFTs can lead to neuronal death. In

the brain, tau proteins are usually modified by adding a phosphate group, a process called phosphorylation, which helps regulate their function. However, in Alzheimer's, tau proteins are hyperphosphorylated, leading to abnormal aggregation and the formation of neurofibrillary tangles.

Amyloid plaques, neurofibrillary tangles, oxidative stress, neuro-inflammation, genetics, and epigenetics all contribute to Alzheimer's disease development. Oxidative stress occurs when an imbalance arises between reactive oxygen species (ROS) production and their removal [73]. These highly reactive molecules can damage cells and tissues, leading to neuronal death and beta-amyloid and tau protein accumulation. Inflammation, the body's natural response to injury or infection, also plays a role in Alzheimer's by activating microglia, the brain's primary immune cells [74]. Microglia release pro-inflammatory cytokines, such as interleukin-1 beta (IL-1β) and tumour necrosis factor-alpha (TNF-α), which can cause neurotoxicity and neuronal damage and promote beta-amyloid accumulation. Inflammation also activates astrocytes, another type of glial cell in the brain, which release cytokines and chemokines that can serve as Alzheimer's disease biomarkers, such as interleukin-6 (IL-6) and monocyte chemoattractant protein-1 (MCP-1). Furthermore, chronic inflammation can disrupt the blood-brain barrier, a specialized membrane that protects the brain from toxins and pathogens in the bloodstream. In Alzheimer's disease, this weakened barrier allows inflammatory cells and toxic substances to enter the brain, exacerbating neuronal damage.

Alzheimer’s disease can be divided into early and late-onset. The genes for the amyloid precursor protein (APP), presenilin 1 (PSEN1), and presenilin 2 (PSEN2) are considered to be genomic biomarkers in early-onset Alzheimer’s disease (EOAD) [75]. APP is a transmembrane protein primarily found in the synapses of neurons in the brain. The protein is involved in

various cellular processes, including synaptic plasticity, neurite outgrowth, and neuronal survival. APP is cleaved by the β and γ secretases to produce the amyloid beta (Aβ) peptides, a key component of amyloid plaques. PSEN1 and PSEN2 are the γ-secretase that cleave APP. PSEN mutations increase the ratio of Aβ42 to Aβ40; the 42 amino acid-long Aβ isoform is more prone to aggregate than the shorter Aβ40 isoform, thereby forming amyloid plaques. This finding has led to the widespread use of mutant forms of APP genes to generate animal models of AD.

APOE is the most potent risk factor and the only confirmed susceptibility locus of LOAD [68]. The most common genotype of APOE is APOE3, which has an odds ratio (OR) estimated at around 3.2; APOE3 has a cysteine (Cys) amino acid at position 112 and an arginine (Arg) at position 158. APOE4, on the other hand, is a variant with an OR estimated at around 14.2; APOE4 arginine (Arg) at positions 112 and 158 due to a single nucleotide polymorphism. APOE4 is the strongest known genetic risk factor for late-onset Alzheimer's disease. Clusterin (CLU), bridging integrator 1 (BIN1), complement component (3b/4b) receptor 1 (CR1), and triggering receptor expressed on myeloid cells 2 (TREM2) along with 50 other risk loci have been associated with LOAD in genome-wide association studies of APOE4. These genes were related to the Aβ pathway, immune system, lipid metabolism, and synaptic function.

Studies have reported increased and decreased DNA methylation levels in AD, which could be partially due to using various tissue samples. A significant increase in DNA methylation has been reported in multiple brain regions, including the hippocampus, entorhinal cortex, dorsolateral prefrontal cortex, temporal cortex, and temporal gyrus [76]. The prefrontal cortex, locus coeruleus, and blood samples observed a DNA methylation decrease [77]. The presence of a change in the methylation signature in the blood of Alzheimer's patients has promising potential to serve as a biomarker for the disease. These changes could potentially be

used as a diagnostic tool to identify individuals at risk for developing Alzheimer's disease or to monitor disease progression.

AD can be classified according to Braak stages, based on the extent and distribution of NFTs in the brain. In the initial stages, Braak I and II, NFTs are localized in the transentorhinal and entorhinal regions, respectively, often without significant clinical symptoms. As the disease progresses to Braak III and IV, NFTs extend to the hippocampus and other limbic structures and then to parts of the association cortex, correlating with the onset and worsening of cognitive impairment. In the advanced stages, Braak V and VI, NFTs are widespread throughout the association cortex and eventually across much of the cerebral cortex, leading to severe cognitive decline and dementia.

1.2. State of the Arts

1.2.1 Style and Coherence of the Thesis

This thesis is presented in a manuscript-style format, with individual research chapters prepared for publication in peer-reviewed journals. Each chapter addresses distinct but interrelated aspects of epigenetic mechanisms and/or machine-learning applications in genomics studies research, aligning with the overarching theme of methylomic and epigenomic studies. Although each chapter is designed as a more in-depth stand-alone manuscript for publication, the research chapters collectively form a cohesive investigation into methylation in the context of transcription factor interactions or their implications in kidney renal papillary cell carcinoma and Alzheimer’s disease. Together, they address the overarching research objective of understanding

transcription factor co-binding and DNA methylation through biological and machine-learning approaches.

1.2.2 Synthesis of Literature on the Application of Machine Learning in Genomics Studies

As the importance of machine learning (ML) in genomics research continues to grow, researchers are leveraging its capabilities to analyze large, high-dimensional datasets. One such instance is the modification Juan A. Botía et al. (2017) proposed to the standard Weighted Gene Co-expression Network Analysis (WGCNA) approach, which incorporates k-means clustering as an additional processing step [78]. WGCNA is widely used for identifying gene co-expression networks (GCN) and deriving gene clusters or modules, helping to identify functionally related genes and understand the molecular mechanisms underlying complex diseases. Integrating k-means clustering generates more accurate network partitions and enhances biological meaningfulness, fostering fruitful downstream analyses.

In miRNA-gene datasets, a research group addressed the curse of dimensionality by implementing an EM algorithm [79]. The group developed a program called miREM, which employs the EM algorithm to enhance the prediction and prioritization of miRNAs from gene sets of interest. By modelling complex relationships between miRNAs and mRNAs, miREM successfully overcame the curse of dimensionality in miRNA-gene datasets. The method couples the EM algorithm with the common approach of hypergeometric probability, resulting in improved predictions compared to existing programs.

Addressing the curse of dimensionality is particularly relevant in genomic studies, and researchers have devised methods to reduce data dimensionality while preserving essential

features and relationships between variables. In one study, researchers implemented a lasso-based clustering method designed explicitly for high-dimensional genomic data with small sample sizes. This approach, known as robust sparse k-means clustering (RSKC), adaptively selects a subset of genes or proteins that contribute to partitioning samples into age-related clusters progressing across the lifespan [50]. The clusters identified by RSKC demonstrate a complex relationship between chronological and brain age, spanning a range of ages. This method solves the challenge of high dimensionality, leading to more accurate and meaningful results that can aid in understanding human brain development.

In a study by Wei et al. (2015), a five-CpG-based assay for ccRCC prognosis was developed using a lasso model based on genome-wide CpG methylation profiling. This classifier, validated in three independent datasets, predicts the overall survival of ccRCC patients independently of standard clinical prognostic factors [80]. The five-CpG-based classifier categorizes patients into high-risk and low-risk groups, with significant differences in clinical outcomes across respective clinical stages and individual 'stage, size, grade, and necrosis' scores. Moreover, methylation at these five CpGs correlates with the expression of five genes: *PITX1, FOXE3, TWF2, EHBP1L1,* and *RIN1*. This demonstrates its potential as a practical and reliable prognostic tool for ccRCC.

Lastly, Baucum M. et al. (2020) used a combination of hidden Markov models (HMMs) and recurrent neural networks to create a hidden Markov recurrent neural network (HMRNN), which improved the accuracy of forecasting in Alzheimer's disease patients [51]. This hybrid model combined the interpretability of HMMs with the flexibility of neural networks, providing a more comprehensive data analysis. The researchers demonstrated that combining different machine learning algorithms can enhance the performance of available genomic tools for

forecasting or classification. Although HMMs and EM differ in design, both rely on iterative approaches to estimate unknown parameters from observed data. HMMs are designed explicitly for sequential data with hidden states, whereas EM is a more general framework that can be applied to different models. Incorporating additional patient covariates can improve parameter estimation and predictive performance in HMMs.

Together, these studies illustrate the importance of machine learning techniques in genomics research and highlight various approaches to overcome the challenges posed by high-dimensional datasets. While each method has its strengths and limitations, incorporating multiple machine learning algorithms can provide more accurate and meaningful results for understanding the molecular mechanisms underlying complex diseases.

1.2.3. Synthesis of the Literature on DNA methylation in kidney renal papillary cell carcinoma and other renal cell carcinomas.

In a comprehensive molecular characterization of 161 primary KIRP tumours, The Cancer Genome Atlas Research Network (2016) discovered that Type 1 and Type 2 KIRP tumours are distinct renal cancer types characterized by specific genetic alterations [62]. Type 1 tumours were linked to MET alterations, while Type 2 tumours were associated with CDKN2A silencing, SETD2 mutations, TFE3 fusions, and increased expression of the NRF2-antioxidant response element (ARE) pathway. A CpG island methylator phenotype (CIMP) was observed in a distinct Type 2 KIRP tumours subgroup, marked by poor survival and gene encoding fumarate hydratase (FH) mutation.

Due to the lack of symptoms in early-stage renal tumours, there is an urgent need for reliable biomarkers to detect the presence of cancer and monitor patients during and after

therapy. Lasseigne et al. (2014) examined genome-wide DNA methylation alterations in renal cell carcinoma (RCC), identifying two panels of DNA methylation biomarkers that reliably distinguish tumours from benign adjacent tissue across all common kidney cancer histologic subtypes and another that does explicitly so for clear cell RCC tumours(ccRCC) [56]. These biomarkers have been independently validated and demonstrate promising potential for clinical application in the early detection of kidney cancer.

While the focus thus far has been on the molecular characteristics and epigenetic alterations in KIRP, it may be beneficial to consider the findings in clear cell renal cell carcinoma (ccRCC) for KIRP research in lieu of a lack of KIRP-specific research. KIRP and ccRCC are subtypes of renal cell carcinoma, and the molecular mechanisms and epigenetic alterations discovered in ccRCC may provide valuable insights for understanding KIRP. In ccRCC, many genes are epigenetically inactivated by promoter hypermethylation, some associated with clinical outcomes. Ricketts et al. (2012) observed hypermethylation of *Fibrillin 2 (FBN2), secreted frizzled-related protein 1 (SFRP1),* and *basonuclin 1 (BNC1)* genes in a large percentage of ccRCC tumours, with *SFRP1* and *BNC1* hypermethylation being significantly associated with poorer survival [81]. The loss of *SFRP1* protein can potentially activate the WNT pathway, leading to poorer survival outcomes in patients with somatic mutations of WNT pathway-regulating genes.

Fisel et al. (2013) investigated the role of DNA methylation in the regulation of monocarboxylate transporter 4 (MCT4) expression in clear cell renal cell carcinoma (ccRCC) [82]. MCT4 is involved in lactate transport across membranes, resulting in antiapoptotic effects. The study found that MCT4 protein expression was upregulated in ccRCC and was associated with cancer-related death. DNA methylation in the *SLC16A3* promoter, which encodes *MCT4*,

was identified as a novel epigenetic mechanism for MCT4 regulation, with higher methylation at individual CpG sites associated with prolonged survival.

Turajlic et al. (2018) conducted a large-scale study on clear-cell renal cell carcinoma (ccRCC), analyzing 1,206 primary tumour regions from 101 patients [83]. The study identified up to 30 driver events per tumour and found that subclonal diversification was associated with known prognostic parameters. The authors categorized ccRCC into seven evolutionary subtypes, ranging from tumours with early fixation of multiple mutational and copy number drivers to highly branched tumours with extensive parallel evolution. The study also suggested that genetic diversity and chromosomal complexity determine patient outcomes in ccRCC.

In conclusion, there is a need for more research into reliable biomarkers in early-stage renal tumours and the integration of potentially relevant findings in ccRCC to KIRP research in lieu of the lack of research in DNA methylation finding in KIRP. By examining these renal cancer subtypes, researchers can gain valuable insights, identify novel therapeutic targets and diagnostic biomarkers, and ultimately improve patient outcomes.

1.2.4. Synthesis of the Literature on DNA Methylation in Alzheimer's Disease.

Numerous studies have examined changes in DNA methylation in Alzheimer's disease (AD) patients. Research indicated 71 CpGs linked with AD pathology burden in 708 autopsied brains within genes such as *ABCA7* and *BIN1*, which contain AD susceptibility variants [84]. Concurrently, another study identified 948 CpG sites across 918 genes associated with late-onset AD (LOAD) in the frontal cortex, suggesting that DNA methylation alterations could influence AD onset and progression [85]. However, these studies could not distinguish between DNA methylation and hydroxymethylation, pointing out the necessity for further research to clarify the impact of these epigenetic modifications in AD.

In related research, K. Lunnon et al. (2014) discovered a differentially methylated region in the ANK1 gene linked to neuropathology in the entorhinal cortex, a key AD-affected area [86]. This highlighted the utility of sequential replication designs in identifying methylomic variations associated with complex diseases such as AD. Another study found 858 differentially methylated sites related to 772 genes, enriched in AD genetic risk loci and associated with normal aging processes, particularly in cell adhesion, immunity, and calcium homeostasis [87]. Furthermore, an independent study validated 11 differentially methylated regions in a set of 117 subjects and identified genes like *ANK1, CDH23, DIP2A, RHBDF2, RPL13, SERPINF1*, and *SERPINF2* with altered RNA expression in AD [84]. Lastly, Tan et al. (2013) explored the role of the *BIN1* gene as a significant risk locus for LOAD, detailing its involvement in tau pathology modulation, endocytosis, inflammation, calcium homeostasis, and apoptosis, as well as the potential methylation of the *BIN1* promoter, illustrating the intricate genetic and epigenetic interplay in AD [88].

One study focused on comprehensively fine-mapping DNA methylation (DNAm) at enhancers in neurons [70], revealing that these enhancer regions in AD neurons were hypomethylated. Hypomethylation of enhancers in AD neurons suggests a disruption in the epigenetic regulation of gene expression, which may contribute to the development and progression of the disease. The same research group found that integrating their epigenetic and transcriptomic data demonstrated a pro-apoptotic cell cycle reactivation in post-mitotic AD neurons, possibly contributing to neuronal dysfunction and loss in AD.

Vitamin B supplements have shown promise in delaying or maintaining cognitive decline in elderly adults, with meta-analysis results indicating significant effects on global cognitive function and homocysteine levels [89]. Overall, the literature review highlights the complex

relationship between methylation and Alzheimer's disease, emphasizing the importance of further research to better understand the roles of these epigenetic modifications in AD pathogenesis and their potential as therapeutic targets.

1.2.5 Synthesis of Literature on YY1 and TAF1

Yin Yang 1 (YY1) is a multifaceted transcription factor central to various cellular activities with a known binding DNA sequence [90]. Structurally, the gene for YY1 is highly preserved across eukaryotic organisms, suggesting an essential biological significance across eukaryotes. YY1 features four zinc fingers at its C-terminus at the protein level, enabling it to bind to DNA securely. Beyond this, the protein contains various domains facilitating interactions with other proteins [90], [91]. Initially, YY1 was identified for its role in repressing the adeno-associated virus, such as inhibiting its replication [92]. YY1 has since evolved in the scientific literature to be a versatile regulator of numerous genes. Reflecting this duality, its name symbolizes its capacity to activate and repress gene transcription, depending on the cellular context.

YY1 is pivotal in various cellular processes, including cell proliferation, differentiation, and apoptosis, making it highly relevant in cancer research. YY1 can function either as a tumour suppressor or an oncogene, depending on the cellular context. One study demonstrated that silencing YY1 increases the activity of the *RKIP* promoter in lung cancer [93], [94], illustrating that YY1 and RKIP regulate each other's expression through feedback loop mechanisms in an inverse reciprocal manner. Additionally, the activity of YY1 is influenced by its cofactors; methylation of *YY1* and these cofactors can affect mRNA expression levels and broader biological functions, such as immune infiltration [93]. This interaction underscores the complexity of YY1's role in cellular regulation and its potential impact on cancer pathology.

YY1 acts as a repressor through three primary mechanisms [30], [91]. Firstly, it can engage in direct competition at binding sites. This involves YY1 competing with activating factors for overlapping DNA binding sites, leading to diminished promoter activity. Notably, the α-actin muscle regulatory elements (MREs) and the serum response element (SRE) of the FBJ/FBR osteosarcoma gene exemplify this overlap. A second mechanism by which YY1 can act as a repressor is interfering with other transcriptional activators. For instance, the c-fos promoter has overlapping sites for YY1 and SRE and features additional YY1 sites. When YY1 attaches to these distal sites, it represses the upstream CRE promoter. YY1 can obstruct the conventional communication pathways of other activators, like the cAMP response element binding (CREB) protein, hampering their activation. Certain coactivators, like E1A, can block YY1-induced repression by disturbing the YY1–CREB interaction. In its third model of repression, YY1 recruits corepressors that either directly induce transcriptional repression or stimulate chromatin restructuring. This restructuring enhances YY1's ability to interact with DNA and repress transcription. YY1's zinc-finger and glycine-rich domains are pivotal for its repression activity. If these regions are altered, the repression capabilities of YY1 fusion proteins are compromised. To boost its repression, YY1 collaborates with several cofactors, including mRPD3, GATA-1, and members of the Smad family. In short, YY1's transcriptional regulation intricacies encompass competitive interactions, direct interference with other activators, and multifaceted collaborations with corepressors and cofactors.

YY1’s activating role can be conceptualized through three primary models [91]. First, the direct activation model suggests YY1 promotes transcription through direct interaction with other transcription factors, such as TATA-binding protein (TBP), TAF, and transcription factor IIB (TFIIB), utilizing specific activation domains within its structure. This model, however,

might represent a simplified version of YY1's intricate regulatory mechanism, with additional cofactors potentially playing a role [91]. In the second model, YY1 can act as an activator through the cofactor-induced inhibition of its repressive function. In this mechanism, structural alterations in YY1, possibly induced by interactions with other cellular elements, unmask the N-terminal activation domain, facilitating a transition from a repressive to an active state [91]. Lastly, the recruitment of coactivators model proposes that YY1 can initiate transcription by recruiting other activating factors to the target promoter [91]. This often involves coactivators with histone acetyltransferase (HAT) activity, enhancing transcriptional activation by facilitating more accessible DNA interactions [95].

TAF1 is one of the TBP-associated factors (TAFs) and a component of the TFIID complex [96], [97], [98]. TAF1 recognizes and binds to specific core promoter elements, such as the initiator (Inr) and the downstream core element (DCE), helping fine-tune the promoter activity based on the gene and cell type. In addition, TAF1 can modulate the binding of TFIID or TBP to the core promoter, acting as both a repressor and activator of basal transcription, depending on the context. It can inhibit the DNA-binding activity of TBP, a component crucial for initiating transcription, through two separate regions at its N-terminus, designated as TAND1 and TAND2. These regions interact with TBP, blocking TATA recognition and competing with other factors that facilitate the TBP-TATA complex formation, thereby regulating transcription activity. Moreover, TAF1 plays a crucial role in recognizing and interacting with various core promoter elements, showcasing its importance in broader promoter recognition, especially in TATA-less promoters. It may function as a promoter-marking factor, helping in transcriptional memory as cells progress through the cell cycle, maintaining its association with active gene promoters even during mitosis.

1.2.6 Gaps in Current Knowledge and Potential Research Directions

One notable gap in genomics research is the limited focus on DNA methylation signatures in KIRP. Although studies on ccRCC offer valuable insights, there is a critical need to broaden investigations to include the molecular and epigenetic characteristics of KIRP. Deepening our understanding of the methylation signature and its impact on KIRP can enhance our biological comprehension of the mechanisms underlying renal cell carcinoma and may lead to novel therapeutic and diagnostic biomarkers. The role of DNA methylation in regulating transcription and transcription factors in KIRP remains relatively underexplored. Previous research on MCT4 regulation in ccRCC has unveiled new epigenetic mechanisms potentially affecting patient survival [82]. Applying the framework in this thesis to other studies, alongside a genome-wide analysis, could uncover methylation variations linked to AD, thereby broadening the network of implicated genes and identifying potential epigenetic modifiers.

Another area that may require further exploration is understanding methylation dynamics in Alzheimer's disease. The literature review emphasizes the importance of additional research to better understand these epigenetic modifications' roles in AD pathogenesis and their potential as therapeutic targets. Additionally, sequential replication design and genome-wide analysis could be employed to detect methylomic variations linked to complex diseases like AD, further expanding the network of genes implicated in the disease and suggesting potential epigenetic modifiers.

Despite considerable research, the roles of YY1 and TAF1 in gene regulation through methylation are not fully elucidated, particularly regarding their interactions with other transcription factors. A gap exists in our understanding of the molecular pathways that govern these interactions, often mediated through epigenetic changes such as DNA methylation.

Investigating how YY1 recruits co-activators and corepressors and how TAF1 affects TFIID's binding to core promoters, especially in TATA-less promoters, remains largely unexplored. Moreover, studying the effects of post-translational modifications on YY1 and TAF1's activity, such as their impact on DNA-binding affinity and protein interactions, presents a rich field for exploration.

Another critical area for research is the dynamic regulatory spectrum of YY1 and TAF1 in response to various cellular stimuli and environmental factors. This encompasses a detailed analysis of how YY1 is modulated during cellular stress, developmental processes, and disease states. It is essential to understand the implications of its domain structures in diverse cellular contexts, including gene silencing and chromatin organization. In the case of TAF1, its multifaceted roles in transcriptional memory and interactions with different core promoter elements offer an expansive landscape for investigation.

Lastly, the therapeutic potential of YY1 and TAF1 is an underexplored avenue. While their critical roles in fundamental cellular processes are acknowledged, leveraging this understanding for disease intervention presents a novel frontier. Future research could focus on developing targeted therapies, such as specific inhibitors or modulators, which could precisely influence the activity of these transcription factors or their interacting partners. Such interventions could provide innovative approaches to modulate gene expression in various pathological conditions, marking a significant step in gene regulation and therapeutic development.

This thesis addresses some of these gaps, highlighting the need for further comprehensive research across various aspects of genomics. By advancing our understanding of DNA methylation, gene expression regulation, and the interactions of transcription factors, we can

uncover novel therapeutic targets and diagnostic biomarkers, ultimately improving disease management and treatment outcomes.

CHAPTER 2: Optimizing KIRP Prognosis Prediction: Leveraging EM-Enhanced K-Means Clustering for DNA Methylation Signature Analysis in KIRP

2.1 Publication Stage and Co-author Statement

This chapter forms the basis of the manuscript titled "Predicting Patient Survival Outcomes by Combining k-means and Expectation Maximization Approaches with Methylomic Data." Preparation of the manuscript has been paused to prioritize other sections of this thesis. I, Gatonguay Siu, am the primary author responsible for the conception of the study, data analysis, coding, and manuscript preparation. Dr. Touati Benoukraf provided guidance on study design, while Dr. Roberto Tirado-Magallanes contributed several key concepts used in my early research and model development.

2.2 Objective and Machine Learning Implementation

2.2.1 Objectives

We hypothesize in this section that integrating an estimation maximization (EM) algorithm with k-means clustering can significantly improve the classification of distinct patient populations based on DNA methylation patterns in Kidney Renal Papillary Cell Carcinoma kidney cancer, thus enhancing our overall ability to understand KIRP and implement different treatment strategies. This approach will allow for discovering key differentially methylated CpG sites, which can serve as potential biomarkers for KIRP prognosis and contribute to developing personalized treatment strategies. To this end:

- The first objective is to create a working, testable algorithm that combines a k-means clustering algorithm and an estimation maximization (EM) algorithm to reduce noise and identify patient clusters with distinct methylation patterns associated with overall survival (OS) in KIRP patients.
- Following this, we will evaluate the performance of the ML approach in classifying KIRP patients based on OS in comparison to traditional k-means clustering of methylation signatures alone.
- Investigate the clinical features and differentially methylated CpG sites significantly enriched in each patient cluster, assessing their potential as biomarkers for predicting KIRP outcomes.
- Demonstrate the utility of the proposed approach in identifying key DNA methylation CpG sites associated with KIRP and improving the accuracy of KIRP prognosis, which could contribute to developing personalized treatment strategies.

2.2.2 Implementation of hybrid k-means EM investigative model

EM is a general-purpose iterative algorithm used in machine learning to maximize a posteriori (MAP) estimates of parameters in statistical models; in our case, the parameter is the probes used in the k-means clustering. Given that we know the outcome of the patients, the EM part of the algorithm would be considered supervised, and the k-means would be regarded as unsupervised. EM is beneficial in situations where some of the variables or parameters are unobserved or hidden. The idea behind EM is to alternate between estimating the missing or hidden variables and updating the model parameters based on these estimates. This process continues until the estimates of the model parameters converge to a maximum or a local maximum; a flowchart summary of the process is illustrated in figure 6.

2.2.2.1 The machine learning CpG biomarker isolation workflow:

Phase 1, called Initialization (e0), involves using all the available probes as a baseline. This means that the (probe × sample) matrix initially contains a row for each DNA CpG probe with no NA value in any samples. The matrix is then grouped into different populations/clusters using k-means clustering. Each cluster is scored based on clinical data and organized into predicted survival outcome groups.

Phase 2, called the Estimation Maximization Inner Loop, involves copying the probes from the current best list and creating a new estimation by randomly adding and removing a few probes.

1. The probes from the current best list are copied to create a trial clone.
2. The trial clone randomly adds and removes a unique set of probes to create a new estimation. Up to half of the current probes are removed, and up to the same number of unused probes are added.
3. Unsupervised k-means clustering is performed on the samples using the probes in the new estimation, resulting in three groups (k = 3).
4. Scoring uses clinical data to compare the new estimations to previous ones. A new best-estimated probe list is saved if one of the following criteria is met (in order of priority):
    a. The clustering of the worst outcome group is more accurate.
    b. The best OS group has a better censored-to-death ratio without hindering the worse outcome group.

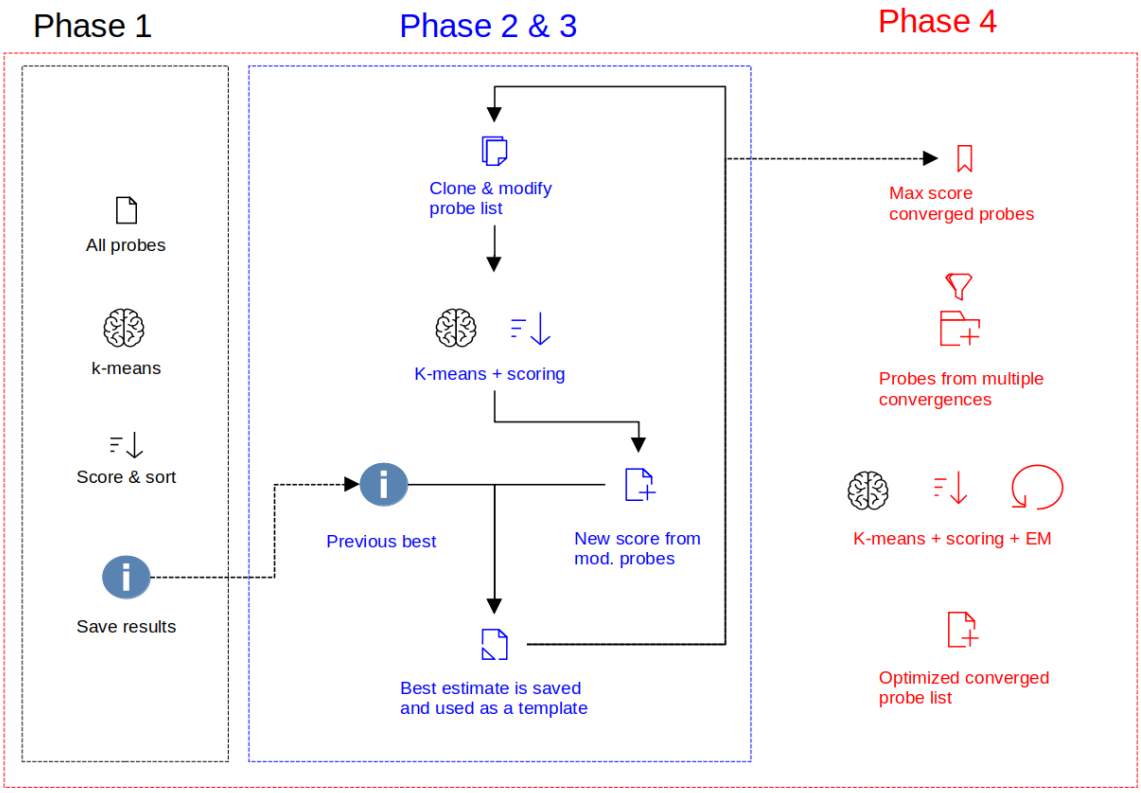
c. Scoring improves because more samples are added to the extreme ends of the groups, indicating that patients are correctly classified into the best and worst predicted survival groups without losing accuracy in prediction.
d. The new estimation has fewer probes, but the scoring remains the same. This maximizes the weight of new probes or removes probes in the k-means clustering step.

5. The best estimation is fed back into the k-means clustering in step A. The process is repeated until a predetermined number of new estimation trials fail to improve the best estimation, indicating that the best list of probes for k-means clustering to segregate survival outcomes in KIRP tumour samples has been reached.

Phase 3 follows a process similar to the Estimation Maximization Inner Loop in Phase 2 but with slight modifications to the scoring criteria. After optimizing the scoring in Phase 2, the focus in Phase 3 shifts to enhancing the model's robustness. To achieve this, the scoring criteria now prioritize adding probes when the scoring is equal to the best estimation.

Phase 4, the Estimation Maximization Optimization, saves the probe list that yields the best results in Phase 3 and removes them from the potential probe pool. The algorithm returns to Phases 1-3, using fewer probes, and iterates ten times (Conv.1 - Conv.10). Finally, the list of probes from the previous ten instances is initialized at the start of Phase 2. In contrast, the rest of the probes remain available to be added and exchanged during the EM cycles, resulting in an optimized list of probes. At the end of Phase 3, with the new probes, our optimized list of CpG probes is finalized.

.



.

Figure 6. The flowchart illustrates the steps to combine k-means and estimation maximization into three phases for a machine learning algorithm. Phase 1, initialization of e0, involves using all the available probes as a baseline for future phases in the algorithm. Phase 2 & 3, the Estimation Maximization inner loop, consists of copying the probes from the current best list and creating a new estimation by randomly adding and removing a few probes until the probes converge onto the maximum possible score. Phase 4, the Estimation Maximization Optimization, involves probe lists from several converged iterations to build a new list of probes that converges and an optimized probe list that more accurately represents a global maximum score.

2.2.2.5 The metric for the scoring and gradient for the investigative model

Each patient will die due to cancer or stop following up with checkups. However, no longer following through with follow-ups does not necessarily mean survival, but the elapsed time since the initial cancer diagnosis may provide helpful information. Therefore, it is crucial to have information about the Last Communication Contact from the Initial Pathologic Diagnosis Date (lc_from_IPDD) as well as the date of death from the initial diagnosis [99], [100].

We used the F1 to identify the cluster with the poorest projected overall survival rate. The F_1 score is a commonly used metric for evaluating the performance of machine learning models in binary classification problems, particularly when the data classes are imbalanced. It ranges from 0 to 1, where 1 represents the best possible score. F_1 is a harmonic mean of precision and recall, two other important evaluation metrics, which ensures a balance between them. Where recall is the proportion of true positive predictions among all actual positive instances, and precision is the proportion of true positive predictions (TP) among all positive predictions. In other words, recall measures how many of the actual positive instances were correctly identified by the model (i.e., the model's ability to find all relevant cases within a dataset), while precision measures how many of the instances predicted as positive by the model were actually positive (i.e., the accuracy of the positive predictions made by the model).

The F_1 scores evaluate the hyperparameters in our predictive machine-learning model. Specifically, we used this score to identify the CpG sites that best group patients in the poorest overall survival group. It's crucial to highlight that whether we utilized the F_1 score or another metric, the general trends remained consistent. However, the specific highlighted CpG sites varied based on the chosen metric.

$$\text{Precision} = D / (D + (l/\alpha)) \quad (1)$$

$$\text{Recall} = D / (D + \sigma) \quad (2)$$

$$F_1 = \frac{2(precision)(recall)}{precision + recall} \quad (3)$$

The F_1 is calculated using Equation 1-3, where 'D' represents the count of patients within the cluster who died due to cancer, 'l' signifies the number of days counted in lc_from_IPDD, 'α' is the median survival period (in days) before patients died of cancer across all patient data, and 'σ' is the number of know patients that succumb to cancer in another cluster.

The score (R_{score}) is used to determine the best predicted overall survivability group and is calculated using equation 2:

$$R_{score} = (D/L) * (d/Beta) \quad (4)$$

In this context, 'D' stands for the count of patients who have died from cancer, 'L' represents the number of samples in the cluster, 'd' is the number of days patients survived before succumbing to cancer, and Beta is the median value of the number of days since the last communication contact from the initial pathologic diagnosis date (lc_from_IPDD) for all clinical data patients. Note that there are cases where patients who have died from cancer did not have recorded lc_from_IPDD days. If there are no reported cancer deaths, 'D' and 'd' are assigned a value of 0.01 to avoid yielding a score of zero. A lower R_{score} is desirable for identifying the most effective probes for survivability prediction.

The final condition under which a set of probes could be regarded as the new optimum is when a smaller probe list is found that achieves the same score. The computational time efficiency is an important constraint of the Expectation-Maximization (EM) algorithm. This algorithm operates on multiple iterations with linear complexity, decreasing performance as input data increases. To address this, we've strived to limit the number of CpG probes involved in the computation to those most likely to be genuinely significant. To ensure this, the final scoring condition checks whether the new set of probes has a smaller total number while maintaining the same score.

2.2.3 Implementation of the Hybrid Predictive Model

The ML predictive model underwent several notable changes to improve efficiency and standardize the metric used to evaluate hyperparameters. In the context of machine learning, hyperparameters are predefined settings that control the training process and are not learned directly from the data. In our model, the hyperparameters are the probes, each representing a scalar observation of a single location along the genome that can have a varied methylation level. These probes are utilized to determine the dimensionality of the k-means clustering process, where the number of clusters is fixed at three. The purpose of this step is to reduce noise in the observations, enabling the model to focus on meaningful patterns in the data.

The centroids of the k-means clusters, which represent the model parameters, are derived from the training data and subsequently used for predictions and classifications. The dataset was divided into training and testing subsets to ensure reliable model evaluation. The training data comprises 90% of the samples selected randomly, while the remaining 10% are held out for the testing phase. This workflow ensures that the model's performance is validated on unseen data, allowing for an accurate assessment of its predictive capabilities.

Phases 1 -3 remain the same. In phase 4, the list of probes from the previous instances is combined to create an extensive list of CpG probes used to determine the centroid locations in hyperdimensional space using k-means clustering. The location of each centroid is saved and held constant to be used to predict the grouping of new samples. The metric for scoring remained unchanged from the previous classification model

2.3 Material and Methods

2.3.1 TCGA Data Acquisition and Selection Criteria

We obtained DNA methylation beta value profiles for all TCGA patients from the National Cancer Institute (NCI) Genomic Data Commons (GDC) using the TCGAbiolinks API [99], [100]. We restricted samples to KIRP (Kidney Renal Papillary Cell Carcinoma) TCGA PanCancer Atlas for the main section of our analysis to ensure uniformity in clinical information, consistent processing, and normalization of copy numbers [101], [102]. Furthermore, we only used files sequenced via Illumina Infinium Human Methylation 450K BeadChip, which facilitated data processing using the manifest files associating each probe with additional information. We extracted the patient's race, weight, age, gender, and pathologic tumour stage from the American Joint Committee on Cancer (AJCC), as well as the survival information for each patient from the clinical data table, available on either the GDC or cbioportal websites [99], [100]. For the gene expression analysis, the data was also downloaded from GDC; however, the RNA-seq data was processed for patient confidentiality; as such, the files already calculated the TPM values for each patient sample; all samples with Illumina 450k DNAm profiles had corresponding RNA-seq quantization data.

2.3.2 NCI GDC Methylation Array Harmonization Workflow and Pre-processing.

The NCI GDC utilizes a Methylation Array Harmonization Workflow to process raw methylation array data from Illumina Infinium DNA methylation arrays (HM27 and HM450) and EPIC platforms. This workflow measures the level of methylation at known CpG sites as beta values by calculating array intensities (Level 2 data) using the formula Beta = M/(M+U), where M represents the methylation intensity signal, and U represents the unmethylated intensity signal measured by the Illumina 450k array. The SeSAMe [103] software package processes the raw methylation array data. From the SeSAMe output files, we utilized the Methylation Beta Value TXT file derived from the two Masked Methylation Array IDAT files.

2.3.3 Beta and M-values Calculation and Their Significance

The Beta-value method has a direct biological interpretation [104], [105], ranging from 0 (unmethylated) to 1 (fully methylated) and represents the average of DNA methylation within a specific cytosine across all reads (i.e. depth of coverage), but they can be affected by background noise. They may be less suitable for detecting small methylation differences. M-values have a more symmetric distribution and perform better in Detection Rate (DR) and True Positive Rate (TPR) for highly methylated and unmethylated CpG sites, making them more suitable for sequencing-based studies. We calculated M-values using the formula:

$$M_i = \log_2[\, \beta_i / (\, 1 - \beta_i \,)] \qquad (5)$$

2.3.4 Determining Optimal Clustering in Methylation Analysis

In our analysis, we utilized hierarchical clustering, grouping samples based on similarity in their methylation profiles. To maintain accuracy, we excluded any probe sites with absent M-values measuring methylation levels in at least one sample. We employed the within-cluster sum of squares (WSS) method to determine the optimal number of clusters by minimizing the variation within each group. We identified three as the ideal number of clusters using the "elbow" method, which locates the point where increasing the cluster count doesn't substantially lower the WSS.

2.3.5 Heatmap Generation and DMR and DEG Criteria

A CpG probe was considered a differentially methylated region if its average M-value difference between normal tissue samples and one of the tumour clusters exceeded ±3Z, measured against the near-normal ML cluster compared to the normal tissue group, along with an adjusted *p*-value of < 0.01. The p-value was calculated based on the paired Kolmogorov-Smirnov test with the Bonferroni-Hochberg correction. Following this, we further restricted the list of probes to those labelled as "Promoter*" in the "Regulatory_Feature_Group" column found in the hm450k manifest file [36] to reduce computational RAM to under 32Gb when building the heatmaps.

In the context of investigating correlative differentially expressed genes (DEGs) for our groups, slight modifications were made. For the DMR magnitude and binary heatmap generation, we randomly sampled the same 75% of the DMR regions to reduce computational RAM to under 32G yet retain information about the distribution of the DMR locations within the genome. The “Relation_to_UCSC_CpG_Island” of the manifest file gave us information about

the location of each of our DMRs (not restricted to the 75% sampling). For the DEGs, less restrictive criteria were used in which any gene would be considered differentially expressed if at least one of the groups had a p-adjusted value of < 0.01 based on the pairing of the Kolmogorov-Smirnov Test, with the Bonferroni-Hochberg correction.

2.3.6 Survival Analysis and Clinical Features

We used the Kaplan-Meier (KM) survival curve to estimate the survival function from data containing censored data points and truncated or missing values. We built a list of patient TCGA IDs for each cluster and saved them on the cbioportal website as unique groups to create the KM survival curves later. Of the 320 samples that fit our initial research criteria, 266 remained in the cbioportal database that could be used in the KM-survival curve.

We also included the clinical features of each patient on the cbioportal website. These features include sex, age, and tumour stage. We downloaded this data in table format and analyzed it based on clusters at the e0 stage, and afterward, the EM k-means algorithm optimized the clustering of each patient sample. The data for the fraction of genome altered and mutation count was also downloaded from the cbioportal website in a boxplot format, in which the P-values were calculated using the Kruskal-Wallis test along with the Benjamini & Hochberg method. Although the Kruskal-Wallis test can use two or more independent groups, we restricted the test to a pairwise comparison.

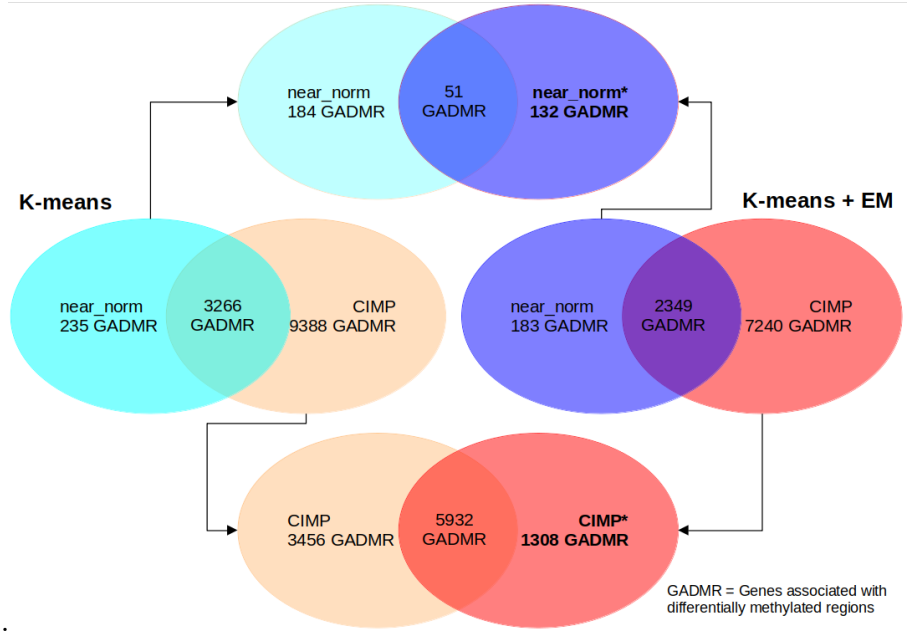
2.3.7 Global Analysis of CpG Methylation, Genes, and GO Terms

We analyzed all differentially methylated CpG sites, taking a comprehensive approach rather than limiting our focus to the CpG sites previously identified as biomarkers by our machine learning (ML) algorithm for patient grouping. Differentially methylated probes had the

same criteria as when building the heatmaps, with one significant change: we used all CpG sites, not just the CpG sites that were labelled as "Promoter*" in the "Regulatory_Feature_Group" column found in the hm450k manifest file [35]. We acquired the gene symbols of the genes associated with differentially methylated regions from the hm450k manifest file. When a probe was linked to multiple genes, all related genes were included in the global enrichment analysis. We compiled a list of all known KIRP-associated genes from the National Center for Biotechnology Information (NCBI) using the keyword search "KIRP" as a reference. After downloading all genes in a table format, we filtered them for KIRP-related genes specific to *Homo sapiens*. This list allowed us to compare the differentially methylated genes against the known genes associated with KIRP. This gave us a simplified KIRP gene enrichment analysis to compare each group and condition.

Subsequently, we examined the differences between the k-means and k-means-EM algorithms in isolating genes associated with differentially methylated regions (GADMR; Figure 4). We utilized gene ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses on David's platform [106]. This allowed us to pinpoint significant changes in GO and KEGG terms isolated within each cluster when adding the EM stage to the ML algorithm, leading to insights in the context of the underlying biological processes, molecular functions, and KEGG pathways that significantly impact overall survival. This comparison employed a significance threshold of $P < 0.05$ for the subset of genes under review.

.



.

Figure 7. Venn Diagrams comparing gene-associated DMR Isolation by k-means and k-means-EM Algorithms: This Figure displays the overlap and differences in gene isolation between the k-means and k-means-EM algorithms in identifying genes associated with differentially methylated regions. Each diagram shows the number of shared genes isolated by both algorithms and those unique to each method.

2.3.8 Distribution analysis of the isolated biomarkers

While the primary aim of our analysis did not directly focus on the biomarkers, we recognized the importance of understanding how the isolated biomarkers' distribution compared with the overall CpG probe distributions in our dataset. We conducted a distribution analysis to assess whether the biomarkers share similar distribution characteristics with the broader set of CpG probes. We employed the Kolmogorov-Smirnov Test to compare the biomarker's averaged methylation values (m-values) distribution with the global averaged CpG probe m-values distribution. This comparison helps determine if the biomarkers' distribution patterns are statistically consistent with or divergent from the general CpG probe population. Furthermore, we applied the Bonferroni-Hochberg correction to the Kolmogorov-Smirnov Test results.

In addition to comparing the biomarkers' M-values with the global CpG probe distributions, we also examined how these biomarkers' M-values align with the distributions of specific CpG probe clusters. This secondary comparison aimed to discern whether the biomarkers follow a distinct distribution pattern within specific clusters of CpG probes, providing deeper insight into the biomarkers' behaviour and characteristics within the subset of the data.

2.3.9 Analysis of dysregulation of transcription factor motifs

An additional angle to our investigation focused on how the differentially methylated CpG markers influence transcription factor binding sites (TFBS) in the context of kidney renal clear cell carcinoma (KIRP). To do so, we used the same isolated DMR as earlier in our analysis, namely the CpG probes that had an average M-value difference between normal tissue samples and one of the tumour clusters that exceeded ±3 standard deviations (±3Z) measured of the near-

normal ML cluster compared to the normal tissue group along with a p-adjusted value of < 0.01 bases on the pairing the Kolmogorov-Smirnov Test, with the Bonferroni Hochberg correction.

The coordinates of these CpG sites were then converted from the human genome assembly hg19 to hg38 to employ them within MethMotif, a tool for studying TFBS occurrences and their methylation states across genomic loci. Using MethMotif's "Batch Query," we mapped these regions and retrieved TFBS and their CpG methylation levels in the HEK293 cell line, chosen for its similarity to kidney tissue samples. Subsequently, we isolated all CpG sites within the motif regions of various transcription factors (TFs). We performed an analysis to elucidate the overall changes in methylation seen at TF binding sites. We investigated specific TF-associated motifs with initial corresponding beta values that overlapped our differential methylated sites in at least 15 CpG markers according to Methmotif. Converting the beta values for these sites from hg38 back to hg19, we were able to compare the changes seen in the motif methylation of various transcription factors across healthy samples, cell line samples, and samples from near-normal cl3, cl2, and CIMP cluster cl1, which was done for all 12 isolated TFs by use of boxplots.

A pairwise ANOVA was conducted across the different populations. The false discovery rate (FDR) was then adjusted using the Benjamini & Hochberg method to identify the most relevant transcription factors (TFs) for further study. To corroborate the expression levels of these TFs, an additional validation step was implemented using BioPortal, which applies RSEM (RNA-Seq by Expectation-Maximization) for mRNA quantification [100].

RSEM is designed to deduce gene and isoform expression levels from RNA-Seq data. It employs statistical models to distribute reads to their probable gene or isoform origins, particularly when reads may align with multiple locations. Through its expectation-maximization

(EM) algorithms, RSEM offers refined estimates of gene and isoform expression, enhancing the accuracy of the mRNA quantification process. The RNASeqV2 data, as referenced in cBioPortal, undergoes normalization and processing using RSEM, aligning with the standards of TCGA. In this context, the "normalized_count" in the .rsem.genes.normalized_results file signifies a transformed value of the "raw_count." This transformation adjusts for technical variations across samples, such as differences in sequencing depth and library size, by normalizing against the 75th percentile of all counts and applying a scaling factor. A flowchart to summarize the methodology is shown below in figure 8.

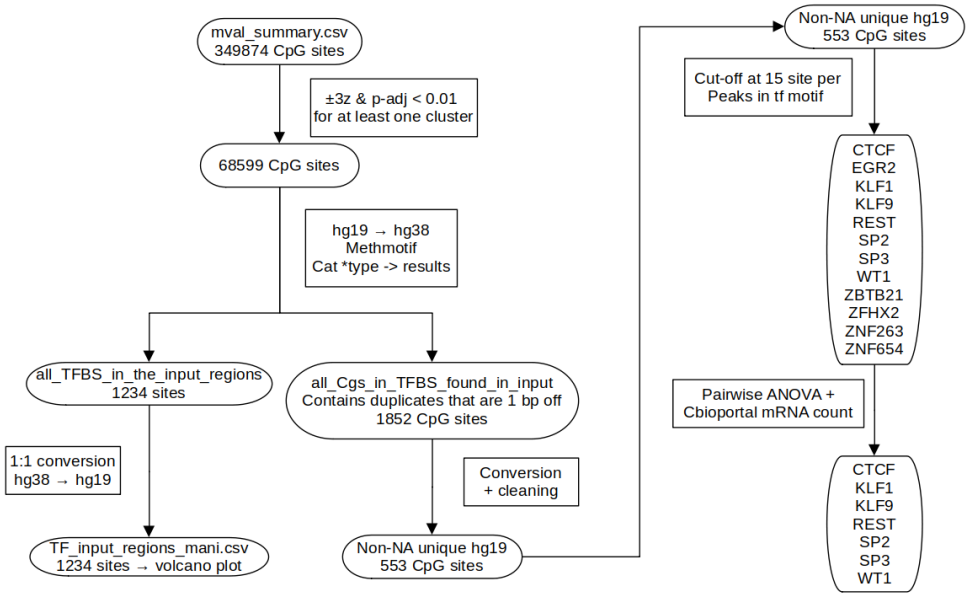


Figure 8. Flowchart of DMR Processing and Transcription Factor Analysis in KIRP. Starting with all non-NA CpG probes, the analysis narrows down to significant DMRs to examine their influence on TFBS in kidney renal clear cell carcinoma. Utilizing MethMotif, overlaps between DMRs and TFBS were identified and analyzed, followed by processing to isolate and examine transcription factors pertinent to the study's focus on differential methylation and its implications.

2.4 Results

We have developed an EM k-means algorithm to improve the prediction of survival outcomes of KIRP patients. The EM k-means algorithm will ideally cluster groups so that differential methylation analysis highlights probes associated with genes and pathways relevant to kidney cancer. To test the EM k-means algorithm, we benchmark it against unsupervised k-means clustering alone, using all non-NA hm450k CpG probes, which were also used in the E_0 (step 1 of the implementation).

2.4.1.A. CpG promoter methylation signatures clustered via all probes at e0.

We characterized a CpG island methylator phenotype (CIMP) (Figure 9.A) that exhibited predominant hyper-methylation across most CpG promoter sites, irrespective of the methylation states in normal samples in our first cluster grouping (cl1). Interestingly, a small portion of the CpGs in this cluster showed hypo-methylation in areas that typically present higher methylation levels in normal tissues. This cluster demonstrated the poorest survival rates marked by a significant downturn in the Kaplan-Meier (KM) survival curve (Figure 9.C). This group contained 10 samples, of which 6 recorded deaths, accounting for 15.79% of all fatalities in the processed dataset. The DNA methylation profile of the second cluster (cl2) depicted a minor segment of hypermethylated CpGs. The third cluster (cl3) revealed an evenly distributed methylation pattern, with a slight inclination towards lower CpG methylation levels. The KM survival curves for cl2 and cl3 were similar, plateauing at 66.16% and 63.90%, respectively.

2.4.1.B. DNA methylation signatures clustered using EM algorithm

In this round, KIRP tumours were again segregated into three DNA methylation subgroups through k-means during the heatmap creation process. CpG biomarker probes were

selected for this process based on the hybrid of k-means and estimation-maximum. The first cluster (Cl1) demonstrated the worst overall survival rate, as shown by a complete drop in the KM survival curve (Figure 9.D). This group included 27 samples with 21 reported deaths, making up 55.26% of all recorded fatalities in the post-processed dataset. Similar to the previous analysis, the DNA methylation profile of Cl1 exhibited a CIMP profile (Figure 9.B). The second cluster (Cl2), like its k-means-only counterpart, showed a section of hypermethylated CpGs, albeit with a more extensive range. The cluster with near-normal methylation signatures displayed the highest survival rates, with no death recorded among the 32 samples, equating to 12.03% of all samples. The intermediate outcome group comprised the remaining 207 samples (77.81%) analyzed.

.

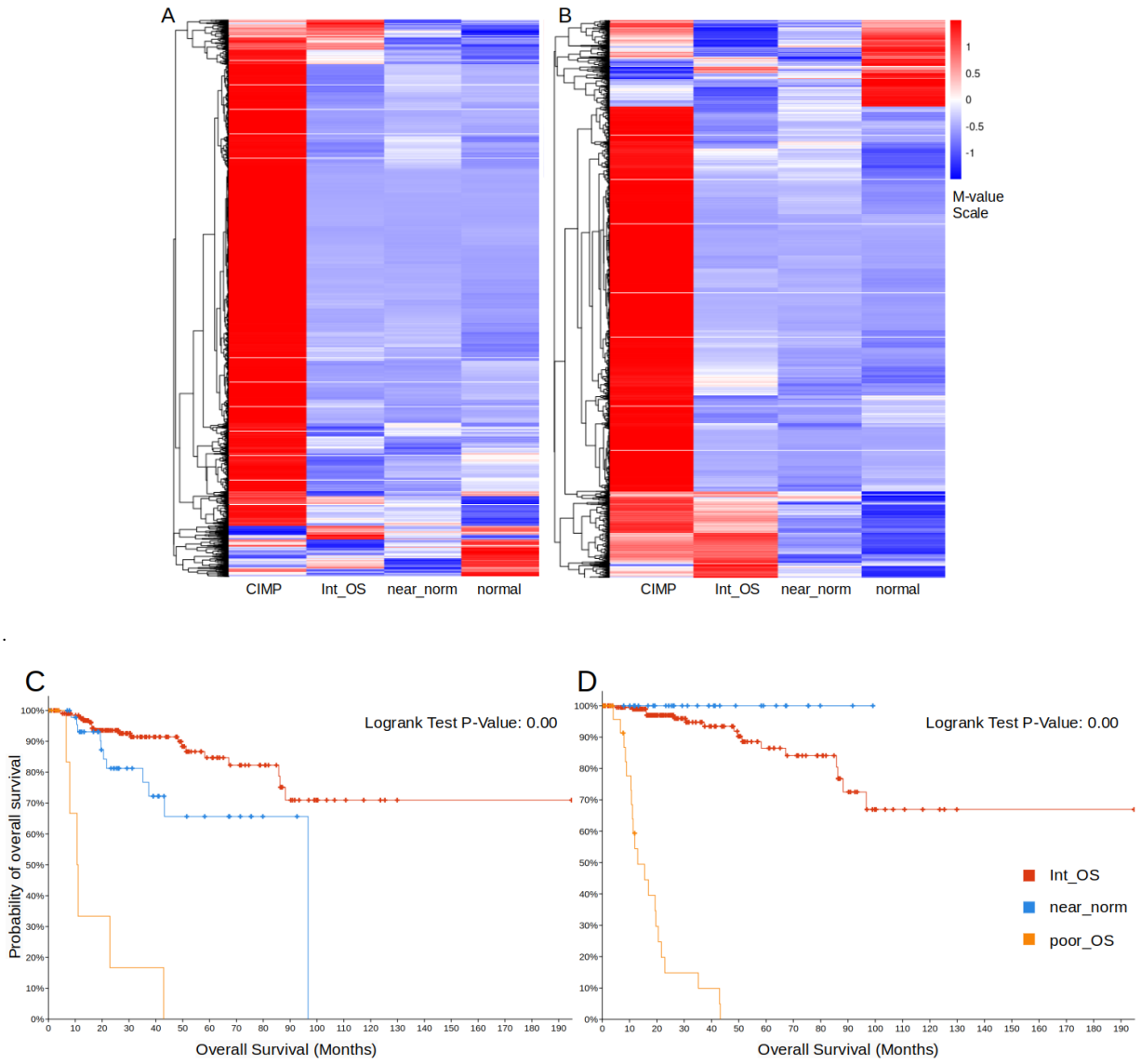


Figure 9. Heatmaps represent each cluster's average M-values for all differentially methylated promoter-associated CpG sites; the names of the clusters are semi-descriptive, with the near-normal being the tumour cluster that was grouped along with the normal tissue samples, the CIMP group being the group displaying a generalize hypermethylation pattern, the int_OS group being the intermediate overall survival group, and the norm is for the normal tissue samples. The heatmap on the left is clustered at the e0 step, with only k-means (A). The heatmap on the right organized the clusters based on the optimized EM k-means algorithm (B). Kaplan-Meier analysis of overall survival (OS) based on k-means clustering of methylation signatures in the TCGA dataset at e0 (C) and after optimization by estimation maximization (D).

2.4.2 Comparative Analysis of Genomic Instabilities

In the exploration of the mutational landscape of three distinct patient groups characterized by their methylation profiles and survival outcomes, the Fraction of Genome Altered (FGA) and mutation rates were analyzed using cbioportal [99], [100]. High FGA and a high number of mutations can both indicate genomic instability and are often observed in cancer cells. While FGA deals with large-scale structural changes in the genome, such as copy number alterations, including amplifications (increased copy number) or deletions (reduced copy number) of chromosomal segments, mutations refer to specific alterations in the DNA sequence. These changes can be small-scale, affecting individual nucleotides (point mutations), such as insertions, deletions, and rearrangements of smaller DNA segments. The comparative analysis of FGA and mutation rates across three patient groups revealed notable patterns (Figure 10.A-B). Group CL1, with a CIMP-high profile, exhibited the highest median FGA at 0.23, coupled with a low mutation log2 value of 5.13, establishing the correlation between a CIMP profile in KIRP patients and large-scale genomic instabilities. In contrast, Group CL2, with intermediate survival outcomes, showed a moderate median FGA of 0.16 and the highest median mutation log2 value of 6.08. Group CL3, with the highest survival and a near-normal methylomic profile, had the lowest median FGA at 0.08, yet an unexpectedly relatively high median mutation log2 value of 5.64. This indicated that although the CL3 group had a low level of large-scale mutations, this group did have an overall high level of point mutations within the genome. The results after optimizing the set of CpG biomarkers by means of the EM k-means clustering groups for analysis resulted in similar trends (Figure 10.C-D). The data for the fraction of genome altered and mutation count was also downloaded from the cbioportal website in a boxplot format, in which the P-values were calculated using the Kruskal-Wallis test along with the Benjamini &

Hochberg method. Although the Kruskal-Wallis test can use two or more independent groups, we restricted the test to a pairwise comparison.

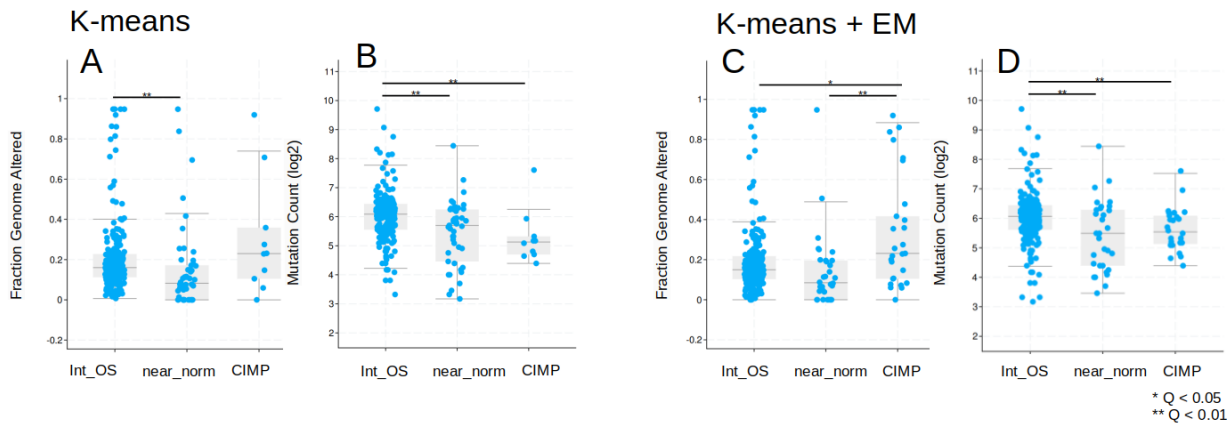


Figure 10. Comparative boxplots of Fraction of Genome Altered (FGA) and mutation rates (log2 values) across three patient groups with distinct methylation profiles and survival outcomes in KIRP (A-B) using only k-means clustering to categorize patients or (C-D) using the ML algorithm to categorize patients.

2.4.2 Differentially expressed gene & differentially methylated regions

Our analysis of differentially methylated regions (DMRs) identified 68,599 sites from the initial Illumina data using the K-means EM algorithm to classify patients. A substantial overlap of DMRs was observed among patients, with a notable negative correlation between DMR isolation and overall survival (Figure 11). The CpG island methylator phenotype (CIMP) group contained 67,874 (98.94%) of these DMRs, with 42,508 uniquely isolated within this group. Patients with intermediate overall survival (OS) exhibited 53,383 (77.82%) DMRs, with 12,167 being unique, while the best OS group had 41,557 (60.58%) DMRs, 498 of which were unique.

Further investigations revealed distinct genomic patterns: DMRs unique to the CIMP group were predominantly within island regions (53.0%), whereas those unique to the intermediate OS group were significantly less frequent in these areas (11.4%). All 3,262 differentially expressed genes (DEGs) were found only in the intermediate OS group, with none meeting the FDR threshold of less than 0.01 in other groups. Variations in log-scaled expression were recorded with standard deviations of 2.37, 2.16, and 2.48 for the CIMP, intermediate OS, and best OS groups, respectively. The average squared differences in log fold change (LFC) from normal tissue were 0.97 for the CIMP group, 2.8 for the intermediate OS group, and 0.84 for the best OS group. Altogether, this highlights distinctive molecular signatures associated with this survival category.

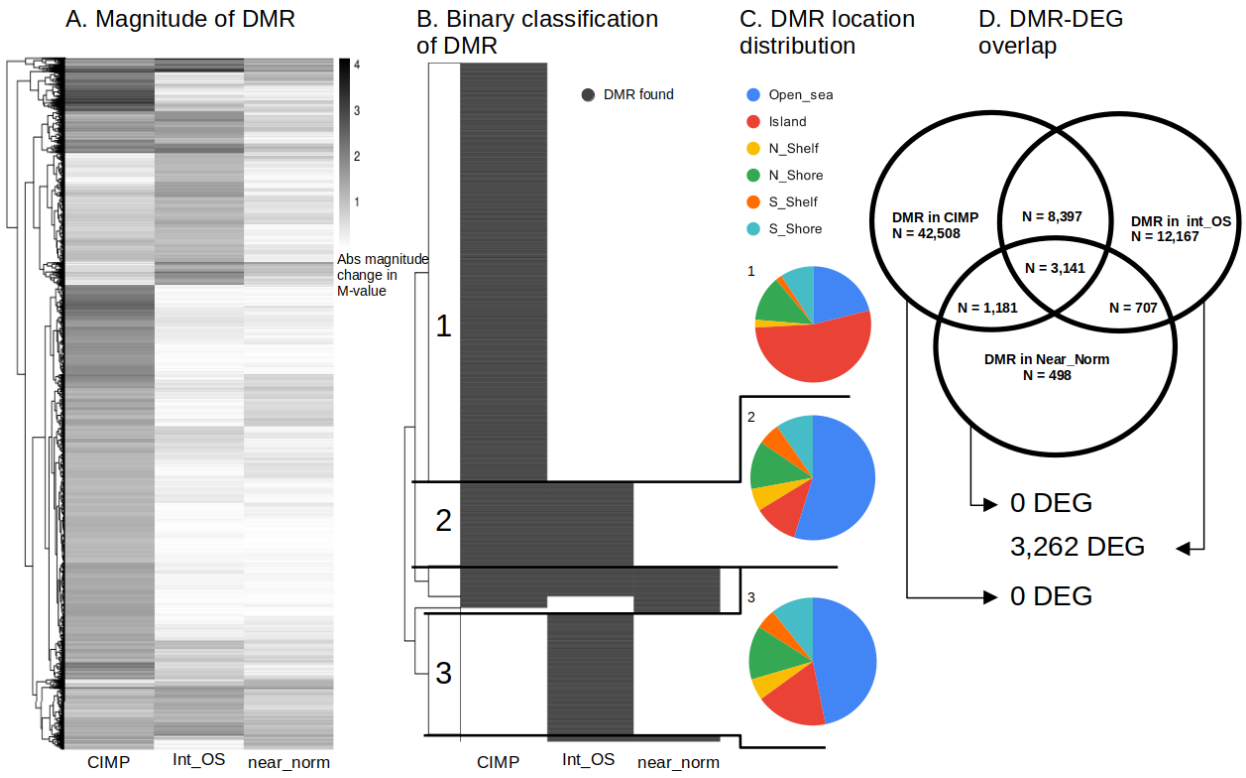


Figure 11. Correlation between DMRs and DEGs. A) depicts a heatmap that simplifies the initial DMR analysis, illustrating the magnitude of M-value shifts in each group compared to normal tissue samples, scaled using M-values. B) presents a second heatmap that adopts a binary approach to represent the overlap of DMRs across the three groups, simplifying the quantization of changes. Panel C) shows pie charts detailing the distribution of DMRs relative to CpG islands. D) features a Venn diagram providing another visual representation of the number of DMRs in each group, including overlaps.; accompanied by counts of DEGs associated with each of the three overarching classifications of patients.

2.4.3 Clinical Feature Distribution Post-Optimization

The distribution of many clinical features remained close to the same between the e0 and optimized clusters. However, some clinical features change to reflect a distribution that more closely reflects realistic distributions, such as the tumour stages association and the ratio of male/female patient samples in each group, which both showed a shift after optimization of the EM k-means algorithm. This shift is especially noticeable in the ratio of male/female patient samples for the worst survival outcome group, which starts at 70% female patients in e0, dropping to 44.4% after optimization (Figure 12).

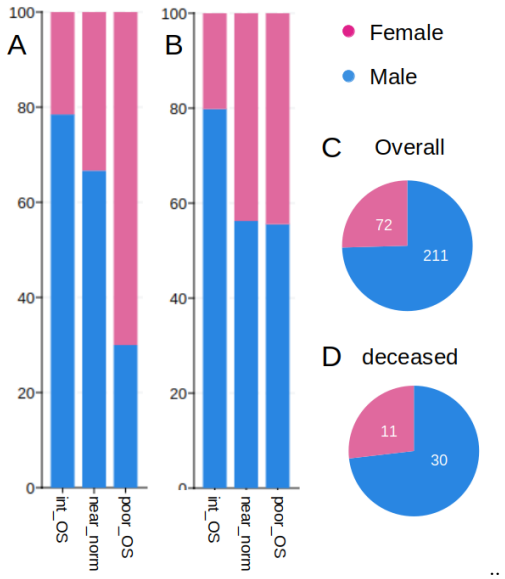


Figure 12. The ratio of male to female patients in each cluster (A) at e0 when clustered using k-means on all probes as input, or (B) when the input for clustering is optimized using the EM k-means algorithm. (C) The ratio of all male to female patients in the TCGA KIRP dataset, and (D) the ratio of all male to female patients recorded as deceased.

The tumour stage is one of the most important factors determining cancer patients' prognosis and survival outcome. Tumour staging is a process that involves determining the extent and spread of cancer in the body. It is based on factors such as the size and location of the tumour, as well as whether or not cancer has spread to other parts of the body. In stage T1, the cancer is localized to a small area and hasn't spread to lymph nodes or other tissues. In stage T2, cancer has grown, but it hasn't spread. In stage T3, cancer grows larger and possibly spreads to lymph nodes or other tissues. This stage is also referred to as metastatic or advanced cancer. The ratio stages of patients within each cluster shift after being adjusted by the EM algorithm. The ratio is calculated by equation 3:

$$\text{Ratio} = s / T \tag{8}$$

Where s is the number of patients in a given stage within a cluster, and n is the total number of patients within the cluster. When an estimation maximization is added to k-means clustering, the distribution of tumour stage by cluster and predicted survival out is improved. The group with the best survival outcome gains a higher ratio of patients in T1, going from 60.00% to 77.78%. The patients in the worst outcome group shift away from the T2 stage and towards the T3 stage after adjustments to the clustering by the EM algorithm. The T2 patient stage ratio goes from 66.67% to 40.00%, whereas the T3 stage ratio shifts from 33.33% to 40.00%. (Figure 13)

.

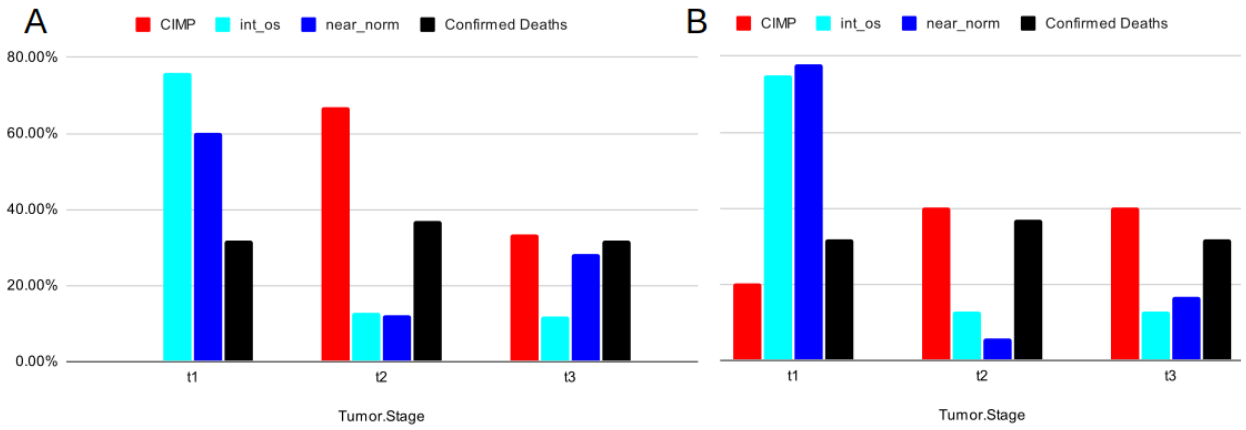


Figure 13. The percent ratio of tumour stages found with just k-means clustering using all CpG probes as input (A), compared to the optimized probe CpG input based on the estimation maximization k-means clustering algorithm (B).

2.4.4 Analysis of Genes associated with differentially methylated regions

We analyzed the genes associated with differentially methylated regions (GADMR) in our three clustered groups. Specifically, the k-means clustering approach categorized patients into three groups, one of which—the near normal methylation group —exhibited 3502 GADMRs that were significantly distinct, each with an adjusted p-value < 0.01 and an average M-value difference exceeding $\pm 3Z$. Using the same M-value threshold and adjusted p-value, the group with intermediate overall survival (int_os) displayed 7624 significant GADMRs, while the group with the CIMP signature showed 12558 significant GADMRs.

To understand the potential impact of these GADMRs on kidney renal papillary cell carcinoma (KIRP), we conducted an enrichment analysis. This analysis compared the GADMRs in each group against a dataset of genes associated with KIRP. The CIMP group's GADMRs overlapped with 362 out of 534 known KIRP genes. In contrast, the intermediate OS group (int_OS) and the near normal group (near_norm) contained fewer overlaps, with 279 and 144 known KIRP genes, respectively. (illustrated in Figure 14. A-C).

Further comparisons were drawn using an enhanced patient grouping methodology that combined k-means clustering with expectation maximization. The results were somewhat parallel; the enhanced method identified 9515 significant GADMRs in the CIMP group, 7396 in the int_OS group, and 2532 in the near normal group. The enrichment followed suit with a similar pattern as the k-means grouping method. The CIMP group's GADMRs overlapped with 298 out of 534 known KIRP genes. The intermediate OS group (int_OS) and the near normal group (near_norm) contained fewer overlaps, with 274 and 109 known KIRP genes, respectively (shown in Figure 14. D-F).

The analysis of the three clustered groups has revealed a substantial variation in the number of significantly isolated GADMR, pointing to a complex interplay between methylation patterns and patient outcomes in KIRP. The notable difference in the number of GADMR among the clusters—particularly between the near normal methylation group (cl3) with fewer GADMR and the CIMP/poor_OS group with a significantly higher number—underscores the potential of methylation patterns to affect overall survival outcome for patients diagnosed with KIRP.

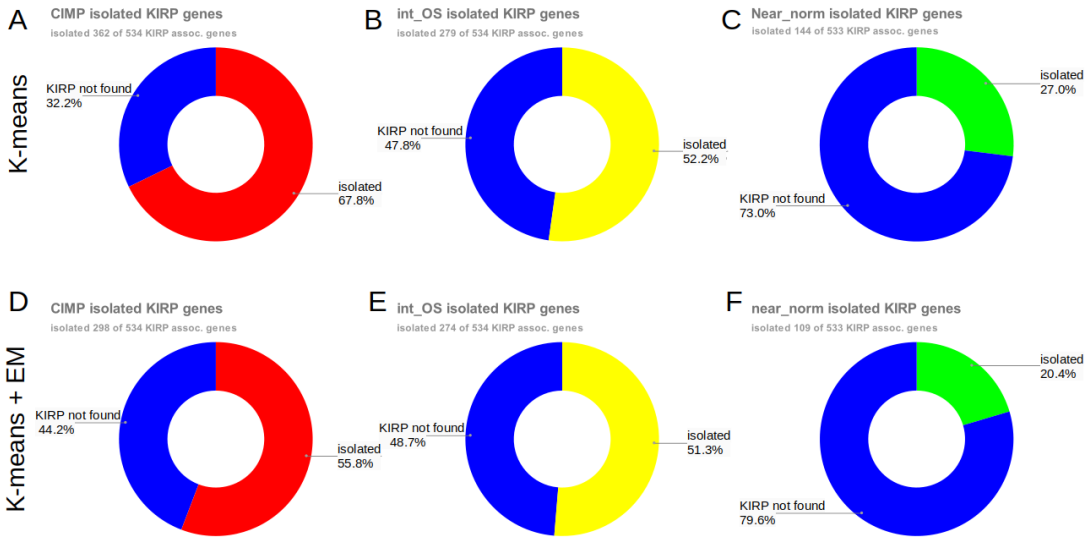


Figure 14. The pie chart represents the relative number of genes associated with differentially methylated regions that overlap with known KIRP genes found using either k-means alone or a k-means EM approach for each of our three predicted survival outcome groups.

2.4.5 Enrichment analysis of changes in the poor overall survival and near normal groups

The changes in GADMR GO and KEGG in the CIMP-like group, in which more patients could be isolated after using the EM stage to reduce noise in the methylation data, are relevant within the context of cancer metabolism. The changes isolated by the GADMR were found within biological processes and pathways that would allow the tumour cells to evade the immune system and resist apoptosis, as well as metastasis (Figure 15).

Conversely, in the near-normal group, characterized by a methylation pattern associated with better overall survival, post-EM filtering revealed GADMR-related biological processes and pathways that bolster the immune system's capacity to combat tumour cells (Figure 15). The critical takeaway from this analysis is the significant role that algorithm choice and subsequent data refinement play in deciphering the complex relationships between DNA methylation patterns, gene functions, and cancer progression. By contrasting the outcomes from k-means and k-means-EM algorithms, we gain nuanced insights into how specific methylation-related changes at the genomic level can influence critical biological pathways that dictate tumour behaviour and patient prognosis.

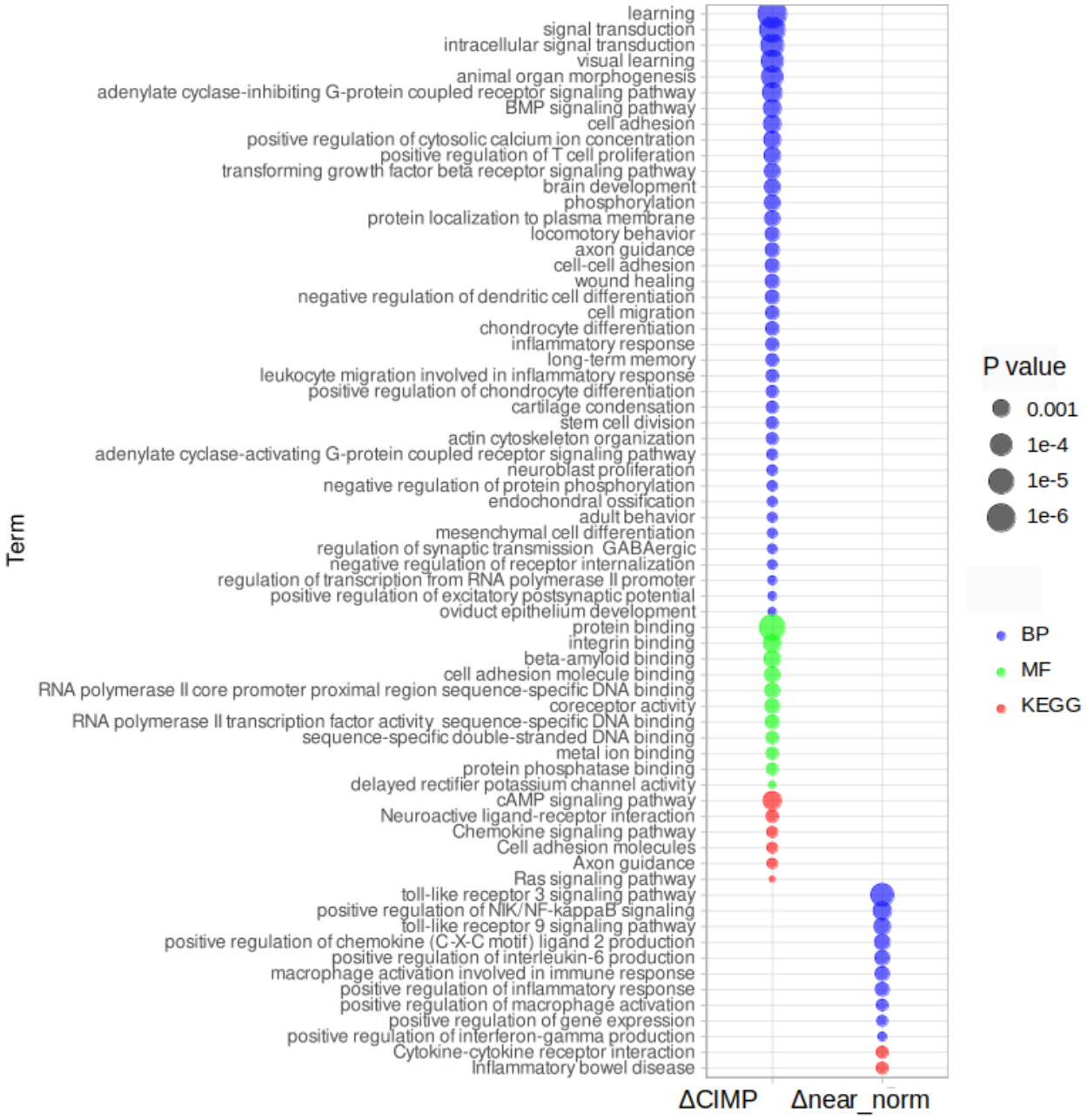


Figure 15. Bubble plot of Gene Ontology (GO) and KEGG Terms for biomarkers used in categorizing patients based on methylation profiles associated with predicted survival outcomes. The bubble plot displays all GO linked to the CpG biomarkers with a P value < 0.1. Colour represents the type of GO: blue represents biological processes, and green represents molecular functions. The KEGG pathways are also represented in red in this figure.

2.4.6. Analysis of ML Biomarkers for different patient populations

The significance of the biomarkers in the poor overall survival (poor_OS) group, as indicated by their distribution being markedly different from the overall averages of the 450k CpG probes with a P-adjusted value < 0.05, highlights the potential of these biomarkers in prognosticating outcomes for patients. This distinct distribution suggests that these biomarkers are not random variations but are statistically meaningful differences that could be linked to the underlying biological processes affecting patient survival. In contrast, the inability of the biomarker distributions in the other groups to reach statistical significance when compared to the overall average M-values implies that the differences in survival outcome may not be specifically associated with the differential methylation of these biomarkers. In essence, the biomarker's importance is limited to categorizing the patients into groups based on their distinct epigenetic profiles.

Consistent with the patterns observed in our other datasets, the differentially methylated CpG sites located within transcription factor (TF) motifs were predominantly hypermethylated in the group exhibiting poor overall survival (OS) and characterized by a CpG island methylator phenotype (CIMP), as illustrated in our volcano plot analysis (Figure 16). The ratio of the differentially methylated CpG site found within the motifs is proportionate to the number of differentially methylated sites found for each group (Ratios; CIMP: 0.013, Int_OS: 0.009, near_norm: 0.013).

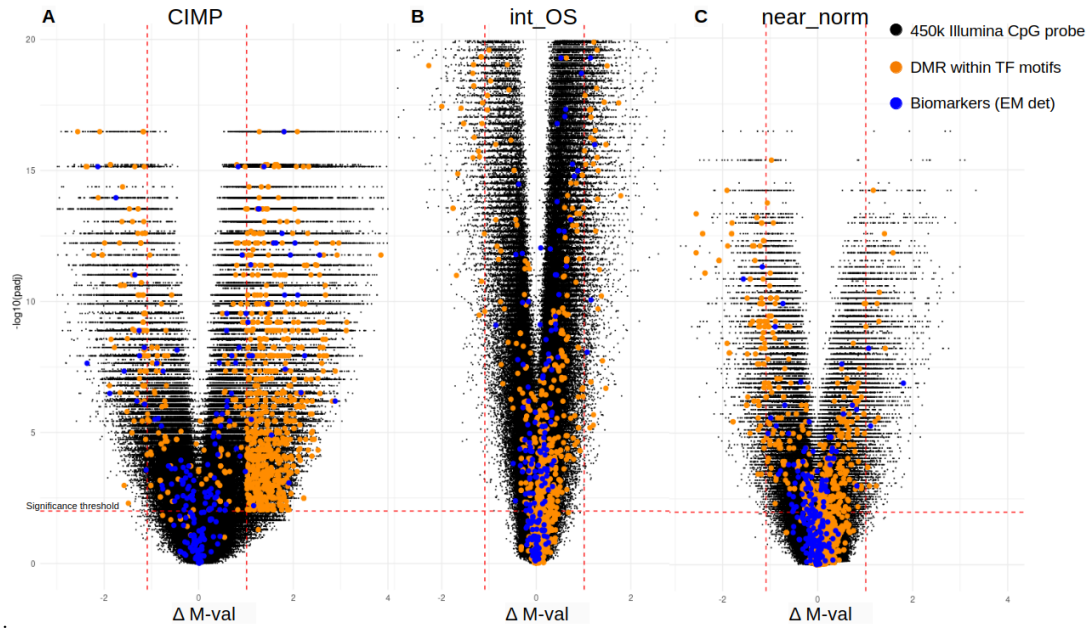


Figure 16. Volcano Plots Showing Changes in M-values Across Clusters. Volcano plots illustrate the changes in each cluster's average M-value (ΔM-val) compared to the normal value. The y-axis represents the significance of the changes, with a threshold set at p = 0.01. The x-axis shows the ΔM-val; negative values represent hypomethylation, whereas positive values represent hypermethylation. A standardized ±3Z threshold marks significant deviated samples using the vertical lines. Blue dots indicate biomarkers identified using the EM algorithm, while the orange dot represents the differentially methylated region (DMR) found within transcription factor (TF) motifs. (A) The CIMP/poor overall survival (OS) cluster is on the left, (B) the intermediate survival outcome cluster is in the middle, and (C) the near-normal/best OS group is on the right.

2.4.7 Dysregulation of Transcription Factor Motifs

The TF motifs exhibiting more than 15 overlapping CpG sites with corresponding shifts in beta values, as indicated by our data, encompass *CTCF, EGR2, KLF1, KLF9, REST, SP2, SP3, WT1, ZFHX2, ZFHX2, ZNF263,* and *ZNF654* (Figure 17). A pairwise ANOVA was conducted across the different populations for each transcription factor represented in the boxplot. FDR was then adjusted using the Benjamini & Hochberg method to identify the most relevant TFs for further study using a p-adjusted cut-off of 0.05. Additionally, we corroborate the expression levels of these TFs. We used BioPortal, which applies RSEM for mRNA quantification, of the transcription factors investigated; the subsets of biologically relevant TF were *CTCF, KLF1, KLF9, REST, SP2, SP3,* and *WT1.* The change in the methylation pattern of these TF motifs shows significance in the CIMP group compared to the HEK293 and normal samples. Significance was usually also reached when comparing the methylation levels with the int_OS or near_norm groups.

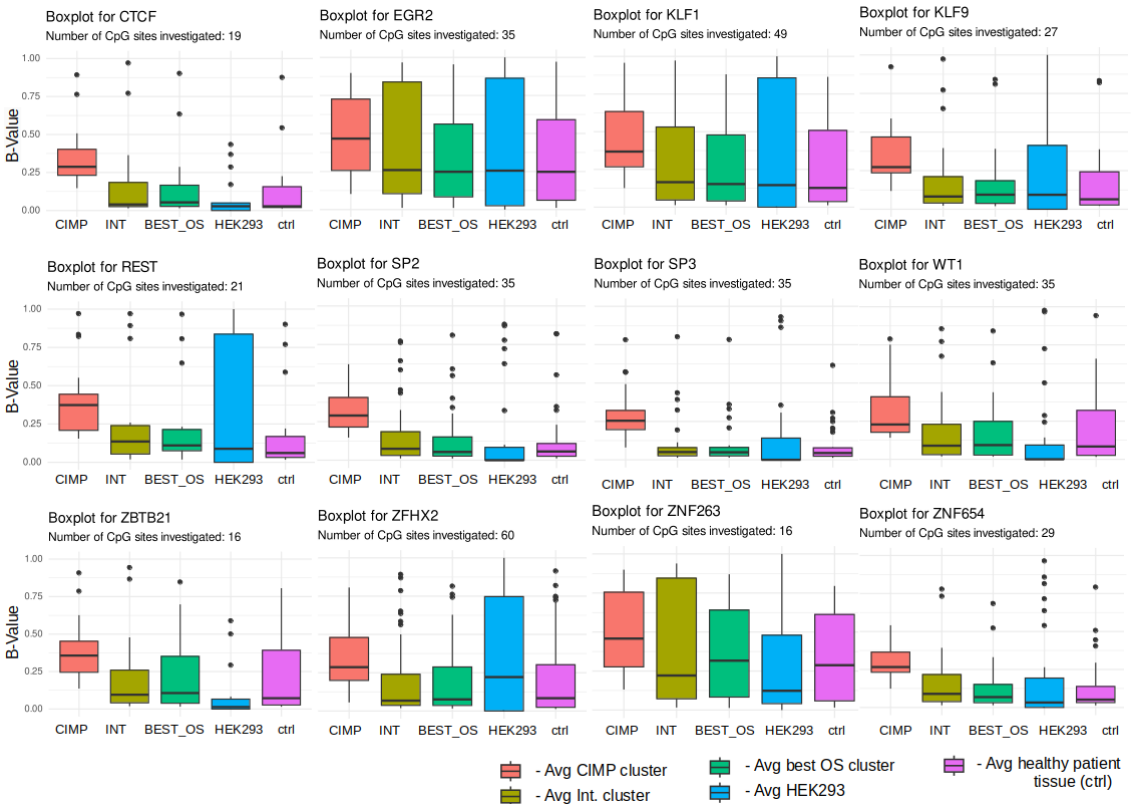


Figure 17. Comparative methylation analysis of Beta values across CIMP high, intermediate OS (int_OS), Near_norm (Best_OS), cell line (HEK293), and normal samples from healthy controls, focusing on TF motifs (*TCF, EGR2, KLF1, KLF9, REST, SP2, SP3, WT1, ZFHX2, ZFHX2, ZNF263*, and *ZNF654*) with significant CpG overlap and beta value shifts.

The overall changes brought on by the implementation of an additional EM stage within the ML algorithms in isolating genes associated with differentially methylated regions (GADMR; Figure 18-19) displayed gene ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses (on *David's platform* [107]) that were heterogeneous in when looking at the individual impact of each isolated TF motifs. In other words, no readily available pattern is discernible.

However, the sum of effects caused by differentially methylated motifs seems relevant in the context of cancer's impact on overall survival. The changes brought on by the additional stage of EM in the k-means algorithm enhanced the isolation of DMRs and allowed for more precise identification of biological processes and pathways shared by patients with poor OS. In this group, the BP and KEEG pathways generally enable tumour cells to evade immune detection, resist cell death, and metastasize with the present methylation aberrations.

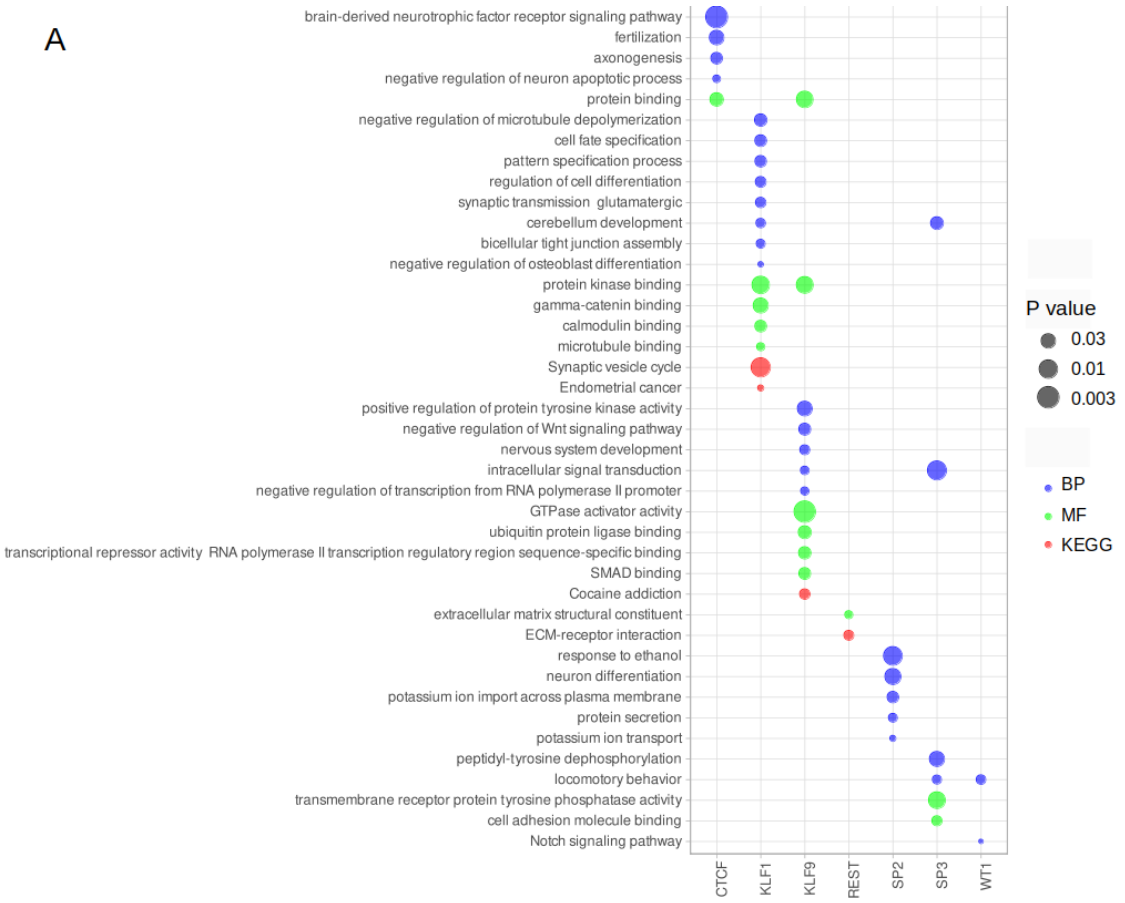


Figure 18. Bubble Plot of key pathways and functions of DMR TF motifs. This plot visualizes the biological processes, molecular functions, and KEGG pathways associated with transcription factors *CTCF, KLF1, KLF9, REST, SP2, SP3,* and *WT1*, which exhibit significant differential methylation within their motifs linked to poor overall survival outcomes. The size corresponds to the p-values, and the colour of the bubbles corresponds to the pathway of function.

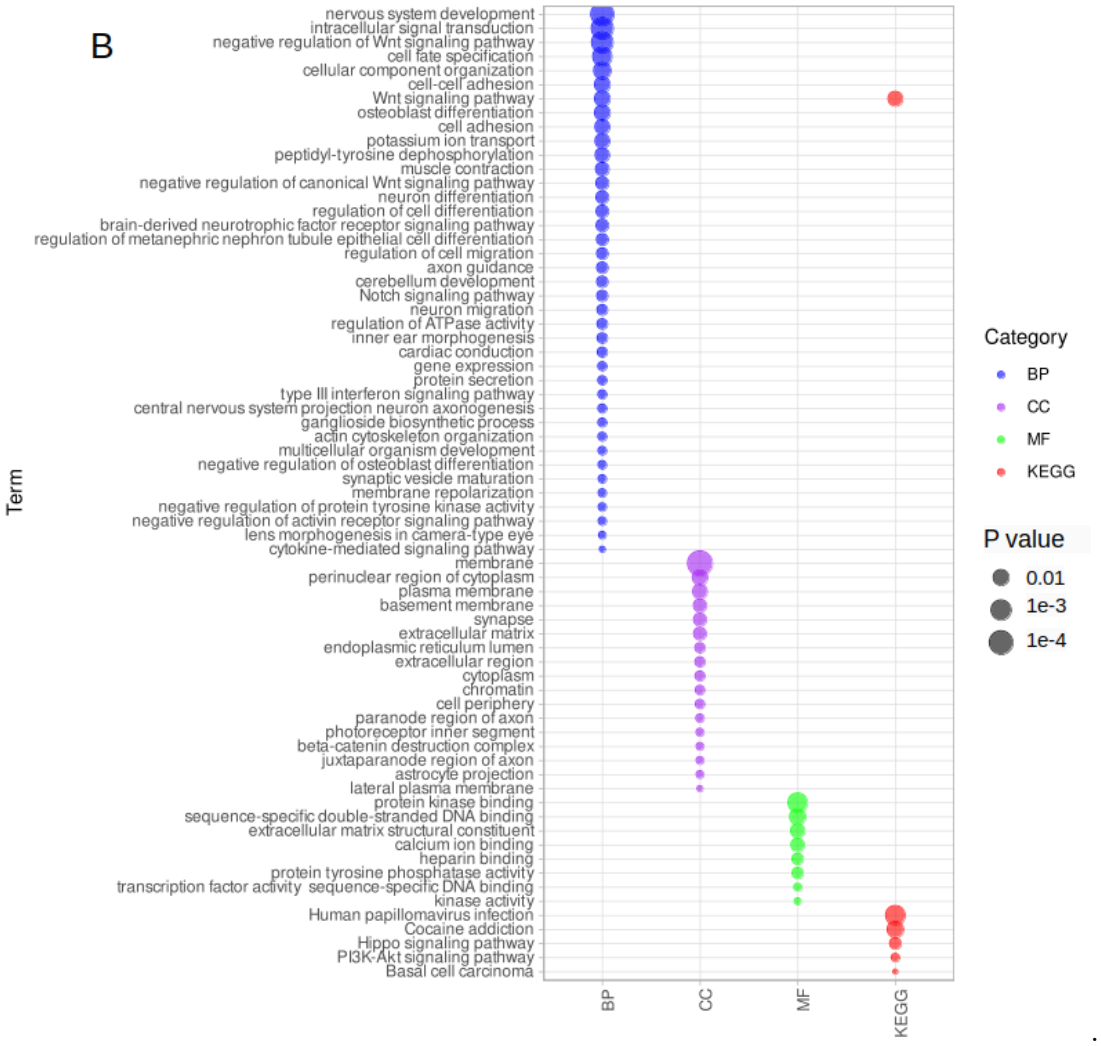


Figure 19. Comprehensive Bubble Plot of GO Enrichment and KEGG Pathway Analysis for DMR-Associated Genes. This plot elucidates the integrated impact on biological processes, molecular functions, cellular components, and KEGG pathways related to the genes associated with differentially methylated regions (DMRs) of transcription factor motifs. Instead of individual TF analysis, it aggregates the effects across genes linked to the motifs of CTCF, *KLF1, KLF9, REST, SP2, SP3,* and *WT1*. These are critical in the context of poor survival outcomes, with differential methylation serving as a key modifier. Bubble size indicates the p-value significance, while the colour highlights the specific enriched pathways or functions category.

2.4.8. Using K-means and EM to create predictive models

CpG biomarker probes were selected based on the hybrid of k-means and estimation-maximum using 90% of patient samples in the training phase. Once the CpG biomarkers were isolated, the location of the centroids was determined; the testing samples could be grouped to their nearest centroid for investigation. The F_1 score in our worst overall survival group increased in the testing data, going from 0.611 in training to 0.800 in testing (supplemental). The KM-survival curve results reflect the same trends as our investigative model. The second cluster (Cl2) demonstrated the worst overall survival rate, as shown by a complete drop in the KM survival curve (Figure 20). This group included 28 samples with 20 reported deaths, making up 52.63% of all recorded deaths in the post-processed dataset. The first cluster starts with the best-predicted survival rate, with 61 patients in this group and only 4 recorded deaths. However, of the last 3 patients that remained in the group after 85 months, 2 succumbed to cancer, breaking the expected trend. The intermediate outcome group (cl2) comprised the remaining 177 samples (66.54%) analyzed and included 14 known patient deaths.

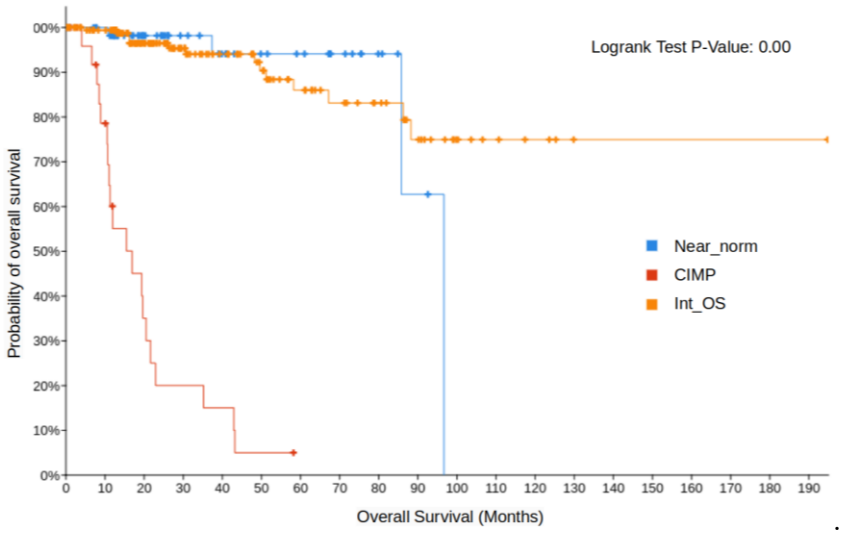
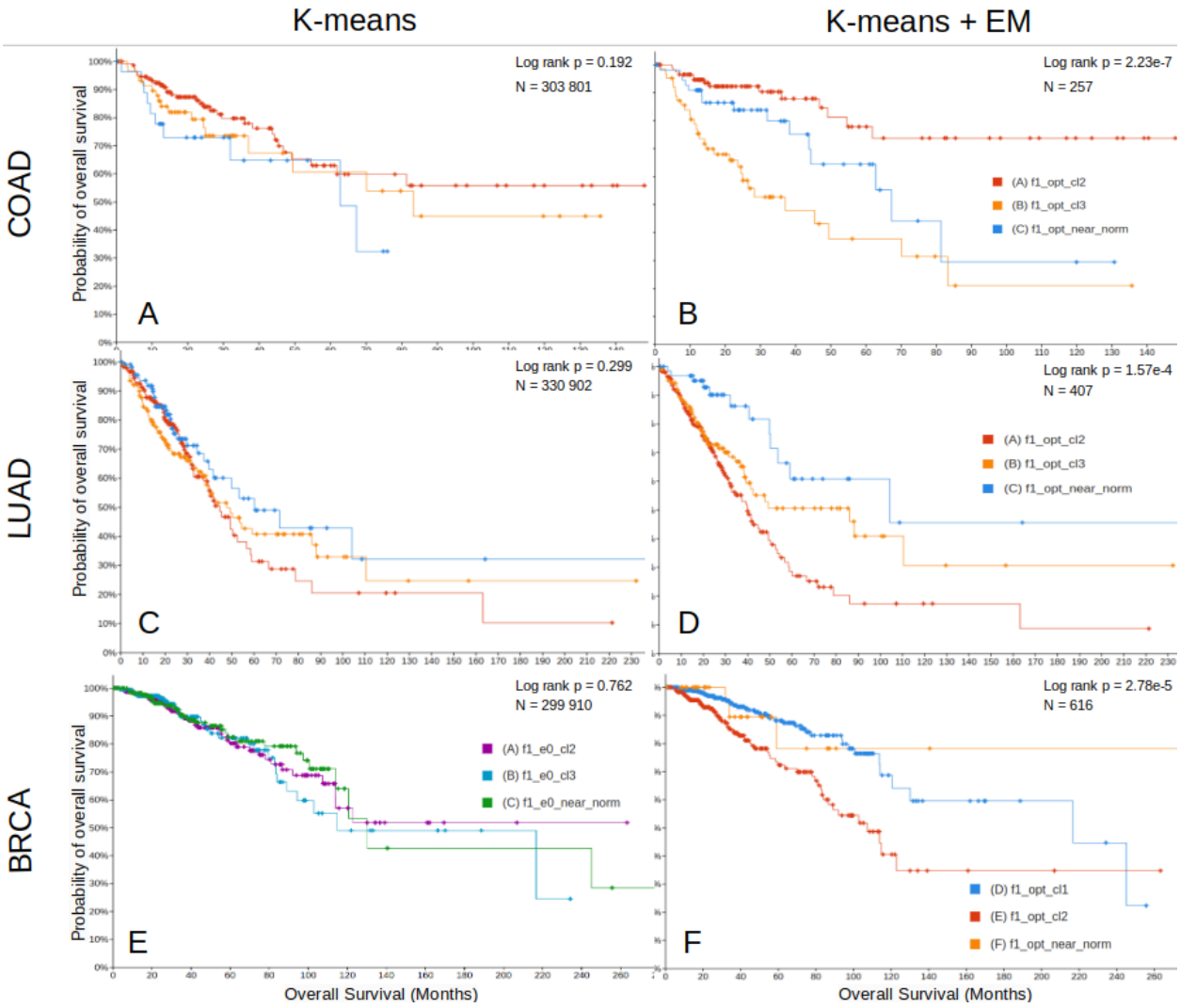


Figure 20. Kaplan-Meier analysis of overall survival (OS) of the combined training and testing samples for the ML predictive algorithm.

2.4.9. Extending the EM K-means Algorithm to Additional Cancer Types

In this study, we extended the application of our EM k-means algorithm beyond KIRP kidney cancer to assess its broader utility in oncological research. Specifically, we aimed to evaluate whether the algorithm could effectively categorize patients across various cancer types based on their methylation patterns and survival outcomes. For this purpose, we selected a subset of cancer types from the TCGA database that are known to exhibit poor survival outcomes associated with CIMP profiles, mirroring the characteristics observed in our KIRP dataset.

The selected cancers—colorectal adenocarcinoma (COAD), lung adenocarcinoma (LUAD), breast invasive carcinoma (BRCA), pancreatic adenocarcinoma (PAAD), lung squamous cell carcinoma (LUSC), and kidney renal clear cell carcinoma (KIRC)—were chosen based on their known CIMP-related survival implications. Applying the EM k-means algorithm to these cancer types, we observed a consistent enhancement in the stratification of patients according to survival outcomes (Figure 21). This improved segregation underscores the algorithm's ability to identify meaningful methylation patterns pertinent to survival across various cancer contexts. Each cancer type maintained its unique methylation signature associated with different survival outcomes; the algorithm successfully highlighted the underlying signatures within each group, facilitating further analysis by other research groups.


K-means
K-means + EM
COAD
LUAD
BRCA
Probability of overall survival
Log rank p = 0.192
N = 303 801
A
Log rank p = 2.23e-7
N = 257
(A) f1_opt_cl2
(B) f1_opt_cl3
(C) f1_opt_near_norm
B
Log rank p = 0.299
N = 330 902
C
Log rank p = 1.57e-4
N = 407
D
Log rank p = 0.762
N = 299 910
(A) f1_e0_cl2
(B) f1_e0_cl3
(C) f1_e0_near_norm
E
Log rank p = 2.78e-5
N = 616
(D) f1_opt_cl1
(E) f1_opt_cl2
(F) f1_opt_near_norm
F
Overall Survival (Months)
Overall Survival (Months)

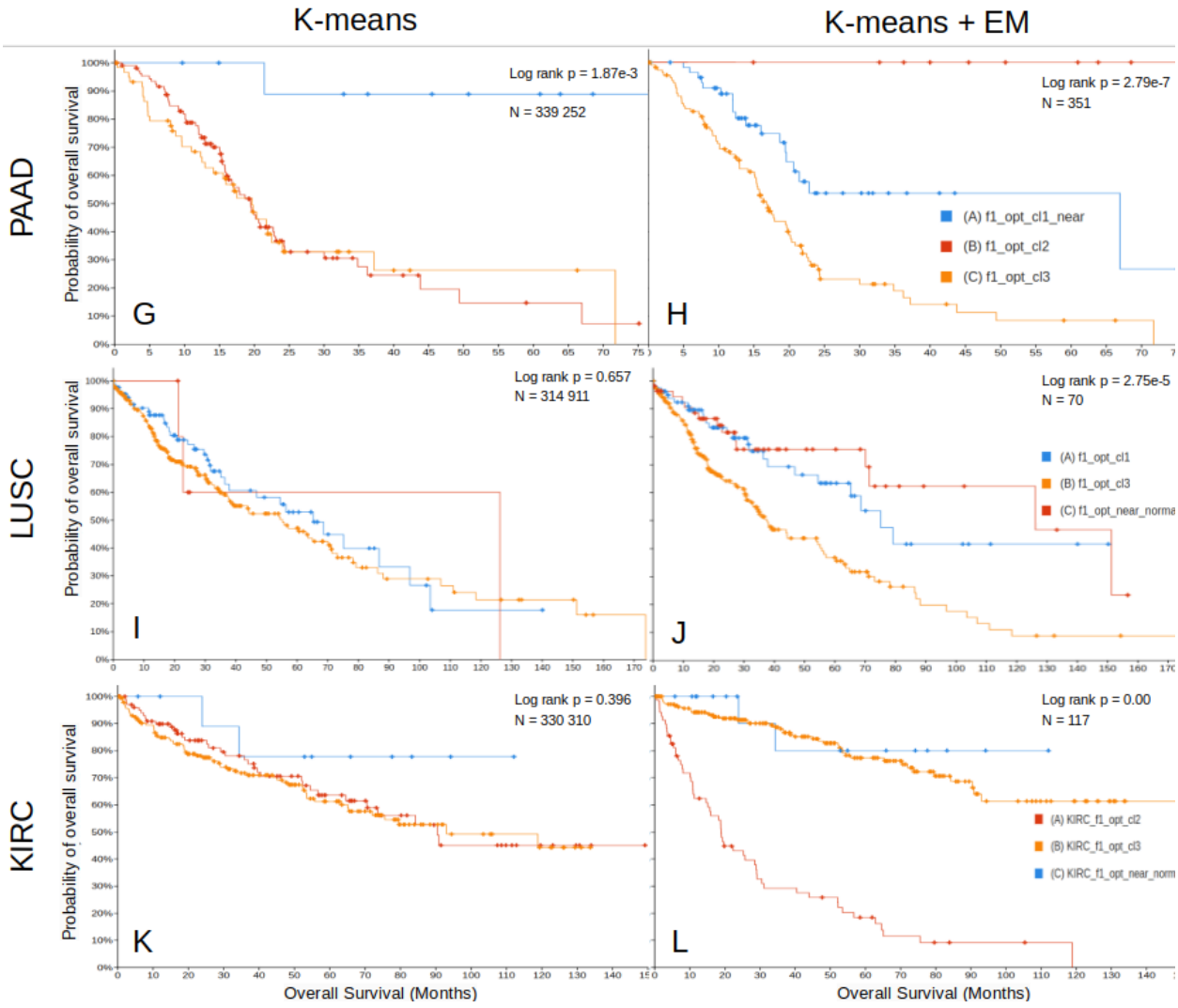


Figure 21. Comparative KM-Survival Curves Across Six Cancer Types Using k-means and Enhanced EM k-means Classification. This Figure presents twelve Kaplan-Meier survival curves for six different cancers: colorectal adenocarcinoma (COAD: A, B)), lung adenocarcinoma (LUAD: C, D)), breast invasive carcinoma (BRCA: E, F), pancreatic adenocarcinoma (PAAD: G, H)), lung squamous cell carcinoma (LUSC: I, J), and kidney renal clear cell carcinoma (KIRC: K, L)). The left-hand Figures (A, C, E, G, I, K) depict patient survival outcomes based on conventional k-means clustering. In contrast, the right-hand Figures (B, D, F, H, J, L) apply k-means on subsets identified using EM to focus on methylation biomarkers. Notably, the EM-enhanced k-means classification consistently demonstrates superior patient stratification, significantly improving the segregation between high-risk

and low-risk survival groups across all cancer types, reflecting the algorithm's efficacy in identifying prognostic methylation signatures.

2.5 Discussion & Conclusion

2.5.1 Investigation of DNA methylation in survival outcomes.

In the kidney renal papillary cell carcinoma (KIRP) study, unique DNA methylation profiles correlated with different survival outcomes were identified. These profiles underscore the significant role that DNA methylation plays across various stages and types of cancer, linking directly back to the unifying theme of my thesis: methylation. Hypermethylated profiles resembling the CIMP (CpG Island Methylator Phenotype) profile are observed in the group with poor overall survival outcomes. This finding supports the association between CIMP methylation patterns and adverse survival outcomes, which is a recurring theme in many cancer types. The distinct epidemiological and clinicopathological features of CIMP tumours, probably stemming from a unique underlying biological mechanism, further emphasize the critical role of methylation patterns in cancer subgroup characterization [108].

The CpG methylation pattern in the intermediate and best overall survival groups is more subtle; the general trend is a further drift away from the normal in the intermediate group compared to the best overall survival group (heatmap). Based on the k-means clustering, the best overall tumour samples survival group is always clustered along with the normal samples, suggesting the closest resemblance in methylation signature.

The analysis of the Fraction of Genome Altered (FGA) and mutation rates within the context of the epigenetic progenitor model of cancer yields intriguing insights regarding the influence of methylation on genomic stability. In the context above, mutation rate refers to small-scale changes affecting individual nucleotides (point mutations) and rearrangements of smaller DNA segments. According to this model, epigenetic alterations precede and facilitate genetic

mutations in oncogenesis. Consistent with this theory, the CIMP methylomic profile associated with the worst predicted survival outcomes exhibited the highest median FGA but the lowest overall mutation rate, pointing to large-scale structural changes that may induce genomic instability, setting the stage for further genetic aberrations in KIRP patients. Conversely, the intermediate-predicted survival group displayed a moderate FGA with a pronounced mutation rate, while the group with the near-normal methylomic signature exhibited the lowest FGA but a relatively high mutation rate. These observations underline a potential correlation between the development of KIRP and specific single-point mutations, whereas the prognosis post-cancer development appears more closely tied to large-scale genomic instabilities induced by aberrant methylation.

The divergence in the relationship between epigenetic modifications and gene expression across different KIRP patient groups suggests distinct modes of interaction. The large number of hypermethylated DMRs in the CIMP group negatively correlated with survival rates and did not coincide with differential gene expression. This observation indicates that alterations in the methylation and changes in gene expression independently modulate survival outcomes. Despite extensive methylation differences, both the CIMP and the near-normal groups show consistent gene expression patterns, with lower averages in the squared differences in log fold change (LFC) and similar standard deviations in the LFC of differentially expressed genes (DEGs), which further dispels the notion that heterogeneity in gene expression is the sole determinant for poor survival in the CIMP group.

In contrast, the intermediate survival group, characterized by a large number of DMRs but significantly isolated differentially expressed genes (DEGs), shows the greatest deviation in gene expression from normal tissue. This pronounced change in gene expression, coupled with

less extensive methylation than seen in the CIMP group, suggests that different molecular mechanisms might be at play, potentially involving a higher level of smaller-scale mutations directly affecting gene expression, crucial for the observed intermediate survival outcomes.

By refining k-means clustering, the Estimation Maximization (EM) k-means algorithm significantly enhances the categorization of KIRP patients, leading patients' methylomic signatures to reflect more accurately their survival outcome, as shown in the Kaplan-Meier survival curves (Figure 9). Drastic shifts in the survival rates often occur toward the end of the Kaplan-Meier curves due to factors like censoring and attrition bias. Censoring arises from incomplete patient information, such as patients leaving the study prematurely or the study concluding before an event occurs. Attrition bias affects results when the loss of participants is non-random and related to the likelihood of the event, introducing bias into the estimates. Based on the clinical features of patients, it is apparent that the ML algorithm improves the categorization of patients within the context of OS. Without the estimation maximization, k-means clustering will bias poor survival outcomes with sex differences. Although still biased toward grouping female patients, after optimization, the ratio of male to female patients moves back toward a more representative population (Figure 12). The EM K-means algorithm reflects more realistically the link between tumour stage and survival outcome based on the CpG methylation pattern. The cluster predicted to have the highest survival rate gained a larger ratio of T1 staged-diagnosed patients. In contrast, the cluster predicted to have the worst survival outcome gains a larger ratio of T3-stage tumour patients (Figure 13). This data also indicates changes in methylation patterns are correlated to different tumour stages. As such, the EM algorithm is mining CpG data that supports what we know to be generally true, namely that later stages of cancer are implicated in poor cancer diagnosis.

2.5.2 Links between TF motif methylation and survival outcomes in KIRP

In our investigation into the link between epigenetic signatures and TF motifs in KIRP patients categorized according to our machine learning algorithm, we identified a critical link between the methylation status of specific TF motifs—namely *CTCF, KLF1, KLF9, REST, SP2, SP3,* and *WT1*. These TFs play pivotal roles in cellular processes essential for maintaining genomic integrity and regulating cell proliferation, differentiation, and apoptosis. Although aberrant methylation patterns are apparent in the poor OS CIMP population, investigating specific disruptions in TF motif binding may give us insight into actionable changes occurring in gene expression.

The Sp family of transcription factors and Krüppel-like factors (KLFs), notably *KLF1* and *KLF6*, play a key role in gene transcription regulation, attributed to their structural features like activation domains and a zinc finger DNA-binding region [109], [110]. Although *Sp3* can act both as an activator and a repressor, depending on the context, *Sp3* is distinguished by an inhibitory domain. These factors are vital in controlling gene expression involved in critical cancer-related processes such as cell cycle, apoptosis, and angiogenesis. The dysregulation of Sp and KLF members, particularly affecting processes like the epithelial-mesenchymal transition (EMT), crucial for metastasis, underlines their impact on cancer behaviour and patient outcomes. Sp3's inhibitory function is especially significant in gene regulation, influencing tumour suppression and growth.

CTCF's role in KIRP extends beyond maintaining genomic stability; it is pivotal in transcription and chromatin organization, influencing tumour suppressor gene expression. Disruptions in *CTCF*, such as mutations or binding site silencing, can facilitate oncogenesis [111]. Additionally, the *WT1* transcription factor, known for its roles in development and

differentiation, has been associated with other forms of cancer [112], [113]. Although *WT1* was first identified as a tumour suppressor in Wilms' tumour, emerging evidence indicates that *WT1* acts as an oncogene in various solid tumours and hematological malignancies. As a transcription factor, *WT1* plays an important role in development, differentiation arrest, apoptosis, and proliferation.

When analyzed collectively, the identified GO terms and KEGG pathways appear to delineate a network of biological processes that could potentially exacerbate the malignancy of KIRP through various mechanisms. For instance, pathways related to 'intracellular signal transduction' and 'Wnt signalling pathway' have been highlighted in our study. Aberrations in these pathways can significantly influence tumour cell behaviour, promoting angiogenesis, immune evasion, resistance to apoptosis, and enhancing their metastatic potential [114]. Disruptions in Wnt signalling, through genetic mutations in the *Wnt* gene or its pathway components, can result in developmental anomalies and contribute to cancer's onset. It is particularly noteworthy as its aberrant activation is often implicated in cancer progression and metastasis [115]. Gene Ontology (GO) terms like 'cell adhesion' and 'cell-cell adhesion' are crucial in understanding how cancer cells. In addition, they are associated with cancer stem cells (CSCs). CSCs are integral to tumour structure and behaviour, heavily relying on adhesion processes for maintaining their niche, migration, and initiating metastasis. Abnormalities in cell adhesion mechanisms can enable CSCs to detach from the primary tumour, survive in circulation, and colonize new tissues, thereby facilitating metastatic spread. This capacity for metastasis, coupled with CSCs' ability to resist conventional therapies, often correlates with poorer prognosis and survival outcomes in patients [116].

The PI3K/AKT signalling pathway is pivotal in cancer biology, particularly in KIRP. This pathway is often hyperactive in cancers with *PTEN* or *PIK3CA* mutations, contributing to tumour growth and evading apoptosis [58], [114]. In the context of KIRP, the methylomic aberrations within the TF motifs may explain the dysregulation in the PI3K/AKT pathway, leading to apoptotic evasion and poor survival outcomes. Similarly, pathways like the 'Hippo signalling pathway are integral to cell fate determination and tissue homeostasis and have been correlated with the survival outcome of patients suffering from cancer [117]. Alterations in these pathways could lead to uncontrolled cell proliferation and a failure in the differentiation processes that normally suppress tumour formation.

This study explored the nuanced correlation between TF motif methylation and patient survival in KIRP, shedding light on the sophisticated epigenetic mechanisms in cancer progression and prognosis. The altered methylation profiles of critical TFs, including *CTCF, KLF1, KLF9, REST, SP2, SP3,* and *WT1*, reveal intricate regulatory disruptions that may compromise key cellular processes vital in oncogenesis, such as genomic stability and apoptosis. Notably, the Sp and KLF families emerge as pivotal entities in this regulatory landscape, offering potential therapeutic targets. The interplay between TF methylation and signalling pathways like PI3K/AKT and Hippo underscores a complex network influencing KIRP malignancy and suggests a strategic avenue for developing nuanced therapeutic interventions.

2.5.3 Trends in GO and KEGG enrichment for improved accuracy in poor OS Isolation

When applying k-means clustering enhanced by the Estimation Maximization (EM) algorithm to our selected probes, we observe more distinct segregation of survival outcomes

among patient clusters compared to using k-means alone (Figure 7). Investigating the change in the GADMR in the larger context of gene ontology and KEGG pathways gives us insight into what these patients share as a common denominator in either classification of high or low overall survival outcomes.

As mentioned in the previous section, aberrations in pathways such as “signal transduction” and its related process such as “intracellular signal transduction” can significantly influence tumour cell behaviour, promoting angiogenesis, immuno evasion, resistance to apoptosis, and enhancing their metastatic potential. Signal transduction pathways are crucial for transmitting signals from the cellular exterior to its interior, impacting various cellular responses [114]. Cell adhesion molecules (CAMs) and related pathways, such as "cell adhesion" and "cell-cell adhesion," are fundamental in maintaining normal tissue architecture and cellular communication. CAMs include diverse groups like integrins and cadherins, which are essential for tissue integrity and pathological processes. They interact with various receptors, integrating external signals with cellular responses. In cancer, changes in CAM expression or function can facilitate cancer cell detachment, invasion, and dissemination through the bloodstream to distant organs [116].

The concept of "stem cell division" underscores the critical role of stem cell-like characteristics within the realm of oncology, particularly highlighting how cancer stem cells (CSCs) influence resistance to therapy and contribute to the diversity seen within tumour structures [31]. These CSCs, like normal stem cells, are capable of extensive self-renewal and differentiation, supporting the tumour's growth and complexity. This mechanism is particularly relevant in understanding how the differentiation processes, such as those regulating chondrocyte maturation, intersect with cancer pathology. Pathways regulating cell differentiation, like

chondrocyte differentiation, can also impact tumour progression and metastasis, particularly in cancers affecting bone or exhibiting osteomimicry.

The "BMP signalling pathway" stands out as a crucial system involved in cellular growth and differentiation processes, with Bone Morphogenetic Proteins (BMPs) serving as pivotal signalling entities within this pathway. These proteins engage in extensive intercellular communication, binding to specialized receptors on their target cells and initiating a cascade of intracellular events that govern cell fate. In tumours, aberrant BMP signalling is linked to enhanced tumour growth and the disruption of normal cellular regulation mechanisms. The pathway's influence extends to various cancer-related processes, including the stem cell-like behaviour of CSCs, underpinning their role in tumour heterogeneity and therapy resistance [118]. Transforming growth factor beta (TGF-ß), a versatile polypeptide, regulates numerous cellular processes, including cell proliferation, differentiation, embryonic development, angiogenesis, and wound healing across different cell types [119]. In advanced cancers, the "TGF-ß receptor signalling pathway" promotes tumour progression and metastasis in later stages. The "cAMP signalling pathway" and "Ras signalling pathway" in the context of cancer provide insights into cell fate decisions and survival. These pathways are critical in mediating cellular responses to external stimuli, and their dysregulation can lead to aberrant proliferation, survival, and metastasis [120], [121].

In the context of finding the similarity of patients with enhancing the overall survival outcomes, understanding the isolated GO terms and KEGG pathways in our best survival outcome group can guide research into therapeutic interventions. The Toll-like receptor (TLR) signalling pathways, such as the "Toll-like receptor 3 signalling pathway," play pivotal roles in the innate immune response. TLRs are critical in detecting pathogen-associated molecular

patterns (PAMPs) and initiating immune responses. Activation of TLR3 pathways leads to the production of type I interferons and other pro-inflammatory cytokines, enhancing the immune system's ability to eliminate pathogens and potentially malignant cells [122], [123]. In the context of cancer, activation of these pathways can enhance the immune surveillance against tumour cells, promoting tumour regression and improving patient survival. Furthermore, these pathways can stimulate the adaptive immune response, providing a long-term immunological memory against tumour antigens.

In the context of KIRP, an important relationship has been observed between *IL-6* methylation levels and patient survival, where elevated IL-6 serum levels are linked to poorer outcomes [124], [125]. IL-6 plays a dual role, aiding in immune defense but also cancer progression when overexpressed. For KIRP patients, the regulation of *IL-6* through methylation is particularly significant. Genes associated with differentially methylated regions (GADMR) that enhance IL-6 production are pivotal since their methylation likely results in reduced IL-6 output. This decrease is beneficial because high *IL-6* levels are associated with advanced disease and lower survival rates. Methylation reduces *IL-6* levels, which likely dampens inflammation that supports tumour growth and enhances the effectiveness of the immune response against the tumour, thereby improving patient survival.

Increased levels of macrophages in the tumour microenvironment of renal cell carcinoma (RCC) are generally associated negatively with patient survival. The research indicates that macrophages within the RCC microenvironment predominantly adopt a tumour-promoting role [126], [127]. They are implicated in various processes that facilitate cancer progression, such as promoting angiogenesis, inhibiting effective immune responses, enhancing tumour cell proliferation, invasion, metastasis, and contributing to the formation of tumour stem cells and

drug resistance. Within the context of our own KIRP research, improved categorization of the best survival outcome group highlights differences in GADMR that are enriched for GO terms such as "macrophage activation involved in immune response," "positive regulation of inflammatory response," and "positive regulation of macrophage activation" are pivotal since an increased survival outcome is associated with a methylation signature that reduces macrophage levels. This methylation-driven alteration in gene expression likely modulates macrophage function and density in the tumour microenvironment, skewing them away from a pro-tumoural activity and potentially restoring their normal immunosurveillance and anti-tumour roles, offering a promising avenue for therapeutic intervention in KIRP.

In conclusion, our research offers a nuanced understanding of the correlations between GO terms, KEGG pathway enrichments, and survival outcomes in kidney renal papillary cell carcinoma KIRP, highlighting the molecular intricacies that dictate patient prognosis. The integration of k-means EM-determined probes has yielded a refined segregation of patient clusters, illuminating the biological commonalities that underpin distinct survival rates. Key pathways involving cell adhesion, signal transduction, and BMP signalling emerge as critical determinants of tumour behaviour, influencing cell detachment, invasion, and differentiation—factors central to cancer progression and metastasis. Our findings underscore the importance of immune-related pathways, particularly the TLR and IL-6 signalling pathways, in modulating the immune response and its impact on tumour progression and patient survival. The differential methylation patterns observed, particularly in genes affecting macrophage activation and IL-6 production, reveal potential epigenetic targets for therapeutic intervention. These methylation patterns are pivotal, highlighting the underlying epigenetic mechanisms that may influence these critical pathways and the clinical outcomes in KIRP.

2.6. Predictive capability of hybrid k-means EM machine learning

In our research, we devised a predictive machine learning (ML) algorithm distinct from our investigative algorithm. One key difference was the choice to exclude the predictive algorithm from a final reduction and reintegration of CpG probes within a k-means EM cycle during the algorithm's final phase. This decision was intentional. Despite the potential of such a cycle to enhance patient grouping within the training data, it also risked augmenting generalized error. In simpler terms, it threatened to reduce the performance of the testing data. The primary goal was establishing a model that would perform equivalently on the training and testing data.

An illustration of the model can be seen in the provided flowchart (Figure 6). Interestingly, the F1 score within our worst overall survival group showed improvement in the testing data, rising from 0.611 during training to 0.800 during testing. Though a phenomenon where we observe a better test error than train error may seem unusual, it can occur given the right conditions and the stochastic nature of small sampling. The right conditions would include model design and hyperparameter selection aimed at minimizing generalized error. The framework of my ML model has strong regularization as phases 2 and 3 iteratively refine the probes in the model to reduce noise and reduce the model's input complexity on the training data. At the same time, phase 4 of the modelling workflow guards against overfitting by combining all the CpG probes from previous iterations to generate an extensive list of probes used in k-means clustering.

Sampling variability within a small sample size of 266 between training and testing sets could also play a role. The training set, comprising 90% of the samples, may include noisier examples or challenging cases due to the stochastic nature of the EM process. In contrast, the

smaller test set may have been less noisy and contain fewer outliers or better aligned with the model's learned parameters, which could explain the higher F1 score. This is particularly plausible given that the phenomenon is limited to a specific group and not the entire dataset.

Initially, the first cluster displayed the best-predicted survival rate. However, this group diverged from its expected trend after 85 months, at which point only three patients remained under observation (Figure 20). When analyzing Kaplan-Meier survival curves, deviations in survival probabilities often occur due to factors like censoring. Censoring is another aspect affecting the late-stage trend deviation. Censoring happens when there's a lack of complete patient information, such as when patients leave the study prematurely or if the study concludes before an event occurs.

Attrition bias is another aspect affecting the late-stage trend deviation when analyzing Kaplan-Meier survival curves. If attrition is not random and is tied to the event likelihood, bias can be introduced into the results. The assumption of a homogeneous population inherent in the KM estimator can impact the curve's final part. We often notice a deviation from the original trend or a steep decline towards the end when few patients remain under study. Several factors contribute to this phenomenon. A significant element is the weight each event—like death or relapse—has on the overall survival probability. With fewer patients left in the study, each event has a larger proportional effect on the estimated survival probability. This is a consequence of the smaller denominator in the survival probability calculation, meaning that even a single event can lead to a noticeable decrease in the curve.

Throughout the study, those who remain are typically those who have been censored, which can distort the survival function estimation towards the study's end. Attrition bias is yet

another aspect affecting the late-stage trend deviation. If attrition is not random and is tied to the event likelihood, bias can be introduced into the results. Furthermore, the assumption of a homogeneous population inherent in the KM estimator can impact the curve's final part. The estimator assumes censored individuals share the same survival prospects as those who continue to be observed. However, if the remaining subjects towards the end of the study aren't representative of the initial group—perhaps they're unusually resilient or have received superior care—this could lead to an abrupt change in the curve. Some researchers may prefer to report median survival times, as these figures tend to be less influenced by the number of patients under observation towards the end of the study.

2.7 Concluding Statement

One of the major challenges in analyzing genomic data is the vast amount of information that needs to be sifted through to make sense of it all. Our study begins with over 450,000 CpG probe sites across multiple patients, each with various characteristics. Statistical analysis is a common approach in epigenomic research, allowing for comparing methylation levels at individual CpG sites or regions between different groups. However, this method can be limited by the need for multiple comparison corrections and potential confounding factors like age, sex, and batch effects. K-means clustering algorithms can also have drawbacks, including sensitivity to outliers and unstable results. To address these challenges, we use an estimation maximization approach to k-means clustering, which helps to reduce noise and yield a smaller set of genes for investigation.

This study was not without its limitations, including difficulty in categorizing the data from patients who contributed both a normal and a primary tumour sample. This led to the potential of these patients to be clustered into two groups but only represented at one point in the

KM survival curve. As such, we have 45 patients with a normal sample and a tumour sample appearing in two different groups throughout the analysis but only represented by their tumour sample in the KM-survival curve. This assumes that a methylation pattern in normal tissue samples will also be seen in the tumour samples for patients with better survival outcomes. Another risk in using a layer or staged machine learning is the risk of over-correction, which is why there are small discrepancies between our investigative and predictive models. The categorization of optimized probes will correspond more heavily to our training data and will likely lose some effectiveness when implemented on new samples or different data sets.

The stark difference in survival outcomes associated with different DNA methylation profiles between tumour groups and normal samples indicates tumours can be classified based on DNA methylation signatures and that the malignancy of some subset of tumours might depend heavily on maintaining their methylation states. Enrichment analysis of differentially methylated promoter sites in poor outcome groups from tumour samples identified some methylation signatures correlated with biological processes that control RNA and DNA transcription, a key regulatory point for tumour metastasis.

Although most age-related DNA methylation patterns are specific to certain tissues, certain genes exhibit conserved DNA methylation patterns across different tissues [128]. Building on our work, future studies can use a similar method on cfcDNA to identify KIRP DNA methylation signatures related to different survival outcomes. This research could assist with the predisposition and diagnostic cancer screenings and narrow the focus to the most critical biological pathways. Despite the limitations, our technique shows the potential of using machine learning algorithms with clustering analysis to identify influential DNA methylation sites linked to survival outcomes in kidney tumours. With further refinement and validation, this approach

may have important implications for developing personalized therapies and improving clinical outcomes for kidney cancer patients.

CHAPTER 3: Optimizing AD Patient Classification and Detection: Modifying the EM-Enhanced K-Means Clustering on Available DNA Methylation Signatures

3.1 Publication Stage and Co-author Statement

This chapter builds on the research from the previous chapter and applies k-means and Expectation Maximization Approaches to reclassify and enhance the analysis of Alzheimer's disease patients. The manuscript, tentatively titled "Application of k-means and Expectation Maximization Approaches with Methylomic Data to Improve Analysis of Alzheimer's Disease Patients," is in the early stages of preparation, as much of the methodology has been integrated into the manuscript from Chapter 2. I, Gatonguay Siu, am the primary author responsible for the study's conception, data analysis, coding, and manuscript preparation. Dr. Touati Benoukraf provided guidance on study design.

3.2 Objective and Machine Learning Implementation

3.2.1 Objectives

We hypothesize in this section that we can modify the EM k-means clustering algorithm from Chapter 2 to significantly improve patients' classification by identifying DNA methylation signatures in Alzheimer's disease. This approach will allow further development into personalized treatment strategies. To this end:

- The first objective is to modify the EM K-means algorithm from Chapter 2 and to test its ability to reduce noise and classify patients with distinct methylation patterns based on the Brakk stage.

- Evaluate the performance of the modified algorithm in identifying patients and their biomarkers with distinct methylation patterns, clinical features, genomic features, gene isolation, and GO enrichment analysis in the context of Alzheimer's disease.

3.2.2 Implementation of estimation maximization

The data initially categorizes patients based on the Braak stage. Braak stages are based on the brain's extent and distribution of neurofibrillary tangles (NFTs). Furthermore, Braak staging is typically done post-mortem and thus is not used for clinical diagnosis or classification of living patients. The Expectation-Maximization (EM) algorithm discussed in this chapter closely resembles the one presented in the previous chapter. However, the only notable change is that phase 4 is the product of five iterations rather than ten and changes to the definition of fitness to account for the differences in diagnosing Alzheimer's compared to cancer (Figure 6).

3.2.3 Determining Optimal Clustering in Methylation Analysis

In our analysis, we utilized hierarchical clustering, grouping samples based on similarity in their methylation profiles. To maintain accuracy, we excluded any probe sites with absent M-values measuring methylation levels in at least one sample. We employed the within-cluster sum of squares (WSS) method to determine the optimal number of clusters by minimizing the variation within each group. We identified three as the ideal number of clusters using the "elbow" method, which locates the point where increasing the cluster count doesn't substantially lower the WSS (Figure 22).

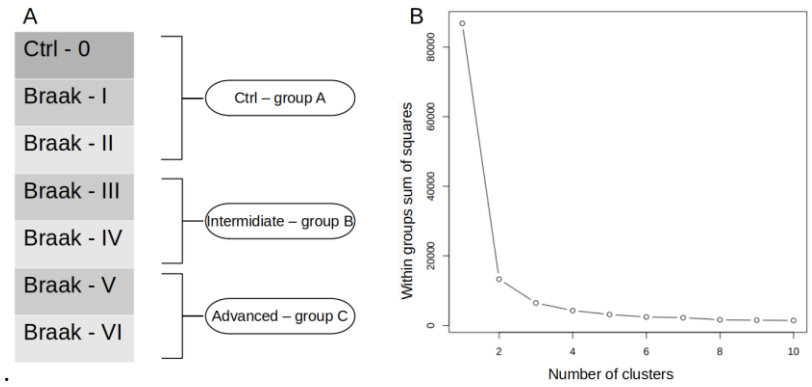


Figure 22. Analysis of Methylation Profiles and Optimal Cluster Determination. (A) Table outlining the grouping of samples based on Braak stages, showing Braak stages 0, I, and II combined into one group, Braak stages III and IV as another group, and Braak stages V and VI as the final group. (B) Elbow graph illustrating the within-cluster sum of squares (WSS) against the number of clusters. The graph identifies the optimal number of clusters as three, indicated by the point where further increases in the number of clusters do not significantly reduce the WSS.

3.2.4 The formulation for the scoring of the EM Algorithm

The patient samples were categorized into three groups according to their Braak stage. Group A contained samples with no Alzheimer's disease symptoms or only early-stage signs, specifically Braak stages I and II, comprising 23 patient samples. Group B included patients at Braak stages III and IV, totalling 16 samples. Group C encompassed patient samples at Braak stages V and VI, amounting to 56 samples (Figure 22).

The scoring metric was established based on the clustering accuracy for each group, with larger groups allowing for greater leniency in sample miscategorization. We employed the following formula to compute a score based on the scoring of each group (eq. 7):

$$A_{score} = \sum R - (T - R) / G \qquad (7)$$

In this equation, A_{score} represents the sum of scores for all groups. T denotes the total number of patients in each group, R signifies the number of correctly categorized samples, and G corresponds to the initial number of samples assigned to each group.

3.1 Material and Methods

3.3.1 Data Acquisition and Selection Criteria

We manually searched the published literature through the Web of Science to identify relevant data for this study. We conducted NCBI keyword searches using terms such as "brain + methylation" and "neur* + Illumina." We downloaded data from Smith Adam et al. (2019) in a study that utilized oxidative bisulphite conversion with the Illumina Infinium Human

Methylation 450K microarray to identify neuropathology-associated differential DNA methylation and DNA hydroxymethylation in the entorhinal cortex. Data can be found using the GEO accession number GSE105109 [10].

The GSE105109 library, downloaded in Dec 2022, contains IDAT files for the Illumina 450k CpG methylation data from the entihornal cortex. For the Illumina 450K profiling, both BS and OxBS methods were used from brain samples in the entorhinal cortex (EC) collected from 96 individuals stored in the MRC London Neurodegenerative Disease Brain Bank.

3.3.2 Pre-processing IDAT to beta values

In this study, we employed the Sesame R package for preprocessing and analyzing Illumina Infinium DNA methylation microarray data derived from IDAT files. Sesame is specifically tailored for Illumina's 450K and EPIC arrays, facilitating the transformation of raw IDAT files into beta values that signify methylation levels at distinct CpG sites [103], [103], [129], [130].

The Sesame package was sourced from Bioconductor and subsequently loaded into the R environment. After installing and activating the Sesame package, raw IDAT files were imported into R using the readIDATpair function. Upon completion of the preprocessing, beta values for individual CpG sites were calculated using the getBeta function. The resulting beta_values matrix encompasses methylation levels, from 0 (unmethylated) to 1 (fully methylated), for each CpG site across all samples incorporated in the analysis [104], [105]. These beta values were subsequently employed for further analyses, including differential methylation analysis and data visualization. The Beta values were converted to M-values (eq. 4), making them more appropriate for sequencing-based studies. M-values have a more symmetric distribution and

perform better in Detection Rate (DR) and True Positive Rate (TPR) for highly methylated and unmethylated CpG sites.

3.3.3 Heatmap Generation and Differential Methylation Analysis

We thoroughly analyzed all DMR, taking a comprehensive approach rather than limiting our focus to the ML-determined biomarkers for patient grouping. When building the volcano plots, we utilized all CpG probes listed in the HM450k manifest file [36], which did not contain any NA values across the samples. The heatmap generation refined the selection of CpG probes for further analysis; we calculated and compared the average M-value scores across each identified cluster. We classified a CpG probe as differentially methylated if the difference in its average M-value between the pseudo-control group and the pseudo-advanced group was greater than or equal to ±2Z, accompanied by an adjusted p-value of less than 0.01. This statistical significance was determined using the Kolmogorov-Smirnov Test with Bonferroni-Hochberg correction for multiple comparisons. The same change in M-value was applied to identify differentially methylated regions (DMRs) within the pseudo-intermediate group. Following this, we further restricted the list of probes to those labelled as "Promoter*" in the "Regulatory_Feature_Group" column found in the hm450k manifest file [36] to reduce computational RAM to under 32Gb when building the heatmaps.

3.3.4 Isolated Genes and Enrichment Analysis

After isolating the gene associated with DMR data of patient clusters using machine learning, we obtained the gene symbols for all DMR within the hm450k manifest file. Differentially methylated probes had the same criteria as when building the heatmaps, with one significant change, we used all CpG sites. All related genes were included in the enrichment

analysis when a probe was associated with multiple genes. We utilized gene ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses on *David's platform* [106] to investigate the GADMR of different AD groups based on machine learning-derived cluster groupings of disease severity.

As a frame of reference, we curated a list of all known genes associated with Alzheimer's disease from NCBI, using the keyword search "Alzheimer's." We downloaded all genes in a table format, then filtered them for only the Alzheimer's genes associated with *Homo sapiens*. This same list created a reference point for our enrichment analysis.

The gene expression data of patients were imported and subjected to enrichment analysis. This enrichment analysis aimed to establish if the GADMR were in agreement with previous findings on genes and enriched terms associated with AD. This also gave rise to the potential to find new insights into the Gene Ontologies and KEGG pathways associated with late-stage Alzheimer's disease. To ensure robust results, a stringent Bonferroni FDR cutoff of less than 0.01 was applied in the analysis to evaluate our machine learning-based grouping method with known Alzheimer's GO and KEGG enriched terms.

3.3.5 Age and DNA Methylation Age Analysis of Alzheimer's Groupings

Utilizing the model devised by Steve Horvath [131], we incorporated our beta values for each patient. We made minor adjustments to the tutorial files to calculate the DNA methylation (DNAm) age. Out of the 353 CpG probes recommended by Horvath, only 300 were accessible for analysis. Consequently, a constant was incorporated into the DNAm age calculation to account for this discrepancy. The following equation was used to determine the constant added:

$$C = DNAm\_age_{avg} - Age_{avg} \quad (10)$$

3.4 Results

We have modified the EM k-means machine learning algorithm from Chapter 2 to improve the classification of Alzheimer's disease patients. The EM k-means algorithm will ideally reduce the heterogeneity in our sample analysis to understand AD's epigenomic landscape better. To test the EM K-means algorithm, we benchmark it against the methylation signature of raw Braak classification of Alzheimer's disease.

3.4.1 DNA CpG methylation sites in Alzheimer's Disease Patients

In this study, Alzheimer's disease (AD) patients were initially categorized based on their Braak stages: Group A with stages 0-II, Group B with stages III-IV, and Group C with stages V-VI. The objective was to create a baseline for further comparison in identifying significantly differentially methylated CpG sites among these groups. However, when using the Braak stage alone and applying the Kolmogorov-Smirnov test followed by Benjamini & Hochberg calculations, no CpG site could be isolated with a p-adjusted value lower than 0.01. Consequently, no associated heatmap for significantly differentially methylated probes was generated, and no further analysis could be conducted with significance.

Our study continued by categorizing AD patient samples into three distinct DNA methylation subgroups using k-means clustering. The input CpG probes for this analysis were refined using the estimation-maximum (EM) algorithm. A gradient descent approach favoured CpG probes that classified patients' samples according to their corresponding Braak stages grouping. The pseudo-advanced Braak stage subgroup displayed a unique CpG methylation

signature characterized by distinct hypermethylated and hypomethylated regions, with fewer intermediate methylated sections than the other groups (Figure 23.A). In contrast, the best and intermediate outcome groups exhibited shared patterns across various CpG site sections but remained distinct due to their unique combinations and methylation levels in other sections. Moreover, the pseudo-intermediate Braak stage subgroup showed more hypomethylated CpG sections than the control/low-level AD group, emphasizing the differences in methylation patterns between these subgroups.

In the volcano plot presented below (Figure 23.B-C), the distribution of EM-determined biomarkers is overlaid across all CpG probes. The pairwise p-values were calculated using the Kolmogorov-Smirnov Test to compare the distributions between the pseudo-intermediate and the global average and between the pseudo-advanced and the global average. The respective p-values are 0.1209 and 0.4213. Consequently, we cannot reject the null hypothesis. This indicates insufficient statistical evidence to assert that the distributions of these samples significantly differ. Compared to the overall average M-values, the failure to achieve statistical significance in the biomarker distributions across other groups suggests that differences in Alzheimer's disease severity may not be directly linked to these biomarkers. Instead, these differences might be associated with other differentially methylated regions (DMRs). The utility of biomarkers may be limited to categorizing patients into specific groups.

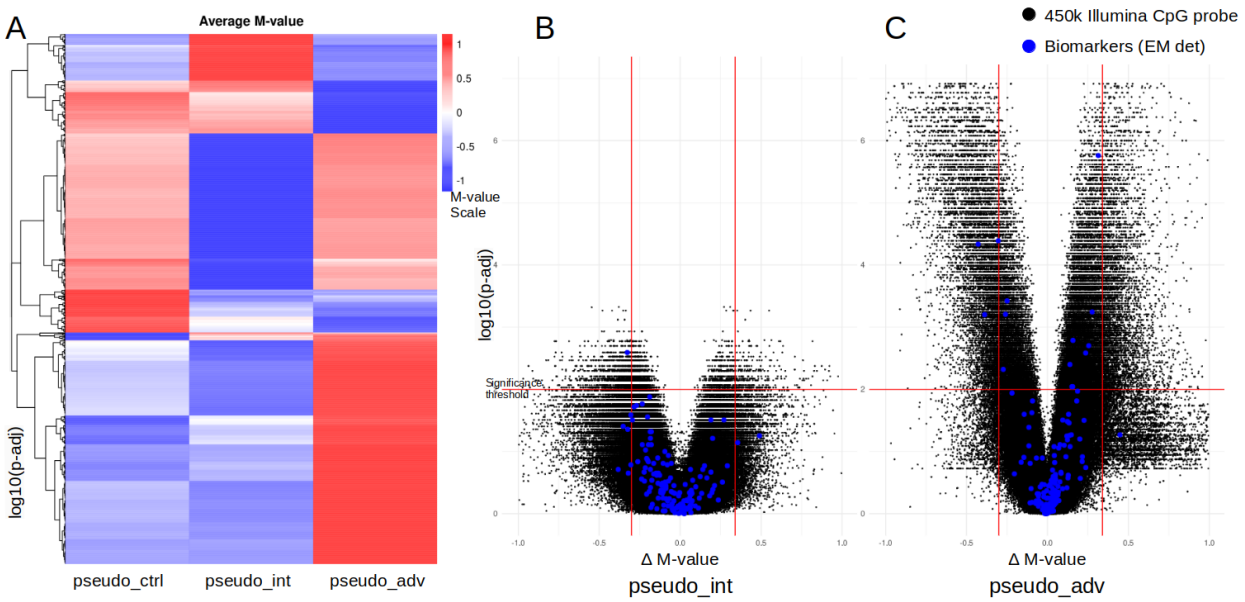


Figure 23. A) Heatmap illustrates the distinct methylation patterns in the ML-classified AD patient subgroups based on average M values for DMR within promoter regions. The left heatmap shows the pseudo-control group, the center represents the pseudo-intermediate patient grouping, and the right represents the pseudo-advanced patient grouping. Volcano plots illustrate each cluster's average M-value (ΔM-val) changes relative to the pseudo_control values. The y-axis represents the significance of the changes, with a threshold set at p = 0.01. The x-axis shows the ΔM-val; negative values represent hypomethylation, whereas positive values represent hypermethylation. A standardized ±2Z threshold marks significant deviation samples using the vertical lines. Plot B) displays the pseudo_intermediate group, while Plot C) shows the pseudo_advanced group.

3.4.2 Alzheimer's Disease GADMR Enrichment Analysis

We examined the genes associated with differentially methylated regions (GADMR) that passed the adjusted p-value significance test in each cluster, representing various Alzheimer's severity levels. The pseudo-intermediate Alzheimer's disease sample cluster (cl2) contained 2836 significantly isolated genes with an adjusted p-value < 0.01 and an average M-value difference > ±2Z when using the pseudo-control group as a reference for methylation levels. The pseudo-advanced Alzheimer's disease sample cluster (cl3) had 8737 significantly isolated genes using the pseudo-control group as a reference for methylation levels. To assess the relevance of these isolated genes, we compared them to known Alzheimer's disease-associated genes. We discovered that cl2's gene list contained 247 1554 known AD genes, while cl3's gene list included 738 (Figure 24).

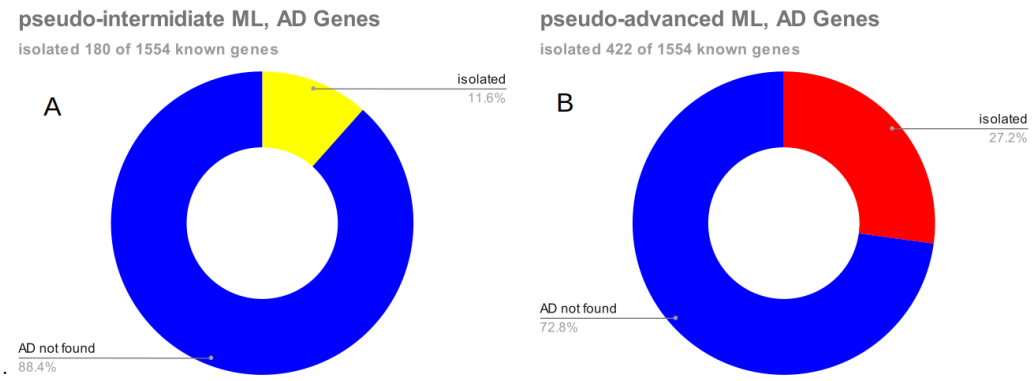


Figure 24. The Pie chart represents the relative number of genes associated with DMR that overlap with known AD genes found using either the k-means EM approach for each of our two severity groups.

3.4.3 Enrichment analysis of differentially methylated genes

To delve deeper into our investigation, we conducted an enrichment analysis on all GADMR. We first analyzed the *Homo sapiens* NCBI gene list for Alzheimer's disease (AD) to establish a baseline of Gene Ontology (GO) and KEGG enrichment terms related to AD. This analysis produced 916 GO terms (Figure 25). For the pseudo-intermediate group (cl2), 32 GO terms were associated with its 2015 GADMR, while the pseudo-advanced group (cl3) had 232 GO terms for its 4770 GADMR. We then examined the GO terms overlapping the ML-determined GADMR and the AD GO terms. We identified 17 overlapping GO terms between the pseudo-intermediate AD group and the general AD GO terms and 145 overlapping GO terms between our pseudo-advanced AD group and the known AD GO term group (Figure 25. A-B). In addition, new GO and KEGG terms were isolated in our pseudo_advanced group. Comparing the cl2_AD overlap with the cl3_AD overlap revealed 13 shared GO terms (Figure 25. C).

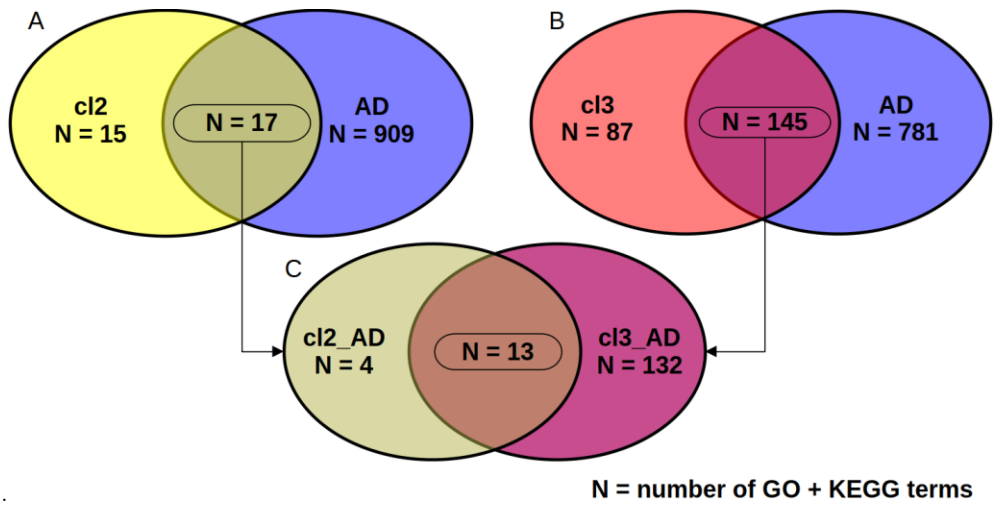


Figure 25. Venn diagram of GO enrichment analysis. (A-B) Overlapping GO terms between the ML-classified pseudo-intermediate (cl2) and pseudo-advanced (cl3) AD clusters and the known AD GO terms. (C) Comparison of the overlapping GO term from (A) cl2_AD and (B) cl3_AD.

3.4.4 Model evaluation based on age and DNAm age analysis

In this study, we evaluated clinically categorized AD patients and compared them with our machine learning (ML)-classified counterparts to assess our ML algorithm. Our findings are summarized in box plots (Figure 26). When categorizing patients by chronological age. We observed a larger spread than when comparing their DNAm age. Interestingly, when classifying patients based on their Braak stage grouping, the intermediate group had a higher average age than the advanced AD patient group in both DNAm and chronological age. Another notable feature of the data was the inconsistent mean relative error (MRE) between the chronological age and DNAm age across all three groups: the control/low AD group had an MRE of 9.6, the intermediate AD group had an MRE of 5.34, and the advanced AD group had an MRE of 7.17. In our ML-classified AD patient groups, we found more consistency. The chronological and DNAm ages of the clusters showed a clear increase with our pseudo-advancement of AD staging risk. Additionally, the MRE for all clusters remained within a closer range: our pseudo-control group (cl1) had an MRE of 7.79, our pseudo-intermediate AD group (cl2) had an MRE of 7.39, and our pseudo-advanced AD group (cl3) had an MRE of 7.23.

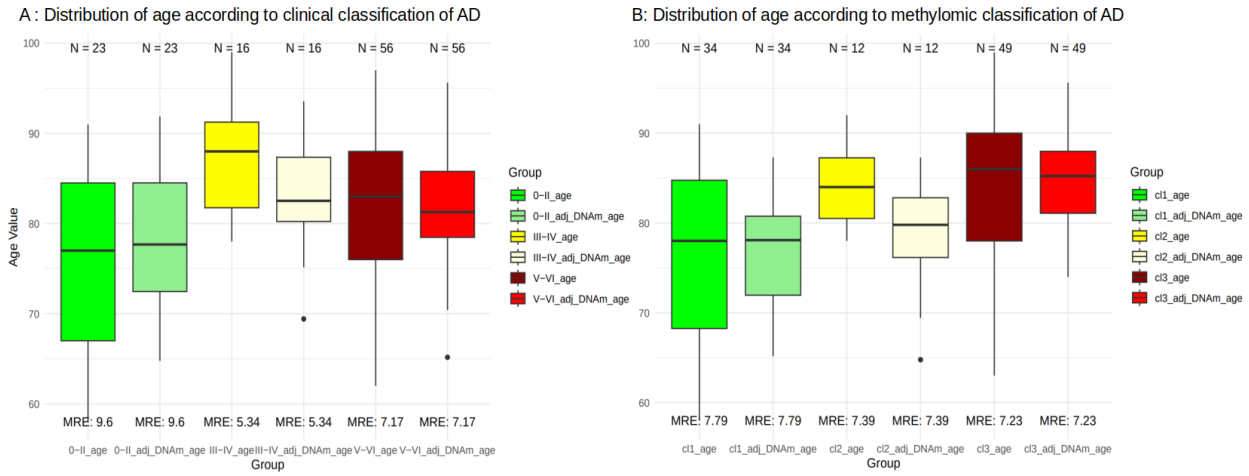


Figure 26. Comparison of clinically categorized and ML-classified AD patients based on chronological and DNAm age. Age value is found on the Y-axis. Each group is represented twice, once with chronological age (age) and once with their adjusted DNAm age. The box's width represents the IQR and indicates the spread of the middle 50% of the data; the mean relative error is displayed below. Skewness: The whiskers extending from the box indicate the range of the data outside the IQR but within 1.5 times the IQR. Outliers are represented as individual points on the plot and found beyond 1.5 times the IQR.

3.5 Discussion

3.5.1: Validating ML Algorithm: Gene Association, Enrichment, and Age Analysis

In this study, we sought to validate our machine learning (ML) algorithm by analyzing genes associated with differentially methylated regions (GADMR) in Alzheimer's disease (AD) patients. Initially, patients were categorized by Braak stages, but this approach failed to isolate any DMR at a significance threshold of < 0.01 p-adjusted. This study employed k-means clustering and the estimation-maximum (EM) algorithm, categorizing AD patient samples into three DNA methylation subgroups corresponding to their methylation signatures and approximate Braak stages.

Our findings revealed that 2015 GADMR were significantly isolated in the pseudo-intermediate AD group (cl2) and that 4770 genes were in the pseudo-advanced AD group (cl3). We compared these genes to known AD-associated genes and found that 11.6% (180) of pseudo_intermediate and 27.2% (422) of pseudo_advanced GADMR matched. Our results support that there is a high level of differential methylation that is associated within the entorhinal cortex that is associated with the progression of AD. Additionally, this finding may add to our understanding of AD as the DMRs that pass significance have implications on gene expression that are yet to be fully understood and, as such, may offer novel insights into AD by uncovering previously unexplored genes and CpG sites related to the disease.

The enrichment analysis underscores the potential mechanistic insights that GADMR might offer in AD pathology, particularly in pseudo-advanced AD groups. The significant overlap between the GADMR-derived GO and KEGG terms with those already established in AD research validates the relevance of these regions in the disease context. Notably, identifying

87 new GO and KEGG terms in advanced AD patients may be associated with AD through differential methylation and suggests novel pathways that may be implicated in disease progression or severity. The shared GO terms across different severity groups highlight common molecular pathways affected in AD, reinforcing the idea of a continuum in molecular changes as the disease progresses. The differential GO terms specific to each group could reflect distinct biological processes driving the disease at different stages, which could be crucial for staging AD or tailoring treatments to specific disease phases.

Adding to our evaluation of our ML algorithm, our age analysis results emphasize the ML algorithm's validity in categorizing AD patients. Comparing genetic and chronological age in clinically categorized and ML-classified patient groups, we noted the ML-classified patients showed a general trend towards increased chronological age and DNAm age based on AD risk/progression, thereby validating the ML algorithm’s classification as it is supported by what is known in age being associated with AD risk and progression. In addition, we also observed an increase in consistency of chronological and DNAm age in our ML-classified groups. The mean relative error values in these groups showed enhanced consistency, further validating the ML algorithm's effectiveness in categorizing AD patients and its potential for improving patient stratification. By identifying novel genes and CpG sites involved in the disease, the ML algorithm contributes to a deeper understanding of AD's epigenetic landscape. It enables a more accurate assessment of risk and progression. This highlights the potential of ML methods in early Alzheimer's disease diagnosis.

3.5.2 Concluding Statement

In this study, we aimed to validate a machine learning (ML) algorithm designed to categorize Alzheimer's disease (AD) patients based on DNA methylation profiles. By employing k-means clustering and the expectation-maximization algorithm, we effectively differentiated patients into subgroups corresponding to methylation signature similarity and approximate AD neuropathological classification. Our analysis identified a significant number of genes associated with differentially methylated regions in the pseudo-intermediate and pseudo-advanced stages. This validation underscores the ML algorithm's capability to detect epigenetic signatures that reflect the progression of AD.

In our study on Alzheimer's disease, we encountered several challenges akin to those previously noted in our work on kidney cancer. A primary difficulty in Alzheimer's research is the disease's heterogeneity. Patients exhibit various pathological features, including amyloid-beta plaques, neurofibrillary tangles, and neuroinflammation, complicating the task of identifying universally applicable biomarkers or molecular signatures. This heterogeneity renders categorization based solely on clinical features problematic for comprehensive analysis. To mitigate these issues, we applied k-means clustering and the expectation-maximization algorithm to stratify patients into subgroups based on their epigenetic profiles. This hybrid method could be considered by future researchers grappling with similar challenges. Moreover, the limitations stemming from the data quality and our sample size warrant consideration. Our dependence on existing datasets likely introduced biases related to sample representativeness and the completeness of methylation profiles. An increase in sample size would likely bolster the statistical power of our analyses and facilitate a more robust validation of our machine-learning model.

CHAPTER 4: Elucidating the Dynamics of YY1 Behavioural Shifts Amidst TAF1 Co-Binding Interactions

4.1 Publication Stage and Co-author Statement

This chapter forms the basis of the manuscript "Elucidating the Dynamics of YY1 Specificity Amidst TAF1 Co-Binding Interactions " prepared for submission to Frontiers; it has not yet been submitted. I, Gatonguay Siu, am the primary author responsible for the study's conception, data analysis, and manuscript preparation. Dr. Touati Benoukraf provided guidance on study design, and Dr. Thomas J. Belbin contributed to data interpretation and manuscript revisions.

4.2 Objective of YY1 and TAF1 analysis

We hypothesize in this section that the direct interaction between the transcription factors Ying-Yang1 (YY1, specifically the subset bound near the promoter region) and TATA Binding-Factor1 (TAF1) critically modulates and amplifies the initiation of gene transcription. This interaction may affect gene expression across diverse biological processes and molecular functions. Building on existing knowledge about YY1's role in gene activation through interactions with other transcription factors, we propose a deeper exploration into the collaborative dynamics of YY1 and TAF1. To this end:

- The first objective is to employ TFregulomeR and additional analytical techniques to analyze variations in the DNA-binding sites of YY1 and TAF1 across diverse cell lines in the context of tag fold change and methylation.
- Subsequently, we aim to analyze gene expression and ontology to pinpoint the specific biological processes notably influenced by the simultaneous expression of YY1 and TAF1.
- Should our findings provide evidence of a significant impact on gene expression due to YY1-TAF1 co-binding, our next step will be to propose a model that explains the formation and functional implications of a complex arising from the collaborative expression of YY1 and TAF1.

4.3 Methods

4.3.1 Methylation Analysis in TF Co-binding Using TFregulomeR

We used the tool TFregulomeR to identify a subset of TFs that show changes in methylation status based on their co-binding activities [132]. This tool provides information on the context in/dependent peaks between TF pairs. We established specific heuristic criteria for our initial data collection to guide subsequent analyses for all the DNA methylation matrices using the DNA methylation and read enrichment scores reports. Among the 414 human cell lines available in TFregulomeR, each cell line required ChIP-seq data for at least three different TFs to be analyzed further. For those cell lines with extensive ChIP-seq data, we limited our comparison to each TF's top five co-binding partners.

We investigated changes in the methylation status of transcription factors by comparing three different scenarios: one where both transcription factors (TFs) were present together

(referred to as the TF pair) and another where only one of the TF pairs was present, excluding the peaks associated with its partner (referred to as exclusive TF peaks). This comparison allowed us to assess how the methylation status differed when the TFs operated independently, as opposed to when they functioned together. Using TFregulomeR, we also obtained DNA methylation data that provided a count of methylated peaks for each base pair (bp) in the DNA sequences where TFs were bound. We excluded instances where neither the TF pair nor the exclusive TF showed more than 25 methylated peaks at the most highly methylated base pair. We also applied a heuristic rule: if a dataset had fewer than 10 total methylated peaks or more than 95% of the peaks exhibited low methylation levels (less than 10% methylation), that dataset was deemed "unmethylated" for subsequent comparisons.

The TF pair and the exclusive TF data were then organized into matrices. Each matrix had three rows, representing low (less than 10%), intermediate (10%-90%), and high (more than 90%) methylation levels. Each column in these matrices corresponded to a base pair within the binding motif. We first normalized the peak counts for each to compare the two scenarios. The absolute difference was calculated between corresponding elements in the two matrices, squared these differences, and summed the result. This approach gave us a pseudo-sum-of-squares difference, allowing for a meaningful comparison between the two conditions. Subsequently, we gathered statistics on these conditional comparisons. Specifically, we identified all TF pairs where the change in methylation exceeded one standard deviation above the average. This resulted in a total of 152 sets of conditions that showed significant variations in methylation. The YY1-TAF1 pair was noticeable in its difference, showing significance in three cell lines: GM12878, H1-hESC, and SK-N-SH.

Our study also employed TFregulomeR for Motif Distribution, Genomic Location, and Gene Ontology (GO) analysis. The tool offers a range of functionalities to ease data manipulation and interpretation. We used the 'data browser' feature to navigate through curated datasets and 'plotLogo' to visualize motif logos. The 'loadPeaks' function was used to import peak regions, while 'exportMMPFM' allowed us to extract motif PWMs and DNA methylation matrices. We focused on 'intersectPeakMatrix' and 'exclusivePeaks' to study interactions between TF cofactors and create the correct conditions for the analysis. The peak information also included the Tag fold change of each fragment.

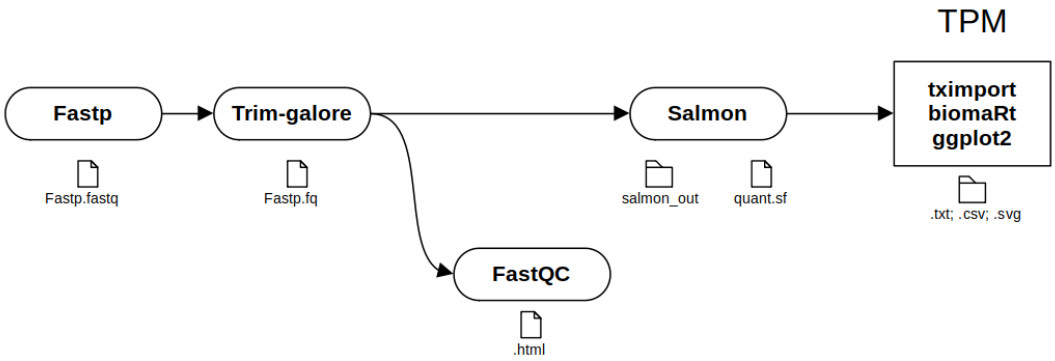
To better understand the biological implications of target genes associated with the promoter-TSS sites, we analyzed our data for enrichment of associated Gene Ontology (GO) terms. We used the list of genes derived from the ChIP-Seq data for each cell line under three conditions using TFRegulomR, taking advantage of the peak information generated by the HTML [13]. This list of genes could also be generated online using MethMotif [21]. This gene list was uploaded to David's website for KEGG and GO biological process enrichment analysis [21]. Tables describing enrichment terms were downloaded and converted into bubble plots using a threshold False Discovery Rate (FDR) of less than 0.05 for gene set enrichment.

4.3.2 Transcriptomic Analysis

We performed a transcriptomic analysis using the same gene list for the GO analysis using publicly available RNA-seq data. Our RNA-seq analysis pipeline is structured to include multiple computational phases such as quality control, read trimming, and transcript quantification. This pipeline is configured to operate on Linux-based systems and uses Bash scripts to automate and execute tasks in parallel. We utilized specific software tools for specialized tasks: FastQC for assessing data quality, Trim Galore! for quality and adapter

trimming of sequences, and Salmon for measuring transcript abundance [133], [134], [135]. A flowchart of the steps taken is provided in the figure below:

.



.

Figure 27. Flowchart of the RNA-seq data analysis pipeline used in the transcriptomic analysis. The process begins with quality control using Fastp, followed by adapter and quality trimming with Trim Galore!, FastQC for quality checks, transcript quantification via Salmon, and statistical analysis and visualization using R.

4.3.3 RNA-seq Data Availability

A list of file prefixes was defined to manage multiple samples. Each file prefix (ENCLB039ZZZ, ENCLB040ZZZ, etc.) was associated with a pair of FastQ files containing the sample's RNA-Seq reads. All the prefixes were related to their publicly available file names. The publicly available files were downloaded from Encode; their GEO IDs are GSE78552, GSE187560, and GSE175034 [136], [137]. Furthermore, an index for the human genome (hg38) generated by Salmon was specified and stored in the INDEX environment variable. The initial hg38 reference genome was obtained from the 'bigZips' directory provided by the UCSC Genome Browser [138]. After downloading and decompressing the hg38.fa.gz file, the reference genome was used to build the Salmon index for transcript quantification.

4.3.4 Tools Used for Processing, Analysis and Interpretation of Data

The first computational step in the pipeline was the quality control and filtering of raw reads using Fastp [139]. This software performs an initial analysis and filters the FastQ files. For each sample (prefix), the pipeline reads the corresponding FastQ files (${prefix}_1.fastq and ${prefix}_2.fastq) and creates filtered output files (${prefix}_1.fastp.fastq and ${prefix}_2.fastp.fastq) then employed Trim Galore! to remove any remaining adapters and low-quality bases [134]. This program is designed to work with paired-end reads and automatically handles read pairing. The trimmed reads are output to a dedicated subfolder, and the filenames of these trimmed files are also standardized for future steps. After trimming, FastQC is used for quality checks on the trimmed reads [133], and outputs are stored in individual folders.

For each sample, an output directory (salmon_output_${prefix}) was created to hold the results of the Salmon quantification process [135]. Salmon was used for transcript quantification with the -l A parameter specifying that the library was automatically detected. Paired-end trimmed FastQ files ($trimmed1 and $trimmed2) were used as input. Salmon used 16 threads (-p 16) and validated mappings (--validateMappings) to ensure high-quality quantification. The output, which included transcripts per million (TPM) and counts among other metrics, was stored in the Salmon output directory.

4.3.5 Data Import, Manipulation, and Statistical Analysis of RNA-Seq Data

We utilized various R libraries to facilitate data manipulation, visualization, and statistical analysis. Specifically, we used tximport for importing transcript-level estimates, biomaRt for transcript-to-gene mapping, and dplyr and tidyr for data wrangling. The ggplot2 and scales libraries were employed for data visualization, and statistical analyses were conducted using the stats package in R [140], [141], [142], [143], [144], [145], [146].

Initially, output files were generated by the Salmon software for the two cell lines under three different conditions [136], [137]. These files were imported into R using the tximport package, allowing us to extract Transcripts Per Million (TPM) directly for subsequent analysis. We utilized the gene symbols readily available from a derived file containing the peak information in TFregulomeR. This resulted in a data frame where rows represented unique gene symbols and columns reflected different conditions. The gene symbols matched the original TPM estimates, generating a final matrix containing gene-level aggregated TPM values. Next, we read predefined gene lists for our cell line and conditions to analyze overall trends in gene expression. We subsetted the original TPM data using these lists to generate three separate data frames, which were then prepared for visualization.

For the statistical analysis, we performed an Analysis of Variance (ANOVA) followed by Tukey's Honest Significant Difference (HSD) test to compare the average gene expression among all conditions. These results were visualized using a boxplot, which employed a logarithmic y-axis to better capture the variance in expression levels. Finally, we generated summary statistics for each condition, which included the number of genes, average and median expression levels, standard deviation, and minimum and maximum expression levels. These summaries provided an overview of the data distribution in each subset and were vital for interpretative analyses.

4.4 RESULTS

4.4.1 Revised the YY1 and YY1-TAF1 co-binding motifs

Using our approach, we could analyze the distinct peaks identified in the Chip-seq data of one transcription factor, exclusive of the other, in addition to co-bound. YY1 demonstrated DNA binding specificity to a distinct sequence, traditionally understood to be the 5'-CCGCCATNTT-3' consensus sequence, as cited in several previous studies [90], [91]. Nevertheless, our research revealed that this sequence initiated one base pair earlier in the 5' direction, coupled with weak conservation at the third position cytosine, thus suggesting a revised consensus motif of 5'-GCNGCCATNTT-3' (as depicted in Figure 28) when not co-bound with TAF1. The motif for the TAF1 exclusive peaks showed some level of cell specificity, with the motif for the GM12878 cell line being 5’-GCCGCCATNTT’-3, the motif of the H1-hESC cell line being 5’-NCCGCATNTT-’3, and the motif of the SK-N-SH cell line being 5’-NNCGCATNTT-’3. Co-bound, the YY1-TAF1 motif also displayed a certain level of cell specificity; in the GM12878, the motif was not conserved in the third position (5’-GCNGCCATNTT-’3), whereas in the H1-

hESC cell-line, the 3 bp position was more strongly conserved as a cytosine. DNA methylation data was added to unveil various methylation levels throughout peaks within the YY1 motif. On the other hand, a markedly low level of methylation could be seen when there was a co-binding of YY1 and TAF1, mirroring the observations predominantly seen in TAF1 peaks. This phenomenon was further accentuated by the heightened conservation of the third base pair cytosine in the motif, potentially facilitating methylation at the CpG pair between the motif's 3-4 base pairs.

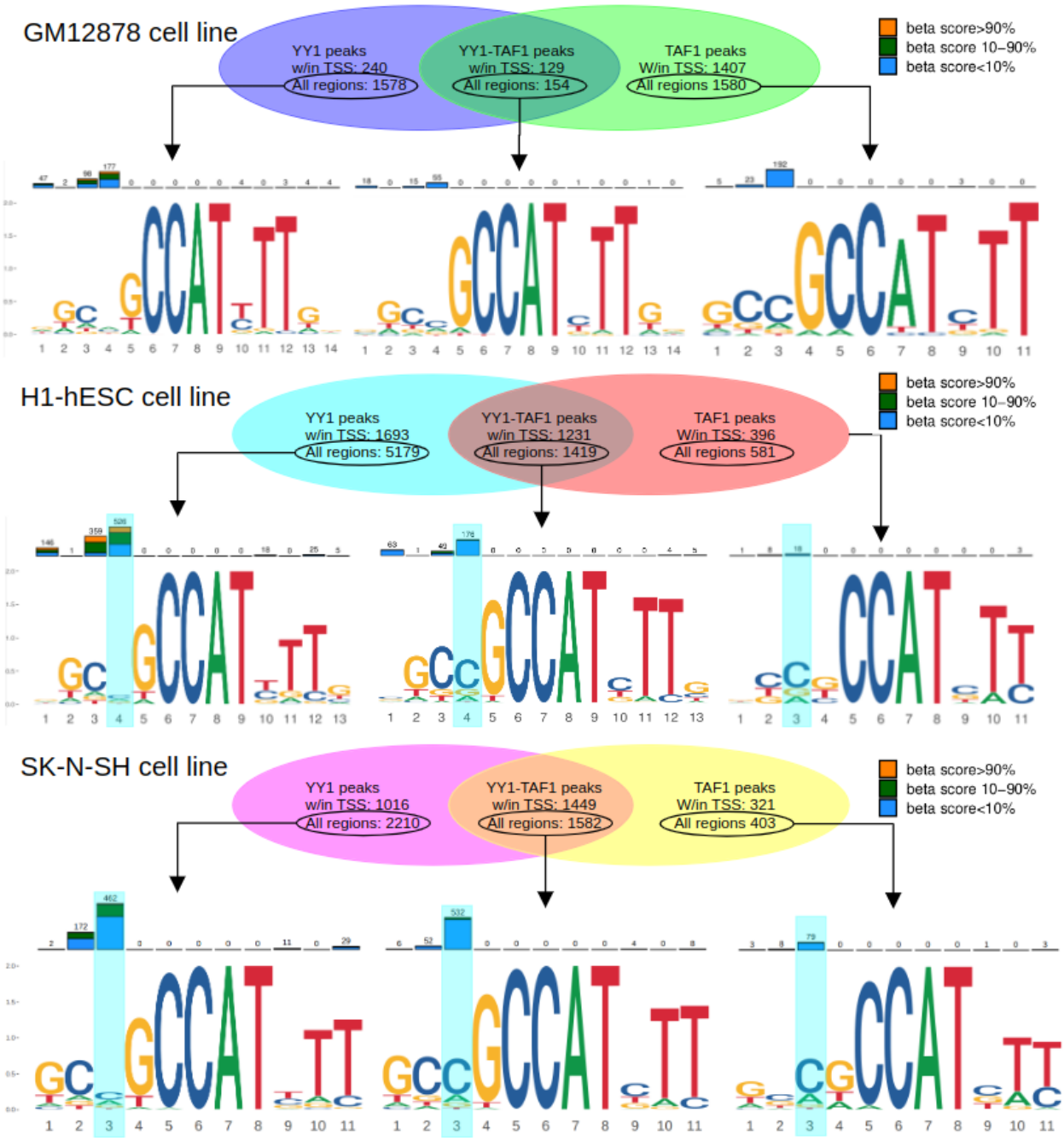


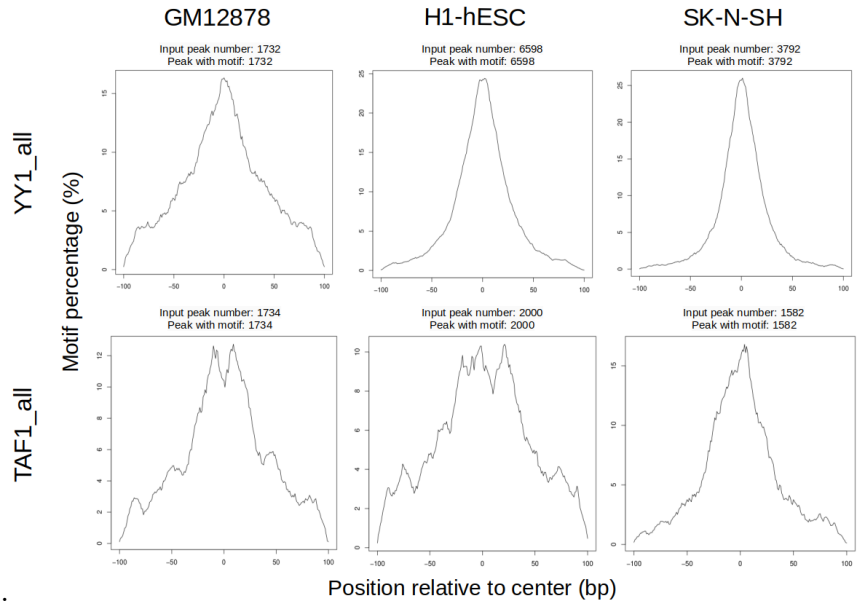
Figure 28. The Diagram Illustrates the Motif and methylation comparison of YY1 and TAF1 from ChIP-seq data across GM12878, H1-hESC, and SK-N-SH. A) YY1-exclusive peaks motif shares the same motif as co-bound YY1-TAF1 within the GM12878 cell line. The first two bp in the motif are not conserved in the TAF1-exclusive motif, transitioning to 5'-NNCGCCAATNTT-3'. B&C) Examining methylation levels within motif peaks highlights variable methylation for YY1 peaks in the H1-hESC and SK-N-SH cell lines, respectively. A significant reduction

in methylation levels is observed during YY1-TAF1 co-binding, mirroring the methylation pattern primarily seen in TAF1 peaks. This suggests a potential cooperative regulatory mechanism between YY1 and TAF1.

4.4.2 Comparative Analysis of YY1 and TAF1 Motif Distributions in ChIP-seq

The motif distribution reveals the frequency and specific genomic locations where a transcription factor attaches to DNA. In the context of ChIP-seq analysis, a sharp, normal distribution of the motif around the peak summit, as observed for YY1, indicates that most motif occurrences are clustered tightly around the peak summit (Figure 29). In other words, the transcription factor binding events are highly concentrated at a specific position relative to the center of the binding peak. The motif distribution for TAF1 shows a binomial pattern within the GM12878 and H1-hESC cell lines, along with a broader spread than YY1 in all three cell lines (GM12878, H1-hESC, and SK-N-SH). This indicates a less localized DNA binding and suggests that TAF1 has a more versatile genomic binding profile, potentially binding to other transcription factors rather than onto the DNA directly and exerting a more flexible regulatory influence.

.



.

Figure 29. This Figure shows the motif distributions for YY1 and TAF1 across GM12878, H1-hESC, and SK-N-SH cell lines. YY1 exhibits a sharp, normal distribution around the peak summit, indicating concentrated binding events at specific genomic locations. In contrast, TAF1 displays a binomial pattern with a broader spread, suggesting a more versatile and less localized binding profile across the examined cell lines.

4.4.3 YY1 and TAF1 ChIP-seq Dynamics: Tag Fold Change and Peak Distribution Analysis

By examining the variations in tag fold change within the peak values of ChIP-seq data, we can gain valuable insight into the behaviour of the transcription factor and the complexities of the formations they constitute. When we focus our analysis on the peaks of the YY1 transcription factors in our selected cell lines and scrutinize them based on tag fold change, we observe a positively skewed distribution (Figure 30). This distribution is characterized by a predominant mode within the 8-13 range, showcasing a potential commonality in the transcription factor's behaviour within this specific range. Delving further into the data, an analysis of the subset where YY1 co-binds with TAF1 reveals a notable shift in the mode, steering toward a platykurtic distribution (Figure 30). This deviation in the distribution pattern highlights a distinct interaction behaviour between YY1 and TAF1. It points to a probable shift in the functional dynamics compared to scenarios where YY1 operates in isolation. This suggests that the co-localization of YY1 and TAF1 might bring about a nuanced modification in their functional interplay, validating the complex formation between the two transcription factors.

.

.

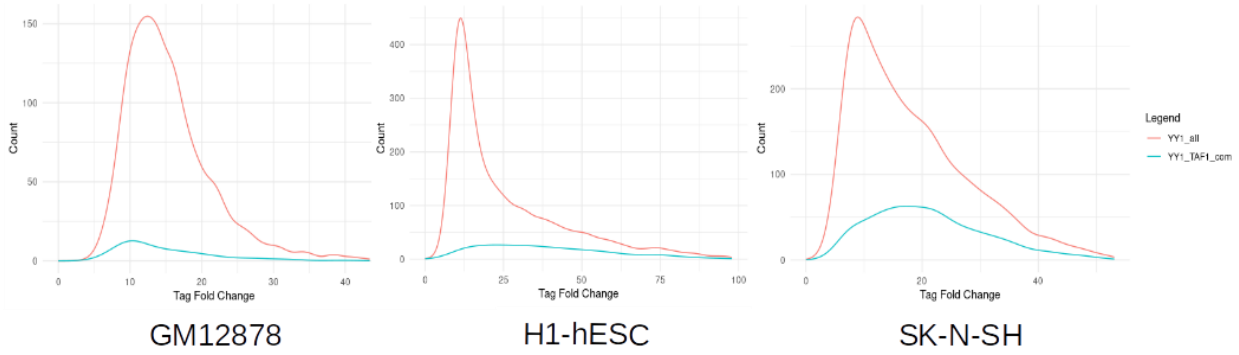
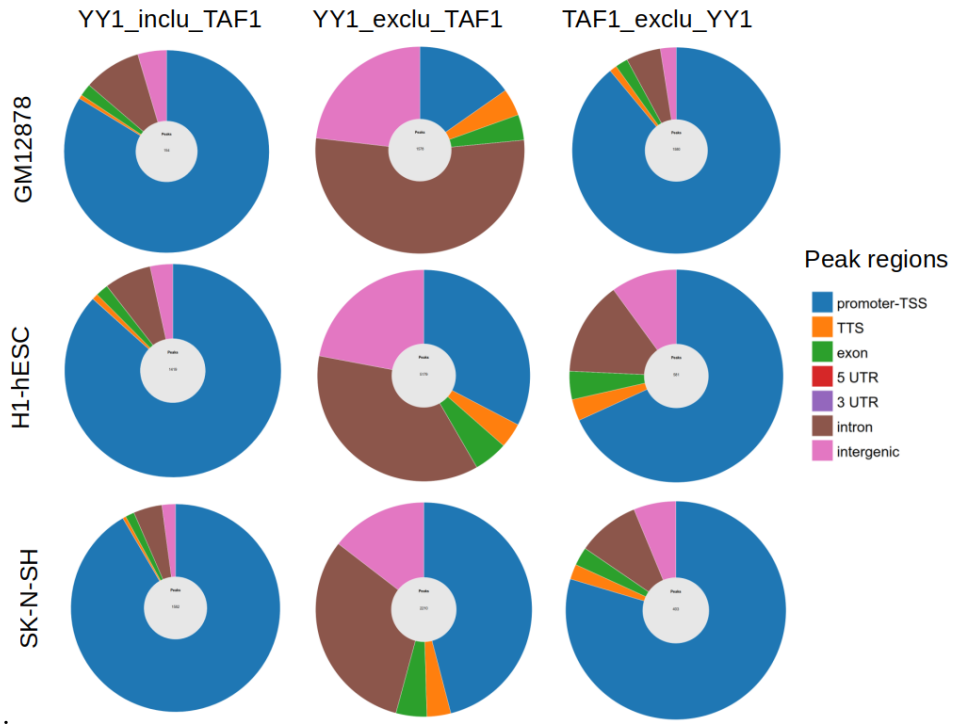


Figure 30. This Figure illustrates the tag fold change distribution for YY1 transcription factors across selected cell lines, showing a positively skewed pattern with a mode in the 8-13 TFC range. An overlapping analysis of YY1-TAF1 co-binding indicates a shift towards a platykurtic distribution with a mode at a higher TFC range.

The locations of binding peaks within the genome offer valuable insights into the functions and activities of transcription factors. Our study presents consistent patterns across three distinct cell lines—GM12878, SK-N-SH, and H1-hESC—concerning the distribution of transcription factor binding sites (Figure 31). Specifically, in our analysis of ChIP-seq peaks unique to YY1 (and not associated with TAF1), we noted a considerable accumulation of peaks in intronic, intergenic, and promoter-transcription start site (TSS) regions.

.



.

Figure 31. These pie charts illustrate the region distribution of ChIP-seq binding peaks for transcription factors YY1 and TAF1 across three cell lines: GM12878, SK-N-SH, and H1-hESC. Each chart represents the percentage of peaks located in intronic, intergenic, and promoter-TSS regions, segmented by individual and shared binding scenarios of YY1 and TAF1.

For example, in the GM12878 cell line, the intronic regions contain the majority of peaks, accounting for approximately 53.42% of the total. In contrast, the SK-N-SH cell line shows a higher prevalence of YY1 binding in the promoter-TSS regions, making up 45.97% of the peaks. In the H1-hESC cell line, the distribution of peaks between intronic and promoter-TSS regions is more balanced, with percentages of 36.30% and 32.69%, respectively.

When analyzing ChIP-seq data for peaks exclusive to TAF1 in GM12878, H1-hESC, and SK-N-SH cell lines, we observed that most of these peaks are concentrated in the promoter-TSS regions, accounting for 89.05%, 68.16%, and 79.65% of the total peaks, respectively. For peaks shared between YY1 and TAF1, the majority were also located within the promoter-TSS regions across all three cell lines: 83.77% in GM12878, 86.75% in H1-hESC, and 91.59% in SK-N-SH. Interestingly, the proportion of peaks in the promoter-TSS regions increased for the shared YY1-TAF1 peaks in the H1-hESC and SK-N-SH cell lines. Conversely, the GM12878 cell line showed a decrease in the percentage of peaks located in these regions when YY1 and TAF1 were found to bind together. These variations in peak localization, especially concerning the shared YY1-TAF1 peaks, could indicate distinct regulatory mechanisms across different cellular contexts.

4.4.4 Enhanced Gene Expression in YY1-TAF1 Co-binding

We obtained a list of target genes via TFRegulomR, leveraging the peak information supplied by the HTML [132] and subsetting the gene list for peaks found within the promoter-TSS sites. Our TPM analysis across the GM12878 and H1-hESC cell lines indicated that the YY1 and TAF1 co-binding was linked to increased gene expression. The most significant variations were observed in GM12878 cells. Genes regulated by co-bound YY1-TAF1 tended to have higher expression levels across all three cell lines than genes regulated exclusively by YY1

or TAF1 (Figure 32). These genes were significant in the pairwise comparison between the YY1-TAF1 common and YY1 exclusive peaks in the GM12878 cell line (adjusted p-value < 0.01). The same trend was observed when investigating the promoter-TSS-associated genes. However, there is no statistically significance difference between the all-inclusive annotation gene-associated peaks for YY1-TAF1 and the promoter-TSS gene-associated peaks for YY1-TAF1.

.

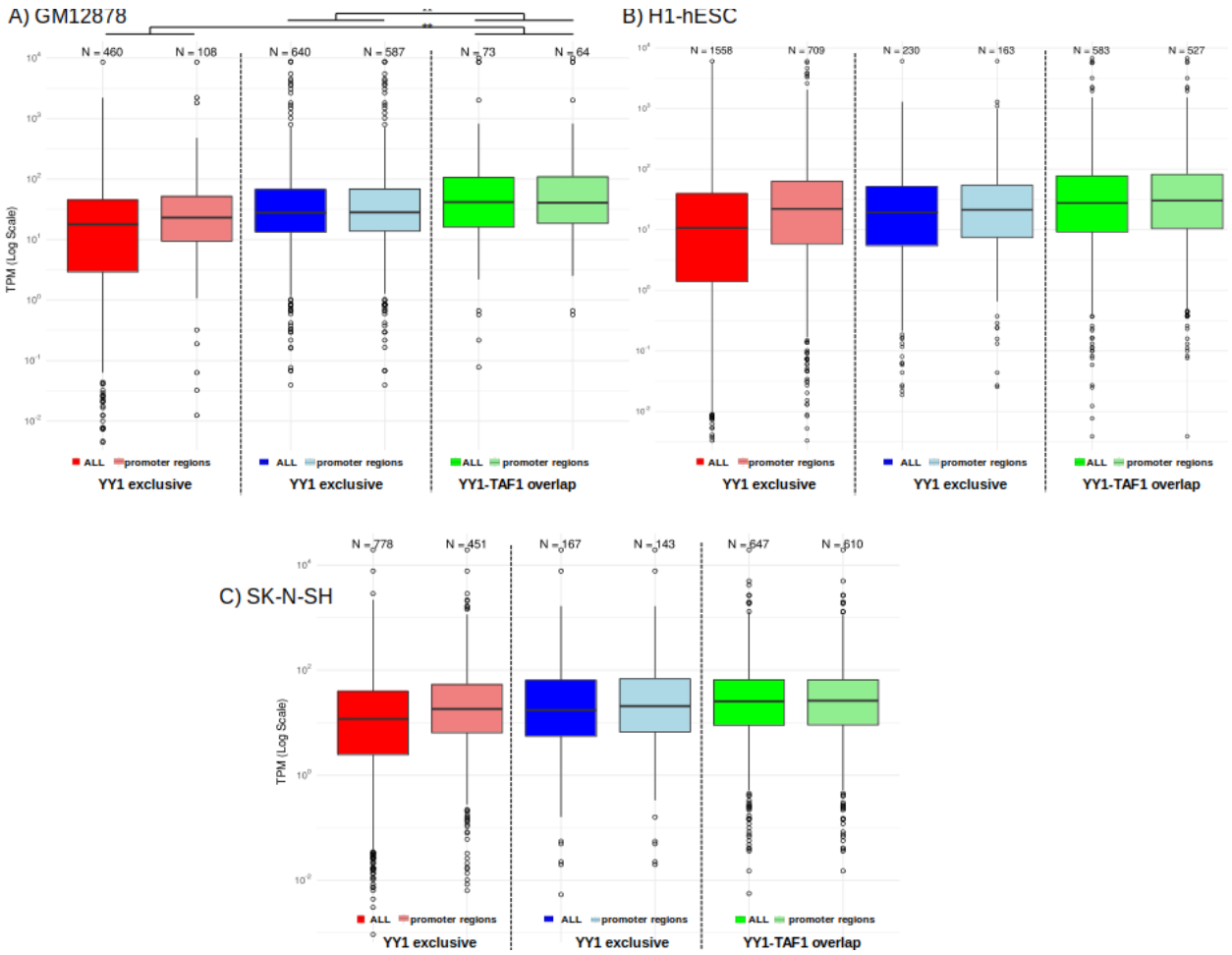
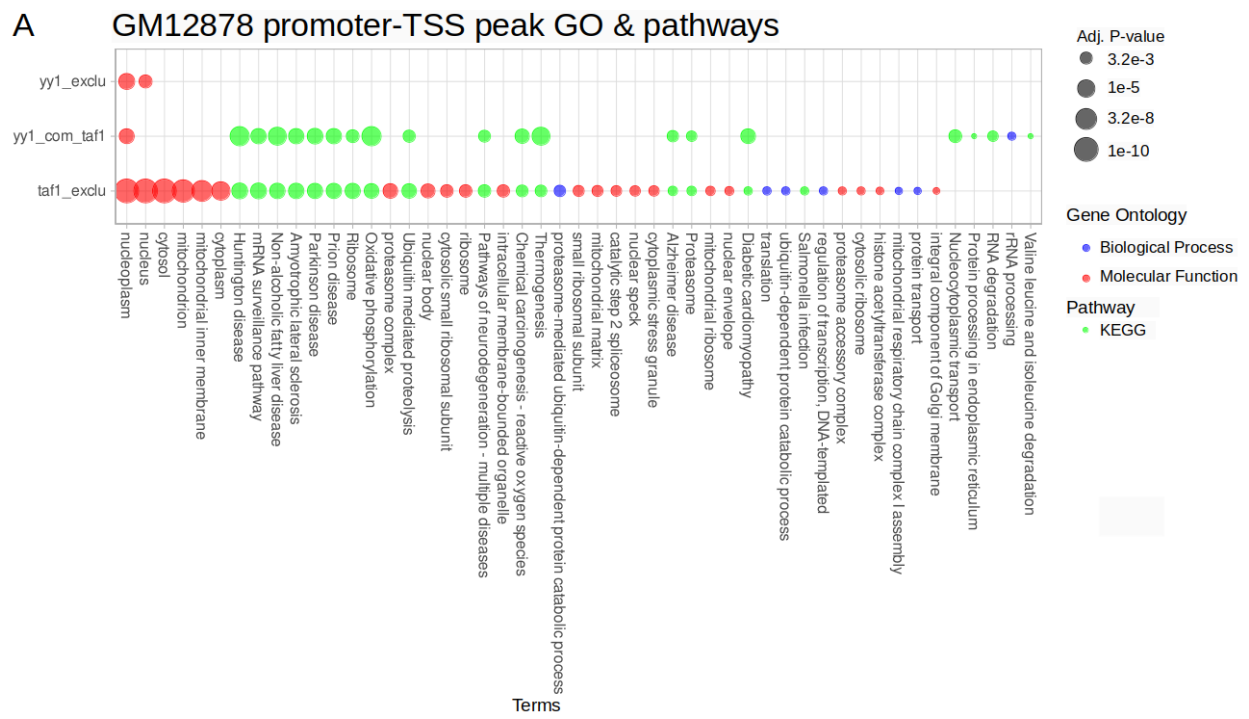


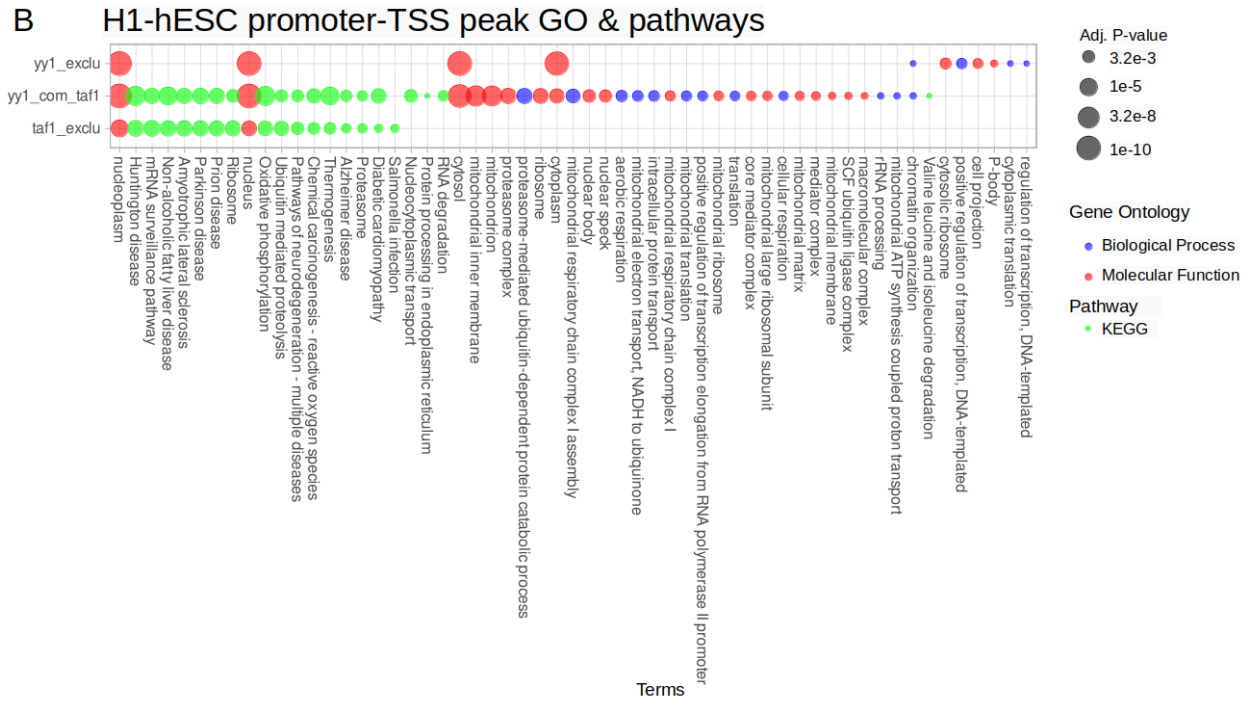
Figure 32. Boxplot Depicting the Transcripts Per Million (TPM) Levels of Genes Associated with YY1 and TAF1 Across GM12878 and H1-hESC Cell Lines. The plot segregates data based on the exclusive and co-bound peaks identified in the ChIP-seq analysis. Also, it provides another set of plots for further narrowing down the genes associated with promoter-TSS peaks in the GM12878 cell line (A), the H1-hESC cell line (B), or the SK-N-SH cell line (C).

4.4.5 Promoter-TSS Peaks Gene Ontology Analysis of YY1 and TAF1

We subsequently uploaded this promoter-TSS peak-gene list to David's platform for enrichment analysis of GO terms [29]. The retrieved enrichment terms were illustrated as a bubble plot, applying a threshold FDR of less than 0.05. Although some characteristics were shared in both cell lines for the YY1 and TAF1 binding condition, the GO enrichment analysis indicated a large degree of cell specificity, as indicated by the extent to which GO enrichment was impacted by the YY1 and/or TAF1 binding. In the GM12878 cell line, the shared YY1-TAF1 peaks were linked to more terms (21 terms, 1 BP, 1 CC, 19 KEGG) than the YY1-exclusive peaks (2 terms, 0 BP, 2 CC, 0 KEGG). The promoter-TSS TAF1-exclusive genes reflected significantly more terms than the other two conditions (44 terms, 6 BP, 22 CC, 16 KEGG); these peaks in GM12878 were connected to many GO terms encompassing diverse cellular compartments and processes, from ribosomal activities to transcription regulation and proteasomal functions. Conversely, although we also found a larger number of GO terms associated with YY1-TAF1 co-bound peaks, there was a decrease in GO terms related to TAF1 exclusive peaks. In the H1-hESC cell line, the number of significantly enriched terms was 11 for YY1-exclusive peaks (4 BP, 7 CC, 0 KEGG), 50 for shared YY1-TAF1 peaks (12 BP, 19 CC, 19 KEGG), and 18 for TAF1-exclusive peaks (0 BP, 2 CC, 16 KEGG); in the SK-N-SH cell line, the number of significantly enriched terms were 24 for YY1-exclusive peaks (5 BP, 6 CC, 13 KEGG), 52 for shared YY1-TAF1 peaks (14 BP, 19 CC, 20 KEGG), and 2 for TAF1-exclusive peaks (0 BP, 2 CC, 0 KEGG).


A
GM12878 promoter-TSS peak GO & pathways
yy1_exclu
yy1_com_taf1
taf1_exclu
nucleoplasm
nucleus
cytosol
mitochondrion
mitochondrial inner membrane
cytoplasm
Huntington disease
mRNA surveillance pathway
Non-alcoholic fatty liver disease
Amyotrophic lateral sclerosis
Parkinson disease
Prion disease
Ribosome
Oxidative phosphorylation
proteasome complex
Ubiquitin mediated proteolysis
nuclear body
cytosolic small ribosomal subunit
ribosome
Pathways of neurodegeneration - multiple diseases
intracellular membrane-bounded organelle
Chemical carcinogenesis - reactive oxygen species
Thermogenesis
proteasome-mediated ubiquitin-dependent protein catabolic process
small ribosomal subunit
mitochondrial matrix
catalytic step 2 spliceosome
nuclear speck
cytoplasmic stress granule
Alzheimer disease
Proteasome
mitochondrial ribosome
nuclear envelope
Diabetic cardiomyopathy
translation
ubiquitin-dependent protein catabolic process
Salmonella infection
regulation of transcription, DNA-templated
proteasome accessory complex
cytosolic ribosome
mitochondrial respiratory chain complex I assembly
histone acetyltransferase complex
protein transport
integral component of Golgi membrane
Nucleocytoplasmic transport
Protein processing in endoplasmic reticulum
RNA degradation
rRNA processing
Valine leucine and isoleucine degradation
Terms
Adj. P-value
3.2e-3
1e-5
3.2e-8
1e-10
Gene Ontology
Biological Process
Molecular Function
Pathway
KEGG

OBJ OBJ


B
H1-hESC promoter-TSS peak GO & pathways
yy1_exclu
yy1_com_taf1
taf1_exclu
nucleoplasm
Huntington disease
mRNA surveillance pathway
Non-alcoholic fatty liver disease
Amyotrophic lateral sclerosis
Parkinson disease
Prion disease
Ribosome
nucleus
Oxidative phosphorylation
Ubiquitin mediated proteolysis
Pathways of neurodegeneration - multiple diseases
Chemical carcinogenesis - reactive oxygen species
Thermogenesis
Alzheimer disease
Proteasome
Diabetic cardiomyopathy
Salmonella infection
Nucleocytoplasmic transport
Protein processing in endoplasmic reticulum
RNA degradation
cytosol
mitochondrion
mitochondrial inner membrane
proteasome complex
proteasome-mediated ubiquitin-dependent protein catabolic process
ribosome
cytoplasm
mitochondrial respiratory chain complex I assembly
nuclear body
nuclear speck
intracellular protein transport
mitochondrial electron transport, NADH to ubiquinone
aerobic respiration
mitochondrial respiratory chain complex I
mitochondrial translation
positive regulation of transcription elongation from RNA polymerase II promoter
mitochondrial ribosome
translation
core mediator complex
mitochondrial large ribosomal subunit
cellular respiration
mitochondrial matrix
mediator complex
mitochondrial membrane
SCF ubiquitin ligase complex
macromolecular complex
rRNA processing
mitochondrial ATP synthesis coupled proton transport
chromatin organization
Valine leucine and isoleucine degradation
cytosolic ribosome
positive regulation of transcription, DNA-templated
cell projection
P-body
cytoplasmic translation
regulation of transcription, DNA-templated
Terms
Adj. P-value
3.2e-3
1e-5
3.2e-8
1e-10
Gene Ontology
Biological Process
Molecular Function
Pathway
KEGG

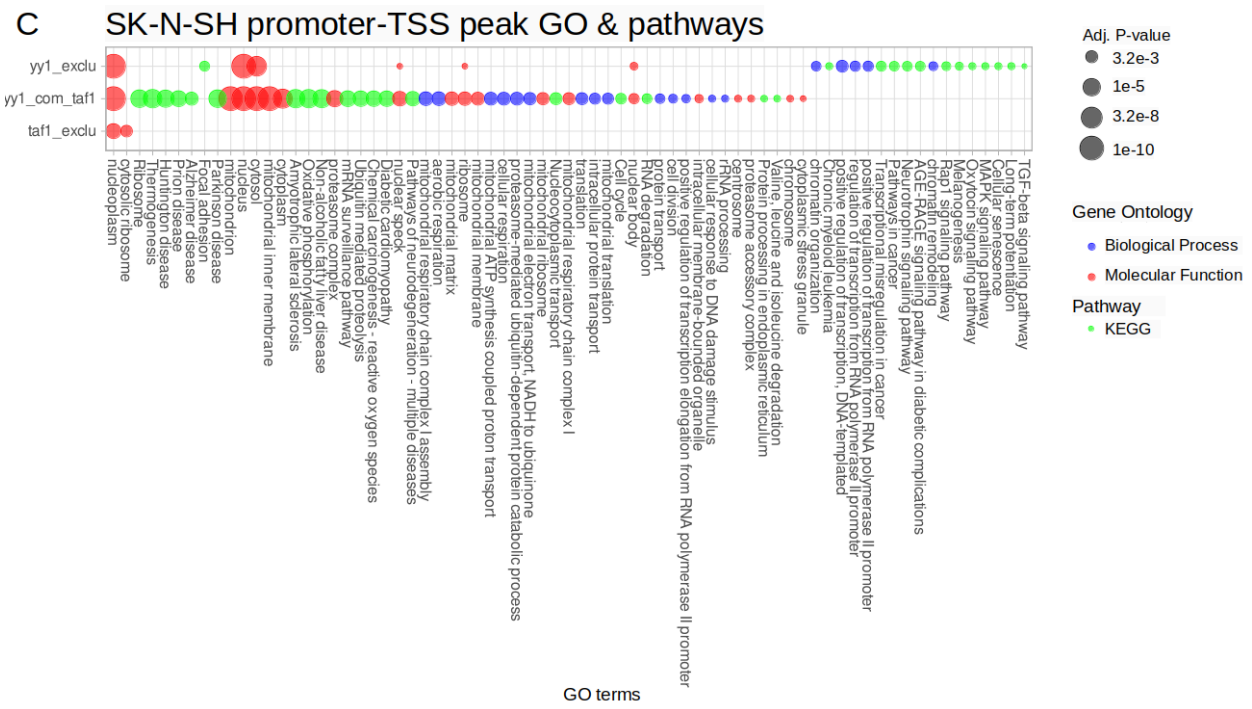


Figure 33. Bubble plot of Gene Ontology (GO) Terms for YY1 and TAF1 Peaks in the GM12878, H1-hESC, and SK-N-SH cell lines. The bubble plot displays all GO linked to the promoter-TSS sites with an FDR < 0.05. All FDRs that were less than 1e-10 were rounded up to 1e-10 to keep the size of the bubbles readable. Colour represents the type of GO: blue represents biological processes, red represents cellular components, and green represents molecular functions. Notably, the co-binding of YY1-TAF1 widens the spectrum GO terms compared to YY1 exclusive peaks. Refer to the supplemental section for a detailed breakdown of specific GO terms.

4.5 Discussion

In the complex hierarchy of gene regulation, understanding the subtle shifts in the behaviour of TFs can provide insights into the regulatory dynamics of spatial and temporal gene expression. The transcription factor YY1 is a pivotal element in multiple cellular processes, characterized by its roles in cell proliferation, differentiation, DNA repair, and apoptosis [90], [91]. Our ChIP-seq data analysis shed light on a significant modification in the YY1 binding consensus sequence when TAF1 was present; specifically, there was enhanced preservation of cytosine at the third position within the motif. This modification suggests a nuanced shift in function whereby the regulatory actions observed when YY1 operates independently differ from those observed when co-binding with TAF1.

4.5.1 Insights of YY1-TAF1 Co-Binding Dynamics

Our analysis adopts a nuanced approach to understanding the interaction between TAF1 and YY1. By constructing the TAF1 motif from data excluding the overlapping peaks with YY1, we observed some similarity to the YY1 peak, with minor deviations in the first or fourth base pairs that are cell type-specific (Figure 28). These deviations highlight the unique binding motifs of TAF1 and YY1 under varying contexts, emphasizing how the cell type and the presence of TAF1 co-binding can influence YY1's regulatory function. Aligning with the current understanding of DNA methylation, where decreased promoter region methylation correlates with increased transcription binding, we noticed reduced methylation levels on the YY1 binding motif during YY1 and TAF1 co-binding. This suggests a favourable environment for their co-binding and subsequent transcriptional activation of target genes.

Further substantiating this theory is the motif distribution of YY1 and TAF1, which indicates TAF1's direct engagement with YY1 during complex formation. A sharp, normal distribution of YY1's localization on the motifs is observed, denoting a precise, localized binding to genomic sites—a hallmark of YY1's established role in focused gene binding and regulation. On the other hand, TAF1 showcases a bi-modal binding landscape, a characteristic vital for initiating and sustaining transcription, as seen in studies [96], [97], [98]. This attribute is prominently observed across various cell lines, including GM12878, H1-hESC, and SK-N-SH. Intriguingly, the distribution of TAF1's binding motifs hints at its capacity to bind to other transcription factors or co-factors.

Following the same line of thought, gene expression analysis showed that sites where co-binding of YY1 with TAF1 occurred increased the average target gene expression (Figure 31). Thus, a compelling hypothesis can be formed by integrating these observations, including the alterations to the binding motif in the co-binding of YY1-TAF1, the observed decrease in methylation levels in the YY1 motif during the co-binding of YY1 and TAF1, and the increase in gene expression at the YY1-TAF1 sites. These elements collectively suggest that methylation affects the co-binding of YY1 and TAF1, which promotes target gene transcription (Figure 33).

The figure below summarizes our findings into a model for a state-dependent co-binding of TAF1 to YY1. When the YY1 binding motif is methylated, having a cytosine or guanine residue at the third residue, TAF1 will not co-bind to the YY1 bound transcription factor. If methylation is impossible due to the third residue within the motif not supporting methylation, TAF1 will also fail to co-bind. When YY1 binds without TAF1, it will have a stronger affinity to act as a repressor. Although the motif for the co-binding YY1-TAF1 peaks is cell-specific, we see stronger cytosine conservation within the third residue, within the peaks that we find YY1-

TAF1 co-binding. In addition, although methylation is possible at this site, the binding of TAF1 onto YY1 depends on an unmethylated state. When TAF1 is co-bound to YY1 in an unmethylated form, the complex has a stronger tendency to act as a transcriptional activator.

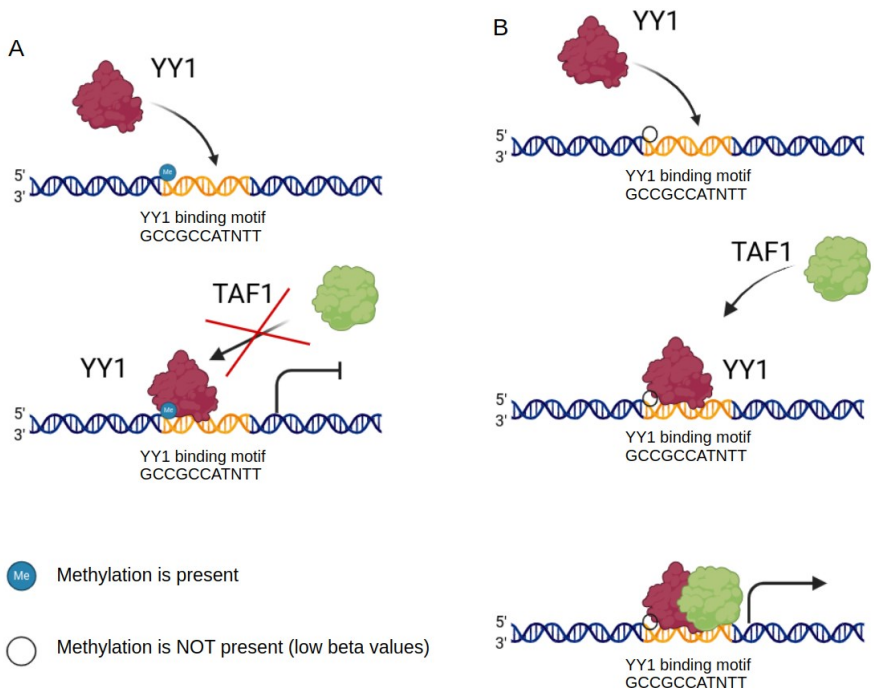


Figure 34. Differential Binding Mechanisms of YY1 and TAF1 in Response to Methylation: The Figure delineates the distinct interactions of Yin Yang 1 (YY1) and TATA-box Binding Protein Associated Factor 1 (TAF1) contingent on the YY1 binding motif's methylation. On the left in Figure A, methylation of the YY1 motif permits the binding of YY1 while precluding TAF1 association, resulting in transcription repression. In cases where methylation cannot occur due to the base pair sequence of the motif, TAF1 still does not bind with YY1. On the right side, cytosine is preferentially found at the third residue of the motif and unmethylated conditions, enabling the sequential binding of YY1 and TAF1, thus promoting YY1's role as a transcriptional activator.

4.5.2 Divergent Roles of YY1-TAF1 Co-Binding in GM12878 and H1-hESC

Our Gene Ontology (GO) analysis in GM12878, H1-hESC, and SK-N-SH cell lines shows that the co-binding of YY1 and TAF1 results in an expanded set of GO terms compared to when YY1 is expressed independently. This expanded range suggests a synergistic effect between YY1 and TAF1 in regulating various cellular processes. We observe that the number of genes and GO terms associated with YY1-TAF1 at promoter-TSS sites varies between cell types, indicating a more extensive set of co-binding sites and cellular functions in cells capable of differentiation. Specifically, in GM12878 cells, YY1-TAF1 co-binding correlates with more narrow functions related to ribosomal machinery, RNA processing, and nucleoplasm activities, highlighting their critical role in protein synthesis and RNA maturation in these cells. Conversely, in the H1-hESC and Sk-N-SH cell lines, the YY1-TAF1 co-binding correlates with a wide range of enrichments. Interestingly, TAF1's association with promoter-TSS genes in the same cell line reveals the opposite pattern as YY1-TAF1 co-bound peaks; a broader array of processes in the GM12878 cell line, and a more narrow range in the H1-hESC and Sk-N-SH cell-lines. This difference suggests that each cell line's cellular needs and differentiation potential influence the binding patterns. GM12878 cells, being more mature, may not require rapid responses to environmental changes, leading to a more pronounced role for TAF1 in gene expression regulation. In contrast, H1-hESC cells, primed for differentiation, might need more stringent control over gene expression to navigate their developmental paths. The SK-N-SH can also differentiate to some extent, which is why it likely follows the same trends as the H1-hESC cell -line. Here, the dual role of YY1, as both an inhibitor and promoter of gene expression, becomes crucial.

4.5.3 Limitation and Concluding Statement

One limitation of our study stems from the limited evidence to specify the nature of the relationship between YY1 and TAF1—namely, whether they form a physical complex, co-bind to DNA, or are merely co-bound in cells. Although our ChIP-seq data reveal overlapping binding peaks for YY1 and TAF1, and Gene Ontology (GO) analysis suggests a possible functional collaboration, these findings do not necessarily confirm direct interactions or complex formation. Without deploying additional investigative techniques like co-immunoprecipitation, fluorescence resonance energy transfer (FRET), or mutational analyses, we can only theorize about the nuances of their interaction and its functional implications.

Future research could benefit from specific experimental techniques to narrow this knowledge gap. For instance, co-immunoprecipitation assays could validate whether YY1 and TAF1 form a cohesive protein complex. Visualization methods like FRET or bimolecular fluorescence complementation (BiFC) could offer insights into the spatial relationship between these proteins within live cells. Furthermore, functional validation through gene knockout or knockdown experiments could elucidate whether the co-binding of both YY1 and TAF1 is essential for the gene expression patterns and biological processes we observed. Based on what is known of TAF, it would come as no surprise if TAF is indirectly bound to the proximal DNA; as such, a series of co-immunoprecipitation or crystallography experiments may also elucidate transcription factors that play a role in our model.

In conclusion, our investigation into the intricate synergistic dynamics between transcription factors YY1 and TAF1 unveils a novel mechanism of co-binding that integrates closely with our unifying theme of methylation analysis. This study expands on how YY1-TAF1 co-binding impacts gene promoter activity and fine-tunes the methylation dynamics at these

sites. Leveraging advanced computational methods alongside RNA-seq, ChIP-seq, and bisulphite sequencing, we have mapped out how these factors are dependent on methylation changes and specificity in the motif sequence. This synergy results in alterations in the YY1 binding consensus sequence and a notable decrease in methylation levels on the YY1 motif, indicating that methylation plays a role in the YY1-TAF1 co-binding. Furthermore, our gene expression analysis uncovers that the YY1-TAF1 co-bound peaks occur with a subset of YY1's peaks primarily located in the promoter-TSS regions, as opposed to intronic and intergenic domains, indicating its pivotal shift in role, going from orchestrating complex regulatory strategies to acting primarily as a promoter. This study is a significant milestone in understanding the intricate interplays between YY1 and TAF1, offering promising avenues for further exploration of therapeutic potentials harboured within their synergistic landscapes. It beckons a bright future for genetic transcription research with profound implications in cellular biology and potential therapeutic advancements.

Chapter 5: Conclusion and Perspectives

5.1 Broad summary and limitations

In our analysis of Kidney Renal Papillary Cell Carcinoma (KIRP), we identified three distinct DNA methylation signatures correlating to different survival outcomes. Particularly, we identified a CpG Island Methylator Phenotype (CIMP) in patients with poor overall survival outcomes. This correlation underpins the recurring theme across various cancers where CIMP methylation patterns are linked to adverse prognoses. The DNA methylation pattern in patients with the best overall survival cluster alongside normal samples when using k-means clustering indicates a methylation signature closest to normal.

The CIMP profile associated with the worst survival outcomes has the highest Fraction of Genome Altered (FGA) and a low mutation rate compared to the other two groups. Conversely, the group with intermediate overall survival had low FGA and a high mutation rate. These findings suggest two different mechanisms of tumorigenesis. The CIMP group's tumours likely result from broad structural changes causing genomic instability and cancer development. On the other hand, the patients in the intermediate group likely have a high number of SNPs or other small DNA mutations. The divergent relationships between epigenetic modifications and gene expression across different KIRP patient groups further support the notion of distinct mechanisms for tumorigenesis. In the CIMP group, numerous hypermethylated DMRs do not coincide with a large number of differentially expressed genes. Conversely, the intermediate overall survival group has a lower level of DMRs but the highest level of DEGs. Despite extensive methylation differences, gene expression patterns in both CIMP and near-normal groups show consistency, with lower variability in log fold change of DEGs, challenging the idea

that gene expression heterogeneity alone determines poor survival in CIMP groups. However, the intermediate survival group, characterized by a significant number of DMRs but distinctly expressed DEGs, shows substantial deviation in gene expression from normal tissue. This pronounced change, coupled with less extensive methylation than the CIMP group, implies that different molecular mechanisms, possibly involving a higher degree of smaller-scale mutations directly impacting gene expression, are crucial for the observed intermediate survival outcomes.

Categorized methylomic signatures more accurately reflect representative clinical features such as tumour stage and sex distribution. In the case of the tumour stage, the refined ML algorithm isolated a larger number of stage 3 tumour patients in the poor survival outcome group. This finding supports that methylation pattern changes correlate with tumour stages and survival outcomes. In the context of the sex distribution, k-means clustering initially biases female patients into the poor survival outcome group. However, adding the EM stage to the ML algorithm reduces this bias.

Our investigation reveals a link between the methylation status of certain TF motifs in our poor survival outcome group—namely, the transcription factor motifs of CTCF, KLF1, KLF9, REST, SP2, SP3, and WT1. These TFs, crucial for cellular processes like genomic integrity maintenance, cell proliferation, differentiation, and apoptosis, show aberrant methylation patterns. Gene ontology analysis of the sum effect of these transcription factors reveals dysregulation affects pathways linked to tumour cell metastatic potential, promotion of angiogenesis, immune evasion, and resistance to apoptosis; the dysregulated gene ontology terms include 'intracellular signal transduction' and the 'Wnt signalling pathway.' Additionally, abnormalities in cell adhesion mechanisms related to 'cell-cell adhesion' are crucial for cancer stem cells (CSCs) and play a critical role in initiating cancer metastasis.

In the study on Alzheimer's disease (AD), we aimed to validate an ML algorithm. With slight modifications, we successfully differentiated patients into subgroups reflecting their methylation signatures and approximated Braak stages of AD neuropathology using entorhinal cortex tissue samples. Once groups use the ML algorithm, statistical power improves, allowing for the identification of 2015 genes associated with differentially methylated regions (GADMR) in the pseudo-intermediate AD group and 4770 genes in the pseudo-advanced AD group. Notably, many of these genes overlapped with known AD-associated genes. Specifically, 11.6% of the genes in the pseudo-intermediate group and 27.2% in the pseudo-advanced group matched with previously identified AD-related genes. This showcases the potential effect of dysregulated methylation levels on AD-associated genes. The overlap with known genes validates the ML algorithm's effectiveness and suggests that the novel genes and CpG sites merit further investigation in the context of understanding AD progression. Enrichment analysis added depth to our findings, linking the GADMR to known and novel GO and KEGG terms associated with AD. This suggests potential new pathways involved in the disease's progression or severity, offering avenues for therapeutic intervention. The discovery of 87 new GO and KEGG terms in advanced AD patients enriches our understanding of AD. These new GO and KEGG terms open up potentially unexplored molecular pathways that might influence the disease's course.

Our age analysis further supported the validity of the ML algorithm, showing that ML-classified patients tended to have increased chronological age and DNA methylation age, which aligns with known risk factors and progression markers in AD. The consistency of chronological and DNA methylation age within these groups reinforced the ML algorithm's accuracy in patient stratification. Moreover, the study's approach reflects broader challenges in Alzheimer's research, such as the heterogeneity of the disease. AD patients display diverse pathological

features that complicate the identification of consistent biomarkers. Our application of a hybrid ML approach, integrating k-means clustering with EM, addresses these categorization challenges, allowing for more nuanced analysis based on epigenetic profiles rather than solely on clinical features.

Our application of k-means clustering combined with the EM algorithm in both KIRP and AD demonstrates the efficacy of hybrid machine-learning approaches in stratifying patient populations based on methylation profiles. This methodology enhances the precision and accuracy of disease classification and supports its broader application in genomics. The successful differentiation of patients into subgroups with distinct methylation signatures in both diseases underscores the potential of these ML techniques to uncover patterns that have yet to be explored.

The correlation between methylation signatures and clinical features, such as tumour stage and sex distribution in KIRP, underscores the clinical relevance of our findings. The refined ML algorithm's ability to accurately isolate stage 3 tumour patients in the poor survival outcome group and reduce initial biases related to sex distribution highlights its potential for clinical diagnostics. In AD, the alignment of DNA methylation age with chronological age in ML-classified patients supports using the ML framework in isolating biomarkers for disease diagnosis and progression. This precision in patient stratification is crucial for enhancing our understanding of the epigenetic landscape of diseases and guiding personalized therapeutic strategies.

The dysregulation of TF motifs in KIRP and their impact on pathways related to metastasis and angiogenesis mirrors the role of TFs in AD, where aberrant methylation affects pathways associated with neuronal survival and synaptic function. These parallels suggest that

targeting specific TF pathways could be a unifying strategy in developing therapies for both diseases. Synthesizing findings from KIRP and AD provides a comprehensive understanding of the interplay between epigenetic modifications and disease mechanisms. Using ML techniques to analyze methylation profiles has enhanced our ability to classify patient subgroups accurately and revealed critical molecular pathways that could be targeted for therapeutic intervention. These insights pave the way for future research to explore novel pathways, integrate multi-omic data, and translate findings into clinical applications, ultimately aiming to improve patient outcomes in both oncology and neurology.

Our final study explores the impact of methylation on transcription factor (TF) activity, particularly examining the dynamics of YY1 and TAF1 co-binding and its implications for gene expression in various cellular contexts. YY1 is a crucial TF involved in multiple cellular processes, including cell proliferation, differentiation, DNA repair, and apoptosis. Our ChIP-seq data revealed a significant modification in YY1's binding consensus sequence in the presence of TAF1, marked by consistent preservation of cytosine at the third position within the motif. This alteration suggests a functional shift when YY1 co-binds with TAF1, differing from its regulatory actions when operating independently.

While there is some similarity in the peaks of YY1 and TAF1, minor deviations are specific to the cell type. TAF1 co-binding can influence YY1's regulatory functions, with reduced methylation levels on the YY1 binding motif during co-binding supporting a more transcriptionally active state. Gene expression analysis corroborated these findings, showing that YY1 and TAF1 co-binding sites exhibited increased target gene expression. Such insights are crucial as they reveal the complex interplay between methylation and transcription factor dynamics, offering potential avenues for therapeutic intervention.

Moreover, our GO analysis in GM12878 and H1-hESC cell lines illustrated that YY1 and TAF1 co-binding leads to an expanded set of GO terms compared to when YY1 is expressed independently, suggesting a synergistic effect in regulating cellular processes. This co-binding effect varied between cell types, with a broader range of GO terms associated with promoter-TSS sites in cells capable of differentiation, reflecting the cell-specific needs and differentiation potentials that influence binding patterns and gene regulation.

Despite these advances, our study is limited by the challenges in confirming the physical and functional interactions between YY1 and TAF1 solely through ChIP-seq and GO analysis. Without additional techniques like co-immunoprecipitation, fluorescence resonance energy transfer (FRET), or mutational analyses, the detailed nature of their interaction and its biological implications remain hypothetical. Future research should use these experimental approaches to validate whether YY1 and TAF1 form a cohesive protein complex and elucidate the precise mechanisms by which their co-binding affects gene expression and cellular processes. Our findings propose a revised model for state-dependent co-binding of TAF1 to YY1. When the YY1 binding motif is methylated at the third residue, TAF1 does not co-bind, and if methylation is impossible due to the motif's composition, TAF1 also fails to co-bind. YY1, binding without TAF1, acts more strongly as a repressor. However, in cell-specific motifs where cytosine is conserved at the third residue and the site is unmethylated, TAF1 co-binding to YY1 acts as a transcriptional activator. This investigation into YY1 and TAF1 dynamics reveals a novel mechanism of transcriptional regulation through co-binding, integrating this phenomenon with the broader theme of DNA methylation analysis. This study significantly advances our understanding of how methylation influences transcription factor activity, impacting gene promoter activity and potentially offering new insights into therapeutic strategies.

5.2 Future Research Directions

Our research has provided significant insights into the role of DNA methylation and transcription factor dynamics in KIRP and AD. However, several unanswered questions and areas require further exploration. One area with vast potential for exploration is the precise mapping of epigenetic mechanisms on the gene expression network. While our study identified correlations between methylation signatures and transcriptional activity, the exact molecular pathways through which these epigenetic modifications influence gene expression remain unclear. Additionally, the interaction between different epigenetic modifications, such as histone modifications and non-coding RNA interactions, in conjunction with DNA methylation, leaves room for further exploration in the next few decades.

Emerging technologies hold promise in epigenomic and genomics research, which could enhance future studies. Future studies may wish to employ a multi-faceted approach combining new computational techniques with experimental validation to build on our findings. Potential studies could include functional genomics using CRISPR/Cas9 technology to selectively modify methylation sites and assess the resulting changes in gene expression and cellular phenotypes [147]. This approach could validate the functional significance of the identified methylation markers and elucidate their roles in disease progression. Single-cell epigenomics, leveraging single-cell sequencing technologies to explore the heterogeneity of DNA methylation patterns at the single-cell level, could provide insights into the clonal evolution of tumours and the heterogeneity of epigenetic states within different cell populations in AD.

Longitudinal studies tracking changes in DNA methylation over time in patients with KIRP and AD are another possible approach to help identify early epigenetic changes associated with disease onset and progression, potentially leading to the development of predictive

biomarkers. Integrating DNA methylation data with other omics data, such as transcriptomics, proteomics, and metabolomics, could build a comprehensive understanding of the molecular networks underlying KIRP and AD. This holistic approach could uncover new therapeutic targets and biomarkers.

Furthermore, emerging technologies like single-molecule sequencing—such as nanopore sequencing and single-molecule real-time (SMRT) sequencing—offer the potential to detect DNA methylation and other modifications at single-molecule resolution directly [148]. These techniques can provide a more accurate and comprehensive view of the epigenome. Spatial transcriptomics, which allows for gene expression mapping within the spatial context of tissues, could reveal how epigenetic modifications influence gene expression in specific cellular niches and tissue architectures when combined with DNA methylation analysis. Machine learning frameworks can utilize various strategies in tandem to facilitate and speed up research on large-scale epigenomic data sets, identify complex patterns, and predict functional outcomes. These tools can also aid in integrating multi-omic data and developing predictive models for disease diagnosis and prognosis.

By merging these technological advancements with interdisciplinary approaches, future research can deepen our understanding of epigenetic mechanisms in KIRP and AD, potentially leading to novel therapeutic strategies and improved patient outcomes.

5.4 Concluding Remarks

In conclusion, our research has advanced the understanding of DNA methylation and its role in disease mechanisms. By focusing on computational bioinformatics using publicly available data, we have uncovered many potential insights into the epigenetic regulation of gene

expression in KIRP and AD. Our findings highlight the potential of DNA methylation markers as diagnostic and prognostic tools and provide a framework for future research that utilizes layered machine learning. Additionally, our investigation into the dynamics of YY1 and TAF1 co-binding provided a more nuanced model of gene regulation during the co-binding of these transcription factors.

References

[1] S. B. Baylin and P. A. Jones, “Epigenetic Determinants of Cancer,” *COLD SPRING Harb. Perspect. Biol.*, vol. 8, no. 9, Sep. 2016, doi: 10.1101/cshperspect.a019505.

[2] W. Timp and A. P. Feinberg, “Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host,” *Nat. Rev. CANCER*, vol. 13, no. 7, pp. 497–510, Jul. 2013, doi: 10.1038/nrc3486.

[3] V. V. Levenson, “DNA methylation as a universal biomarker,” *EXPERT Rev. Mol. Diagn.*, vol. 10, no. 4, pp. 481–488, May 2010, doi: 10.1586/ERM.10.17.

[4] A. H. Kit, H. M. Nielsen, and J. Tost, “DNA methylation based biomarkers: Practical considerations and applications,” *BIOCHIMIE*, vol. 94, no. 11, pp. 2314–2337, Nov. 2012, doi: 10.1016/j.biochi.2012.07.014.

[5] L. D. Moore, T. Le, and G. Fan, “DNA Methylation and Its Basic Function,” *NEUROPSYCHOPHARMACOLOGY*, vol. 38, no. 1, pp. 23–38, Jan. 2013, doi: 10.1038/npp.2012.112.

[6] D.-M. Franchini, K.-M. Schmitz, and S. K. Petersen-Mahrt, “5-Methylcytosine DNA Demethylation: More Than Losing a Methyl Group,” *Annu. Rev. Genet.*, vol. 46, no. 1, pp. 419–441, 2012.

[7] A. Hermann, R. Goyal, and A. Jeltsch, “The Dnmt1 DNA-(cytosine-C5)-methyltransferase Methylates DNA Processively with High Preference for Hemimethylated Target Sites,” *J. Biol. Chem.*, vol. 279, no. 46, pp. 48350–48359, 2004.

[8] M. Okano, D. W. Bell, D. A. Haber, and E. Li, “DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development,” *Cell*, vol. 99, no. 3, pp. 247–257, 1999.

[9] Y. Zhang *et al.*, “Dnmt2 mediates intergenerational transmission of paternally acquired metabolic disorders through sperm small non-coding RNAs,” *Nat. Cell Biol.*, vol. 20, no. 5, pp. 535–540, 2018.

[10] A. R. Smith *et al.*, “Parallel profiling of DNA methylation and hydroxymethylation highlights neuropathology-associated epigenetic variation in Alzheimer’s disease,” *Clin. EPIGENETICS*, vol. 11, Mar. 2019, doi: 10.1186/s13148-019-0636-y.

[11] J. Zhao *et al.*, “A genome-wide profiling of brain DNA hydroxymethylation in Alzheimer’s disease,” *ALZHEIMERS Dement.*, vol. 13, no. 6, pp. 674–688, Jun. 2017, doi: 10.1016/j.jalz.2016.10.004.

[12] P. A. JONES, “Functions of DNA methylation: islands, start sites, gene bodies and beyond,” *Nat. Rev. Genet.*, vol. 13, no. 7, pp. 484–492, 2012.

[13] M. V. C. Greenberg and D. Bourc’his, “The diverse roles of DNA methylation in mammalian development and disease,” *Nat. Rev. Mol. Cell Biol.*, vol. 20, no. 10, pp. 590–607, 2019.

[14] C. G. Duncan *et al.*, “Dosage compensation and DNA methylation landscape of the X chromosome in mouse liver,” *Sci. Rep.*, vol. 8, no. 1, pp. 10138–17, 2018.

[15] R. A. Irizarry *et al.*, “The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores,” *Nat. Genet.*, vol. 41, no. 2, pp. 178–186, Feb. 2009, doi: 10.1038/ng.298.

[16] L. A. Pennacchio, W. Bickmore, A. Dean, M. A. Nobrega, and G. Bejerano, “Enhancers: five essential questions,” *Nat. Rev. Genet.*, vol. 14, no. 4, pp. 288–295, Apr. 2013, doi: 10.1038/nrg3458.

[17] H. Hermeking, “MicroRNAs in the p53 network: micromanagement of tumour suppression,” *Nat. Rev. CANCER*, vol. 12, no. 9, pp. 613–626, Sep. 2012, doi: 10.1038/nrc3318.

[18] F. Spitz and E. E. M. Furlong, “Transcription factors: from enhancer binding to developmental control,” *Nat. Rev. Genet.*, vol. 13, no. 9, pp. 613–626, Sep. 2012, doi:

10.1038/nrg3207.
[19] K. T. Prep, *MCAT Biology and Biochemistry Review*. New York, NY: Kaplan Publishing, 2019.
[20] D. Shlyueva, G. Stampfel, and A. Stark, “Transcriptional enhancers: from properties to genome-wide predictions,” *Nat. Rev. Genet.*, vol. 15, no. 4, pp. 272–286, Apr. 2014, doi: 10.1038/nrg3682.
[21] S. A. Lambert *et al.*, “The Human Transcription Factors,” *CELL*, vol. 172, no. 4, pp. 650–665, Feb. 2018, doi: 10.1016/j.cell.2018.01.029.
[22] D. Voet, J. G. Voet, and C. W. Pratt, *Fundamentals of Biochemistry*, 5th ed. New York, NY: John Wiley & Sons, 2023.
[23] J. LEE, K. GALVIN, and Y. SHI, “EVIDENCE FOR PHYSICAL INTERACTION BETWEEN THE ZINC-FINGER TRANSCRIPTION FACTORS YY1 AND SP1,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 90, no. 13, pp. 6145–6149, Jul. 1993, doi: 10.1073/pnas.90.13.6145.
[24] N. M. Luscombe, S. E. Austin, H. M. Berman, and J. M. Thornton, “An overview of the structures of protein-DNA complexes,” *GENOME Biol.*, vol. 1, no. 1, Feb. 2000.
[25] R. MARMORSTEIN, M. CAREY, M. PTASHNE, and S. HARRISON, “DNA RECOGNITION BY GAL4 - STRUCTURE OF A PROTEIN DNA COMPLEX,” *NATURE*, vol. 356, no. 6368, pp. 408–414, Apr. 1992, doi: 10.1038/356408a0.
[26] K. Kluska, J. Adamczyk, and A. Krezel, “Metal binding properties, stability and reactivity of zinc fingers,” *Coord. Chem. Rev.*, vol. 367, pp. 18–64, Jul. 2018, doi: 10.1016/j.ccr.2018.04.009.
[27] S. E. Ahmadi, S. Rahimi, B. Zarandi, R. Chegeni, and M. Safa, “MYC: a multipurpose oncogene with prognostic and therapeutic implications in blood malignancies,” *J. Hematol. Oncol.J Hematol Oncol*, vol. 14, no. 1, Aug. 2021, doi: 10.1186/s13045-021-01111-4.
[28] T. ELLENBERGER, “GETTING A GRIP ON DNA RECOGNITION - STRUCTURES OF THE BASIC REGION LEUCINE-ZIPPER, AND THE BASIC REGION HELIX-LOOP-HELIX DNA-BINDING DOMAINS,” *Curr. Opin. Struct. Biol.*, vol. 4, no. 1, pp. 12–21, Feb. 1994, doi: 10.1016/S0959-440X(94)90054-X.
[29] D. Yesudhas, M. Batool, M. A. Anwar, S. Panneerselvam, and S. Choi, “Proteins Recognizing DNA: Structural Uniqueness and Versatility of DNA-Binding Domains in Stem Cell Transcription Factors,” *GENES*, vol. 8, no. 8, Aug. 2017, doi: 10.3390/genes8080192.
[30] M. Li and J. C. I. Belmonte, “Ground rules of the pluripotency gene regulatory network,” *Nat. Rev. Genet.*, vol. 18, no. 3, pp. 180–191, Mar. 2017, doi: 10.1038/nrg.2016.156.
[31] R. A. Young, “Control of the embryonic stem cell state.,” *Cell*, vol. 144, no. 6, pp. 940–954, Mar. 2011, doi: 10.1016/j.cell.2011.01.032.
[32] M. Hernandez-Hernandez, E. G. Garcia-Gonzalez, C. E. Brun, and M. A. Rudnicki, “The myogenic regulatory factors, determinants of muscle development, cell identity and regeneration,” *Semin. CELL Dev. Biol.*, vol. 72, pp. 10–18, Dec. 2017, doi: 10.1016/j.semcdb.2017.11.010.
[33] T. I. Lee and R. A. Young, “Transcriptional Regulation and Its Misregulation in Disease,” *CELL*, vol. 152, no. 6, pp. 1237–1251, Mar. 2013, doi: 10.1016/j.cell.2013.02.014.
[34] H. Akiyama, M. Chaboissier, J. Martin, A. Schedl, and B. de Crombrugghe, “The transcrintion factor Sox9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of Sox5 and Sox6,” *GENES Dev.*, vol. 16, no. 21, pp. 2813–2828, Nov. 2002, doi: 10.1101/gad.1017802.
[35] P. W. Laird, “Principles and challenges of genome-wide DNA methylation analysis,” *Nat. Rev. Genet.*, vol. 11, no. 3, pp. 191–203, 2010.
[36] R. Pidsley, C. C. Y. Wong, M. Volta, K. Lunnon, J. Mill, and L. C. Schalkwyk, “A data-driven approach to preprocessing Illumina 450K methylation array data,” *BMC GENOMICS*, vol. 14, May 2013, doi: 10.1186/1471-2164-14-293.
[37] T. A. Down *et al.*, “A Bayesian deconvolution strategy for immunoprecipitation-based DNA

methylome analysis," *Nat. Biotechnol.*, vol. 26, no. 7, pp. 779–785, Jul. 2008, doi: 10.1038/nbt1414.
[38] D. Serre, B. H. Lee, and A. H. Ting, "MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome," *NUCLEIC ACIDS Res.*, vol. 38, no. 2, pp. 391–399, Jan. 2010, doi: 10.1093/nar/gkp992.
[39] P. J. Park, "ChIP-seq: advantages and challenges of a maturing technology," *Nat. Rev. Genet.*, vol. 10, no. 10, pp. 669–680, Oct. 2009, doi: 10.1038/nrg2641.
[40] T. S. Furey, "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions," *Nat. Rev. Genet.*, vol. 13, no. 12, pp. 840–852, Dec. 2012, doi: 10.1038/nrg3306.
[41] P. Machanick and T. L. Bailey, "MEME-ChIP: motif analysis of large DNA datasets," *BIOINFORMATICS*, vol. 27, no. 12, pp. 1696–1697, Jun. 2011, doi: 10.1093/bioinformatics/btr189.
[42] Q. X. X. Lin, S. Sian, O. An, D. Thieffry, S. Jha, and T. Benoukraf, "MethMotif: an integrative cell specific database of transcription factor binding motifs coupled with DNA methylation profiles," *NUCLEIC ACIDS Res.*, vol. 47, no. D1, pp. D145–D154, Jan. 2019, doi: 10.1093/nar/gky1005.
[43] I. Yevshin, R. Sharipov, S. Kolmykov, Y. Kondrakhin, and F. Kolpakov, "GTRD: a database on gene transcription regulation2019 update," *NUCLEIC ACIDS Res.*, vol. 47, no. D1, pp. D100–D105, Jan. 2019, doi: 10.1093/nar/gky1128.
[44] Y. Zhang *et al.*, "Model-based Analysis of ChIP-Seq (MACS)," *GENOME Biol.*, vol. 9, no. 9, 2008, doi: 10.1186/gb-2008-9-9-r137.
[45] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nat. Rev. Genet.*, vol. 16, no. 6, pp. 321–332, 2015.
[46] P. Larrañaga *et al.*, "Machine learning in bioinformatics," *Brief. Bioinform.*, vol. 7, no. 1, pp. 86–112, 2006.
[47] A. L. Tarca, V. J. Carey, X. Chen, R. Romero, and S. Drăghici, "Machine learning and its applications to biology," *PLoS Comput. Biol.*, vol. 3, no. 6, pp. e116–e116, 2007.
[48] T. Ching *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *J. R. Soc. Interface*, vol. 15, no. 141, pp. 20170387–20170387, 2018.
[49] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, no. C, pp. 8–17, 2015.
[50] J. L. Balsor *et al.*, "A Practical Guide to Sparse k-Means Clustering for Studying Molecular Development of the Human Brain," *Front. Neurosci.*, vol. 15, pp. 668293–668293, 2021.
[51] M. Baucum, A. Khojandi, and T. Papamarkou, "Hidden Markov models as recurrent neural networks: an application to Alzheimer's disease," 2020.
[52] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA- CANCER J. Clin.*, vol. 68, no. 6, pp. 394–424, Dec. 2018, doi: 10.3322/caac.21492.
[53] D. Hanahan and R. Weinberg, "The hallmarks of cancer," *CELL*, vol. 100, no. 1, pp. 57–70, Jan. 2000, doi: 10.1016/S0092-8674(00)81683-9.
[54] E. N. Marieb and K. Hoehn, *Human Anatomy & Physiology*, 11th ed. New York, NY: Pearson, 2018.
[55] C. Rye, R. Wise, V. Jurukovski, J. DeSaix, J. Choi, and Y. Avissar, *Biology*. Houston, Texas: OpenStax, 2016.
[56] B. N. Lasseigne, T. C. Burwell, M. A. Patil, D. M. Absher, J. D. Brooks, and R. M. Myers, "DNA methylation profiling reveals novel diagnostic biomarkers in renal cell carcinoma," *BMC Med.*, vol. 12, no. 1, pp. 235–235, 2014.
[57] Z. Zou, T. Tao, H. Li, and X. Zhu, "mTOR signaling pathway and mTOR inhibitors in cancer:

progress and challenges," *CELL Biosci.*, vol. 10, no. 1, Mar. 2020, doi: 10.1186/s13578-020-00396-1.
[58] H. Guo *et al.*, "The PI3K/AKT Pathway and Renal Cell Carcinoma," *J. Genet. GENOMICS*, vol. 42, no. 7, pp. 343–353, Jul. 2015, doi: 10.1016/j.jgg.2015.03.003.
[59] Q. Xu, M. Krause, A. Samoylenko, and S. Vainio, "Wnt Signaling in Renal Cell Carcinoma," *CANCERS*, vol. 8, no. 6, Jun. 2016, doi: 10.3390/cancers8060057.
[60] M. Oya and M. Murai, "Renal cell carcinoma: Relevance of pathology," *Curr. Opin. Urol.*, vol. 13, no. 6, pp. 445–449, Nov. 2003, doi: 10.1097/00042307-200311000-00004.
[61] J. Hu *et al.*, "A Novel Pyroptosis-Related Gene Signature for Predicting Prognosis in Kidney Renal Papillary Cell Carcinoma," *Front. Genet.*, vol. 13, Mar. 2022, doi: 10.3389/fgene.2022.851384.
[62] C. G. A. R. Network, "Comprehensive molecular characterization of papillary renal-cell carcinoma," *N. Engl. J. Med.*, vol. 374, no. 2, pp. 135–145, 2016.
[63] T. M. Malta *et al.*, "Machine learning identifies stemness features associated with oncogenic dedifferentiation," *Cell*, vol. 173, no. 2, pp. 338–354, 2018.
[64] K. A. Hoadley *et al.*, "Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer.," *Cell.*, vol. 173, no. 2.
[65] B. Kolb, I. Q. Whishaw, and G. C. Teskey, *An Introduction to Brain and Behavior*, 5th ed. New York, NY: Macmillan Education, 2016.
[66] K. T. Prep, *MCAT Behavioral Sciences Review*, 2020th ed. New York, NY: Kaplan Publishing, 2019.
[67] K. Morgan, "The three new pathways leading to Alzheimer's disease," *Neuropathol. Appl. Neurobiol.*, vol. 37, no. 4, pp. 353–357, Jun. 2011, doi: 10.1111/j.1365-2990.2011.01181.x.
[68] P. Rana *et al.*, "Evaluation of the Common Molecular Basis in Alzheimer's and Parkinson's Diseases," *Int. J. Mol. Sci.*, vol. 20, no. 15, Aug. 2019, doi: 10.3390/ijms20153730.
[69] J. Cummings, G. Lee, A. Ritter, M. Sabbagh, and K. Zhong, "Alzheimer's disease drug development pipeline: 2020," *ALZHEIMERS Dement.-Transl. Res. Clin. Interv.*, vol. 6, no. 1, 2020, doi: 10.1002/trc2.12050.
[70] P. Li *et al.*, "Epigenetic dysregulation of enhancers in neurons is associated with Alzheimer's disease pathology and cognitive symptoms," *Nat. Commun.*, vol. 10, May 2019, doi: 10.1038/s41467-019-10101-7.
[71] G. D. Schellenberg and T. J. Montine, "The genetics and neuropathology of Alzheimer's disease," *Acta Neuropathol. (Berl.)*, vol. 124, no. 3, pp. 305–323, Sep. 2012, doi: 10.1007/s00401-012-0996-2.
[72] Y. Yang and J.-Z. Wang, "Nature of Tau-Associated Neurodegeneration and the Molecular Mechanisms," *J. ALZHEIMERS Dis.*, vol. 62, no. 3, pp. 1305–1317, 2018, doi: 10.3233/JAD-170788.
[73] Z. Chen and C. Zhong, "Oxidative stress in Alzheimer's disease," *Neurosci. Bull.*, vol. 30, no. 2, SI, pp. 271–281, Apr. 2014, doi: 10.1007/s12264-013-1423-y.
[74] E. E. Spangenberg and K. N. Green, "Inflammation in Alzheimer's disease: Lessons learned from microglia-depletion models," *Brain. Behav. Immun.*, vol. 61, pp. 1–11, Mar. 2017, doi: 10.1016/j.bbi.2016.07.003.
[75] M. Nikolac Perkovic *et al.*, "Epigenetics of Alzheimer's Disease," *BIOMOLECULES*, vol. 11, no. 2, Feb. 2021, doi: 10.3390/biom11020195.
[76] N. Coppieters, B. V. Dieriks, C. Lill, R. L. M. Faull, M. A. Curtis, and M. Dragunow, "Global changes in DNA methylation and hydroxymethylation in Alzheimer's disease human brain," *Neurobiol. AGING*, vol. 35, no. 6, pp. 1334–1344, Jun. 2014, doi: 10.1016/j.neurobiolaging.2013.11.031.
[77] L. Chouliaras *et al.*, "Consistent decrease in global DNA methylation and hydroxymethylation in the hippocampus of Alzheimer's disease patients," *Neurobiol. AGING*, vol. 34, no. 9, pp. 2091–2099, Sep. 2013, doi:

10.1016/j.neurobiolaging.2013.02.021.
[78] J. A. Botía *et al.*, “An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks,” *BMC Syst. Biol.*, vol. 11, no. 1, pp. 47–47, 2017.
[79] L. H. Abdul Hadi *et al.*, “miREM: an expectation-maximization approach for prioritizing miRNAs associated with gene-set,” *BMC Bioinformatics*, vol. 19, no. 1, pp. 299–299, 2018.
[80] J.-H. Wei *et al.*, “A CpG-methylation-based assay to predict survival in clear cell renal cell carcinoma,” *Nat. Commun.*, vol. 6, no. 1, pp. 8699–8699, 2015.
[81] C. J. Ricketts, V. K. Hill, and W. M. Linehan, “Tumor-specific hypermethylation of epigenetic biomarkers, including SFRP1, predicts for poorer survival in patients from the TCGA Kidney Renal Clear Cell Carcinoma (KIRC) project,” *PloS One*, vol. 9, no. 1, pp. e85621–e85621, 2014.
[82] P. FISEL *et al.*, “DNA Methylation of the SLC16A3 Promoter Regulates Expression of the Human Lactate Transporter MCT4 in Renal Cancer with Consequences for Clinical Outcome,” *Clin. Cancer Res.*, vol. 19, no. 18, pp. 5170–5181, 2013.
[83] S. Turajlic *et al.*, “Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal,” *CELL*, vol. 173, no. 3, p. 595+, Apr. 2018, doi: 10.1016/j.cell.2018.03.043.
[84] P. L. De Jager *et al.*, “Alzheimer’s disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci,” *Nat. Neurosci.*, vol. 17, no. 9, pp. 1156–1163, Sep. 2014, doi: 10.1038/nn.3786.
[85] K. M. Bakulski *et al.*, “Genome-Wide DNA Methylation Differences Between Late-Onset Alzheimer’s Disease and Cognitively Normal Controls in Human Frontal Cortex,” *J. ALZHEIMERS Dis.*, vol. 29, no. 3, pp. 571–588, 2012, doi: 10.3233/JAD-2012-111223.
[86] K. Lunnon *et al.*, “Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer’s disease,” *Nat. Neurosci.*, vol. 17, no. 9, pp. 1164–1170, Sep. 2014, doi: 10.1038/nn.3782.
[87] S. A. Semick *et al.*, “Integrated DNA methylation and gene expression profiling across multiple brain regions implicate novel genes in Alzheimer’s disease,” *Acta Neuropathol. (Berl.)*, vol. 137, no. 4, pp. 557–569, Apr. 2019, doi: 10.1007/s00401-019-01966-5.
[88] M.-S. Tan, J.-T. Yu, and L. Tan, “Bridging integrator 1 (BIN1): form, function, and Alzheimer’s disease,” *TRENDS Mol. Med.*, vol. 19, no. 10, Art. no. 10, Oct. 2013, doi: 10.1016/j.molmed.2013.06.004.
[89] S. Li, Y. Guo, J. Men, H. Fu, and T. Xu, “The preventive efficacy of vitamin B supplements on the cognitive decline of elderly adults: a systematic review and meta-analysis,” *BMC Geriatr.*, vol. 21, no. 1, Jun. 2021, doi: 10.1186/s12877-021-02253-3.
[90] Y. Shi, J. Lee, and K. Galvin, “Everything you have ever wanted to know about Yin Yang 1,” *Biochim. Biophys. ACTA-Rev. CANCER*, vol. 1332, no. 2, pp. F49–F66, Apr. 1997, doi: 10.1016/S0304-419X(96)00044-3.
[91] S. Gordon, G. Akopyan, H. Garban, and B. Bonavida, “Transcription factor YY1: structure, function, and therapeutic implications in cancer biology,” *ONCOGENE*, vol. 25, no. 8, pp. 1125–1142, Feb. 2006, doi: 10.1038/sj.onc.1209080.
[92] Y. Shi, E. Seto, L.-S. Chang, and T. Shenk, “Transcriptional repression by YY1, a human GLI-Kruppel-related protein, and relief of repression by adenovirus E1A protein,” *Cell*, vol. 67, no. 3, pp. 377–388, 1991, doi: 10.1016/0092-8674(91)90189-6.
[93] S. Baritaki and A. Zaravinos, “Cross-Talks between RKIP and YY1 through a Multilevel Bioinformatics Pan-Cancer Analysis,” *CANCERS*, vol. 15, no. 20, Oct. 2023, doi: 10.3390/cancers15204932.
[94] S. Vivarelli *et al.*, “Computational Analyses of YY1 and Its Target RKIP Reveal Their Diagnostic and Prognostic Roles in Lung Cancer,” *CANCERS*, vol. 14, no. 4, Feb. 2022, doi: 10.3390/cancers14040922.
[95] M. Thomas and E. Seto, “Unlocking the mechanisms of transcription factor YY1: are

chromatin modifying enzymes the key?," *GENE*, vol. 236, no. 2, pp. 197–208, Aug. 1999, doi: 10.1016/S0378-1119(99)00261-9.
[96] M. C. Thomas and C.-M. Chiang, "The general transcription machinery and general cofactors," *Crit. Rev. Biochem. Mol. Biol.*, vol. 41, no. 3, pp. 105–178, Jun. 2006, doi: 10.1080/10409230600648736.
[97] R. K. Louder, Y. He, J. Ramon Lopez-Blance, J. Fang, P. Chacon, and E. Nogales, "Structure of promoter-bound TFIID and model of human pre-initiation complex assembly," *NATURE*, vol. 531, no. 7596, p. 604+, Mar. 2016, doi: 10.1038/nature17394.
[98] A. B. Patel *et al.*, "Structure of human TFIID and mechanism of TBP loading onto promoter DNA," *SCIENCE*, vol. 362, no. 6421, SI, p. 1376+, Dec. 2018, doi: 10.1126/science.aau8872.
[99] E. Cerami *et al.*, "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data," *CANCER Discov.*, vol. 2, no. 5, pp. 401–404, May 2012, doi: 10.1158/2159-8290.CD-12-0095.
[100] J. Gao *et al.*, "Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal," *Sci. Signal.*, vol. 6, no. 269, Apr. 2013, doi: 10.1126/scisignal.2004088.
[101] J. N. Weinstein *et al.*, "The Cancer Genome Atlas Pan-Cancer analysis project," *Nat. Genet.*, vol. 45, no. 10, SI, pp. 1113–1120, Oct. 2013, doi: 10.1038/ng.2764.
[102] G. F. Gao *et al.*, "Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data," *CELL Syst.*, vol. 9, no. 1, p. 24+, Jul. 2019, doi: 10.1016/j.cels.2019.06.006.
[103] W. Zhou, T. J. Triche Jr., P. W. Laird, and H. Shen, "SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions," *NUCLEIC ACIDS Res.*, vol. 46, no. 20, Nov. 2018, doi: 10.1093/nar/gky691.
[104] P. Du *et al.*, "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis," *BMC Bioinformatics*, vol. 11, Nov. 2010, doi: 10.1186/1471-2105-11-587.
[105] C. Xie, Y.-K. Leung, A. Chen, D.-X. Long, C. Hoyo, and S.-M. Ho, "Differential methylation values in differential methylation analysis," *BIOINFORMATICS*, vol. 35, no. 7, pp. 1094–1097, Apr. 2019, doi: 10.1093/bioinformatics/bty778.
[106] B. T. Sherman *et al.*, "DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update)," *NUCLEIC ACIDS Res.*, vol. 50, no. W1, Art. no. W1, Jul. 2022, doi: 10.1093/nar/gkac194.
[107] B. T. Sherman *et al.*, "DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update)," *NUCLEIC ACIDS Res.*, vol. 50, no. W1, pp. W216–W221, Jul. 2022, doi: 10.1093/nar/gkac194.
[108] J. M. Teodoridis, C. Hardie, and R. Brown, "CpG island methylator phenotype (CIMP) in cancer: Causes and implications," *CANCER Lett.*, vol. 268, no. 2, pp. 177–186, Sep. 2008, doi: 10.1016/j.canlet.2008.03.022.
[109] R. Limame, K. O. de Beeck, F. Lardon, O. De Wever, and P. Pauwels, "Kruppel-like factors in cancer progression: three fingers on the steering wheel," *ONCOTARGET*, vol. 5, no. 1, pp. 29–48, Jan. 2014, doi: 10.18632/oncotarget.1456.
[110] S. Safe and M. Abdelrahim, "Sp transcription factor family and its role in cancer," *Eur. J. CANCER*, vol. 41, no. 16, pp. 2438–2448, Nov. 2005, doi: 10.1016/j.ejca.2005.08.006.
[111] C. G. Bailey *et al.*, "CTCF Expression is Essential for Somatic Cell Viability and Protection Against Cancer," *Int. J. Mol. Sci.*, vol. 19, no. 12, Dec. 2018, doi: 10.3390/ijms19123832.
[112] X. Li *et al.*, "Wilms' tumour gene 1 (WT1) enhances non-small cell lung cancer malignancy and is inhibited by microRNA-498-5p," *BMC CANCER*, vol. 23, no. 1, Sep. 2023, doi: 10.1186/s12885-023-11295-2.

[113] B. Zhou *et al.*, “WT1 facilitates the self-renewal of leukemia-initiating cells through the upregulation of BCL2L2: WT1-BCL2L2 axis as a new acute myeloid leukemia therapy target,” *J. Transl. Med.*, vol. 18, no. 1, Jun. 2020, doi: 10.1186/s12967-020-02384-y.
[114] A. Adjei and M. Hidalgo, “Intracellular signal transduction pathway proteins as targets for cancer therapy,” *J. Clin. Oncol.*, vol. 23, no. 23, pp. 5386–5403, Aug. 2005, doi: 10.1200/JCO.2005.23.648.
[115] Q. Xu, M. Krause, A. Samoylenko, and S. Vainio, “Wnt Signaling in Renal Cell Carcinoma,” *CANCERS*, vol. 8, no. 6, Art. no. 6, Jun. 2016, doi: 10.3390/cancers8060057.
[116] J. S. Hale, M. Li, and J. D. Lathia, “The malignant social network Cell-cell adhesion and communication in cancer stem cells,” *CELL Adhes. Migr.*, vol. 6, no. 4, pp. 346–355, Aug. 2012, doi: 10.4161/cam.21294.
[117] U. Lee, E.-Y. Cho, and E.-H. Jho, “Regulation of Hippo signaling by metabolic pathways in cancer*,” *Biochim. Biophys. ACTA-Mol. CELL Res.*, vol. 1869, no. 4, Apr. 2022, doi: 10.1016/j.bbamcr.2021.119201.
[118] M. Blanco Calvo, V. Bolos Fernandez, V. Medina Villaamil, G. Aparicio Gallego, S. Diaz Prado, and E. Grande Pulido, “Biology of BMP signalling and cancer,” *Clin. Transl. Oncol.*, vol. 11, no. 3, pp. 126–137, Mar. 2009, doi: 10.1007/S12094-009-0328-8.
[119] N. S. Nagaraj and P. K. Datta, “Targeting the transforming growth factor-β signaling pathway in human cancer,” *EXPERT Opin. Investig. DRUGS*, vol. 19, no. 1, pp. 77–91, Jan. 2010, doi: 10.1517/13543780903382609.
[120] G. B. Bolger, “The cAMP-signaling cancers: Clinically-divergent disorders with a common central pathway,” *Front. Endocrinol.*, vol. 13, Oct. 2022, doi: 10.3389/fendo.2022.1024423.
[121] S. K. Rajasekharan and T. Raman, “Ras and Ras mutations in cancer,” *Cent. Eur. J. Biol.*, vol. 8, no. 7, pp. 609–624, Jul. 2013, doi: 10.2478/s11535-013-0158-5.
[122] N. Nomi, S. Kodama, and M. Suzuki, “Toll-like receptor 3 signaling induces apoptosis in human head and neck cancer via survivin associated pathway,” *Oncol. Rep.*, vol. 24, no. 1, pp. 225–231, Jul. 2010, doi: 10.3892/or_00000850.
[123] T. Matijevic and J. Pavelic, “The dual role of TLR3 in metastatic cell line,” *Clin. Exp. METASTASIS*, vol. 28, no. 7, pp. 701–712, Oct. 2011, doi: 10.1007/s10585-011-9402-z.
[124] B. E. Lippitz and R. A. Harris, “Cytokine patterns in cancer patients: A review of the correlation between interleukin 6 and prognosis,” *ONCOIMMUNOLOGY*, vol. 5, no. 5, Art. no. 5, 2016, doi: 10.1080/2162402X.2015.1093722.
[125] L. Peterfi, M. V. Yusenko, and G. Kovacs, “IL6 Shapes an Inflammatory Microenvironment and Triggers the Development of Unique Types of Cancer in End-stage Kidney,” *ANTICANCER Res.*, vol. 39, no. 4, Art. no. 4, Apr. 2019, doi: 10.21873/anticanres.13294.
[126] H. Zhang and G. Zhu, “Beyond Promoter: The Role of Macrophage in Invasion and Progression of Renal Cell Carcinoma,” *Curr. STEM CELL Res. Ther.*, vol. 15, no. 7, pp. 588–596, 2020, doi: 10.2174/1574888X15666200225093210.
[127] A. R. de V. Chevez, J. Finke, and R. Bukowski, “The Role of Inflammation in Kidney Cancer,” in *INFLAMMATION AND CANCER*, vol. 816, B. Aggarwal, B. Sung, and S. Gupta, Eds., in Advances in Experimental Medicine and Biology, vol. 816. , 2014, pp. 197–234. doi: 10.1007/978-3-0348-0837-8_9.
[128] C. I. Weidner *et al.*, “Aging of blood can be tracked by DNA methylation changes at just three CpG sites,” *GENOME Biol.*, vol. 15, no. 2, 2014, doi: 10.1186/gb-2014-15-2-r24.
[129] W. Ding, D. Kaur, S. Horvath, and W. Zhou, “Comparative epigenome analysis using Infinium DNA methylation BeadChips,” *Brief. Bioinform.*, vol. 24, no. 1, Jan. 2023, doi: 10.1093/bib/bbac617.
[130] T. J. Triche Jr., D. J. Weisenberger, D. Van Den Berg, P. W. Laird, and K. D. Siegmund, “Low-level processing of Illumina Infinium DNA Methylation BeadArrays,” *NUCLEIC ACIDS*

Res., vol. 41, no. 7, Apr. 2013, doi: 10.1093/nar/gkt090.
[131] S. Horvath, “DNA methylation age of human tissues and cell types,” *GENOME Biol.*, vol. 14, no. 10, 2013, doi: 10.1186/gb-2013-14-10-r115.
[132] Q. X. X. Lin, D. Thieffry, S. Jha, and T. Benoukraf, “TFregulomeR reveals transcription factors’ context-specific features and functions,” *NUCLEIC ACIDS Res.*, vol. 48, no. 2, Jan. 2020, doi: 10.1093/nar/gkz1088.
[133] “FastQC.” Jun. 2015. [Online]. Available: https://qubeshub.org/resources/fastqc
[134] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EJ Bioinforma.*, vol. 17, no. 1, pp. 10–12, Jan. 2011, doi: 10.14806/ej.17.1.200.
[135] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, “Salmon provides fast and bias-aware quantification of transcript expression,” *Nat. METHODS*, vol. 14, no. 4, p. 417+, Apr. 2017, doi: 10.1038/nmeth.4197.
[136] ENCODE Project Consortium, “total RNA-seq from GM12878 (ENCSR000AEE),” *Gene Expr. Omn. GEO*, vol. GSE78552, Mar. 2016, [Online]. Available: https://www.encodeproject.org/ENCSR000AEE/
[137] ENCODE Project Consortium, “total RNA-seq from H1 (ENCSR588EJX),” *Gene Expr. Omn. GEO*, Nov. 2021, [Online]. Available: https://www.encodeproject.org/ENCSR588EJX/
[138] W. Kent *et al.*, “The human genome browser at UCSC,” *GENOME Res.*, vol. 12, no. 6, pp. 996–1006, Jun. 2002, doi: 10.1101/gr.229102.
[139] S. Chen, Y. Zhou, Y. Chen, and J. Gu, “fastp: an ultra-fast all-in-one FASTQ preprocessor,” *BIOINFORMATICS*, vol. 34, no. 17, pp. 884–890, Sep. 2018, doi: 10.1093/bioinformatics/bty560.
[140] Z. Gu and D. Huebschmann, “rGREAT: an R/Bioconductor package for functional enrichment on genomic regions,” *Bioinformatics*, 2022, doi: 10.1093/bioinformatics/btac745.
[141] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, “Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt,” *Nat. Protoc.*, vol. 4, pp. 1184–1191, 2009.
[142] H. Wickham, R. François, L. Henry, K. Müller, and D. Vaughan, *dplyr: A Grammar of Data Manipulation*. 2023. [Online]. Available: https://CRAN.R-project.org/package=dplyr
[143] H. Wickham *et al.*, “Welcome to the tidyverse,” *J. Open Source Softw.*, vol. 4, no. 43, p. 1686, 2019, doi: 10.21105/joss.01686.
[144] H. Wickham, D. Vaughan, and M. Girlich, *tidyr: Tidy Messy Data*. 2023. [Online]. Available: https://CRAN.R-project.org/package=tidyr
[145] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org
[146] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2023. [Online]. Available: https://www.R-project.org/
[147] P. I. Thakore, J. B. Black, I. B. Hilton, and C. A. Gersbach, “Editing the epigenome: technologies for programmable transcription and epigenetic modulation,” *Nat. METHODS*, vol. 13, no. 2, pp. 127–137, Feb. 2016, doi: 10.1038/NMETH.3733.
[148] B. A. Flusberg *et al.*, “Direct detection of DNA methylation during single-molecule, real-time sequencing,” *Nat. METHODS*, vol. 7, no. 6, pp. 461-U72, Jun. 2010, doi: 10.1038/NMETH.1459.